Signaling and detecting uncertainty in audiovisual speech by children and adults

Emiel Krahmer and Marc Swerts

Communication & Cognition, Tilburg University, The Netherlands
{e.j.krahmer, m.g.j.swerts}@uvt.nl

Abstract

We describe two experiments on signaling and detecting uncertainty in audiovisual speech by adults and children. In the first study, utterances from adult speakers and child speakers (aged 7-8) were elicitated and annotated with a set of six audiovisual features. It was found that when adult speakers are uncertain about their answer they are more likely to produce filled pauses, delays, high intonation, eyebrow movements, smiles and funny faces. The basic picture for the child speakers is similar, in that the presence of an audiovisual cue in an answer correlates with uncertainty, but the differences are relatively small and only significant for the features delay, eyebrow and funny face. In the second study both adult and child judges watched answers from adult and child speakers selected from the first study to find out whether they were able to correctly estimate a speakers' level of uncertainty. It was found that both child and adult judges give more accurate scores for answers from adult speakers than from child speakers and that child judges overall provide less accurate scores than adult judges.

1. Introduction

In recent years, a number of studies have researched the production and perception of uncertainty in interaction.[1] Smith and Clark (1993), for instance, have studied the way speakers signal uncertainty in factual question-answering situations. They used the *feeling of knowing* (FOK) paradigm, originally due to Hart (1965), and found that when speakers are uncertain about the correctness of their answer they signal this using a variety of prosodic cues such as filled pauses, longer delays, and rising intonation. Following up on this, Brennan & Williams (1995) focussed on the perception of uncertainty using a variant of the FOK paradigm which they dubbed the *feeling of another's knowing* (FOAK). They found that the uncertainty cues from Smith and Clark's work indeed have communicative relevance, since listener's use them to adequately estimate the level of certainty of a speaker's answer.

Arguably these studies are incomplete in at least two ways. First, they only pay attention to auditive cues to uncertainty. It appears that speakers also signal their level of uncertainty visually, and that such combined audio-visual signaling leads to a more accurate perception of uncertainty (Swerts et al. 2003). Second, previous work only focussed on adult speakers. The main research question addressed in the current paper is whether young elementary school children (age 7-8) signal and detect uncertainty in the same way as adults do. Children in this age group have already developed a fairly complete "theory of mind" (see e.g., Flavell 1999 and the references cited therein). Unlike pre-schoolers, for instance, young elementary school children can reflect on their knowledge in a meta-cognitive way. But, it might be that children signal their uncertainty in a different way or to a different extent than adults do, given that children may use a somewhat different prosodic and gestural repertoire than adults.

The remainder of this paper is structured as follows. In section 2 we describe and compare the results of two feeling of knowing experiments, one with adults and one with children. In section 3, we report on four feeling of another's knowing experiments with children and adults *both* as speakers *and* as judges. We end with some concluding remarks in section 4.

2. Experiment I: Signaling Uncertainty

2.1. Method

2.1.1. Overview Following Smith and Clark (1993), we used Hart's (1965) three step procedure to collect certain and uncertain speaker utterances from both children and adults. In the first step, speakers were asked a series of factual questions, and their answers were recorded with a digital camera. In the second step, they had to indicate for the same questions as in the first round, how certain they are that they would recognize the answer in a multiple choice test. This score is the actual "feeling of knowing" (FOK) score. The third step in the procedure was the actual multiple choice test. Afterwards utterances recorded in the first step were labeled using a set of audio-visual features.

2.1.2. Subjects Twenty adults and twenty-one children participated as speakers. The adults (11 males, 9 females) were colleagues and students from Tilburg University, all between 20 and 50 years old. The children (9 boys, 12 girls) were in Group 4 (i.e., 2nd grade in the American school system) of 't Schrijverke ("the little writer"), an elementary school in Goirle (a small town adjacent to Tilburg). They were all between 7 and 8 years old.

2.1.3. Stimuli For both adults and children the stimuli consisted of a series of factual questions (40 for adults, 30 for children). For adults, the questions were selected from the Dutch version of the "Wechsler Adult Intelligence Scale" (WAIS), a standard intelligence test for adults. We only selected those questions which would trigger a one-word response (e.g. Who wrote Hamlet? What is the capital of Switzerland?), and added a supplementary list from the game Trivial Pursuit. For children, the questions were taken from the Dutch version of the "Wechsler Intelligence Scale for Children" (WISC). Again, we only selected those questions that allowed for a single word answer, and supplemented them with questions from the Dutch version of Trivial Pursuit for children (e.g., How much is a dozen? Who discovered America?) Speakers were always confronted with this list of questions in one of two random orders.

[1]The research described in this paper was conducted as part of the VIDI-project "Functions Of Audiovisual Prosody (FOAP)", sponsored by the Dutch NSF (NWO), see www.foap.tk. Swerts is also affiliated with the Flemish Fund for Scientific Research (FWO-Flanders) and Antwerp University. Thanks to Judith Schrier (Antwerp) and Jorien Scholze, Kim Smulders and Nicole Hobbelen (Tilburg) for their help in carrying out the experiments. Thanks to Lennard van de Laar and Pashiera Barkhuysen for their help with the annotation and the experimental set-up.

Figure 1: *Four stills taken from the adult and the child experiments: on the left are responses where the speakers have a high FOK, on the right where they have a low FOK.*

2.1.4. Experimental procedure Child and adult speakers underwent the same procedure, modulo some small differences detailed below.

First, speakers were asked the series of questions by the experimenter, whom they could not see, and the speakers' responses were filmed. Second, after this test, the same sequence of questions was presented again, but now they had to indicate how sure they were that they would recognize the correct answer if they would have to find it in a multiple-choice test. For adults, a 7-point Likert scale was used. For children in this age group, a standard 7-point Likert scale might cause problems, hence we opted for a 5-point Likert scale using a facial representation of the items.[2] For the purpose of comparison, both Likert scales were recoded to the interval [0,1], with 0 = "absolutely not" and 1 = "definitely yes". Throughout this paper, these scores are referred to as FOK scores. A speaker's utterance is said to be **uncertain** if the corresponding FOK score is low, and **certain** if the FOK score is high. See Figure 1 for some stills illustrating high and low FOK responses. Third, the final test was a paper-and-pencil test in which the same sequence of questions was now presented in a multiple-choice in which the correct answer was mixed with three plausible alternatives. For instance, the question "What is the capital of Switzerland?" listed Bern (correct) with three other large Swiss cities: Zürich, Genève and Basel.

As an illustration, consider 4 actual responses from the child experiment (translated from Dutch) to the question "Who discovered America?": a. *Columbus*; b. *Saddam Hussein*; c. *Pirates*; d. *Uh I don't know*. This example shows cases of correct (a), incorrect (b and c), and non-answers (d).

Table 1 present the average FOK scores for adults and children respectively as a function of different response categories. The first thing to note is that the overall pictures are strikingly similar. Both children and adults gave the highest FOK scores to correct answers and the lowest to non-answers in the open question test. Similarly, in the multiple choice test the correct answers have a much higher FOK score than the incorrect ones. This indicates that the children understood the experiment and could perform it as intended.

[2]This so-called "smileymeter" is a validated test scale that is used, among other things, for usability tests of children's software (Read et al. 2002). To reduce the chances of misunderstandings to a minimum, we used an extended training session (consisting of 10 questions) in which the scale and the experimental question were explained and illustrated.

Table 1: *Average FOK scores for children and adults for different response categories.*

Experiment	Response	FOK scores for Adult	Child
Open Question	Answers	0.90	0.90
	Correct Answers	0.94	0.96
	Incorrect Answers	0.70	0.54
	Non-answers	0.43	0.36
Multiple Choice	Correct Answers	0.88	0.86
	Incorrect Answers	0.55	0.56

2.1.5. Labeling and annotation All utterances from the first test (adults and children) were transcribed orthographically and manually labeled with a number of auditive and visual features by four independent transcribers on the basis of an explicit labeling protocol. We labeled the presence or absence of the following verbal and visual features:

Filled pause Whether the utterance contained one or more filled pauses ('uh', 'uhm'), or whether these were absent.

Delay Whether a speaker responded immediately, or took some time to respond.

High intonation Whether a speaker's utterance ended in a high or a low boundary tone.

Eyebrow movement If one or more eyebrow movements departed from neutral position during the utterance.

Smile If the speaker smiled (even silently) during the response.

Funny face Whether the speaker produced a 'marked facial expression', of the type illustrated in the right stills in Figure 1.

2.2. Results

Tables 2 and 3 display the labeling results for adult answers and non-answers respectively. Table 2 shows that the presence of a verbal or visual feature in answers always coincides with a significantly lower FOK score, whereas table 3 shows that the presence of such a feature in non-answers leads to higher FOK scores, be it that not all of the differences are significant, probably because of the relatively limited number of data (all tests for significance done with paired t-tests). Tables 4 and 5 describe the labeling results for child answers and non-answers. The picture for child answers and non-answers is similar to that for the adults, but the results are overall less pronounced. Looking at the answers in Table 4, we see that in most cases the presence of a verbal or visual feature is associated with a lower FOK score, albeit that the difference is only significant for the features delay, eyebrow and funny face. It is surprising that filled pauses (a strong cue for adult uncertainty) plays only a marginal role for uncertainty signaling in children. It is also noteworthy that smile corresponds positively (but non-significantly) with FOK score. For the child speakers' non-answers (Table 5), the presence of filled pauses, delays or high intonation is associated with somewhat higher FOK scores, but the only significant difference, however, is for smile.

Table 2: *Average adult FOK scores for answers ($n = 704$) as a function of presence or absence of audiovisual features.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filled pause	0.83	0.93	–0.10***
Delay	0.75	0.93	–0.18***
High Intonation	0.87	0.92	–0.05***
Eyebrow	0.82	0.92	–0.10***
Smile	0.87	0.91	–0.04*
Funny Face	0.74	0.91	–0.16**

*$p < .05$; **$p < .01$; ***$p < .001$

Table 3: *Average adult FOK scores for non-answers ($n = 96$) as a function of presence or absence of audiovisual features.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filled pause	0.71	0.38	0.33***
Delay	0.54	0.34	0.20***
High Intonation	0.57	0.41	0.16*
Eyebrow	0.51	0.41	0.10
Smile	0.50	0.41	0.09
Funny Face	0.45	0.43	0.02

*$p < .05$; **$p < .01$; ***$p < .001$

Table 4: *Average children FOK scores for answers ($n = 496$) as a function of presence or absence of audiovisual features.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filled pause	0.88	0.89	–0.01
Delay	0.70	0.94	–0.24***
High Intonation	0.88	0.92	–0.04
Eyebrow	0.81	0.91	–0.10***
Smile	0.92	0.89	0.03
Funny Face	0.66	0.92	–0.26***

***$p < .001$

Table 5: *Average children FOK scores for non-answers ($n = 131$) as a function of presence or absence of audiovisual features.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filled pause	0.41	0.35	0.06
Delay	0.37	0.33	0.04
High Intonation	0.42	0.34	0.08
Eyebrow	0.36	0.36	0.00
Smile	0.43	0.34	0.09*
Funny Face	0.36	0.36	0.00

*$p < .05$

2.3. Discussion

In the first experiment the FOK paradigm was used to elicitate certain and uncertain utterances from adult and child speakers. From the labeling analysis it appears that particular audiovisual surface forms of the utterances produced by the adult speakers are indicative of the amount of confidence speakers have about the correctness of their response. For answers, lower FOK scores correlate with the presence of delays, filled pauses, high intonation, eyebrows, funny faces and smiling. For non-answers, the relationships between FOK scores and the different audiovisual features is the mirror image of the outcome with answers. Arguably, a speaker who utters a low FOK non-answer is relatively certain that he or she does not know the answer, and hence does not need to start 'looking' for the answer. These results generalize the earlier finding of Smith and Clark (1993) that answers and non-answers differ in speaker behaviour. Interestingly, the overall picture for child speakers is somewhat similar but much less pronounced than that for adults. For child answers, lower FOK scores generally correlate with the presence of audiovisual cues (excepting smile), but this trend is only significant for the features delay, eyebrows and funny faces. For child non-answers, the relation between FOK scores and audio-visual features indeed appears to mirror the results for answers, but only for smile does this lead to a significant result.[3] In sum, it seems fair to conclude that adult speakers use audio-visual cues more consistently to signal their level of certainty than child speakers.

3. Experiment II: Detecting Uncertainty

3.1. Method

3.1.1. Overview The first study focussed on speakers' production of uncertainty, to gain insight into audiovisual correlates of FOK. In the second study we investigate the perceptual relevance of such cues. For this, we use earlier work by Brennan and Williams (1995) as our main source of inspiration. In this experimental set-up, judges are presented with speaker utterances and they have to estimate the speaker's level of certainty. This estimation is referred to as a "feeling of another's knowing" (FOAK) score.

3.1.2. Subjects 80 native speakers of Dutch participates as judges, 40 adults and 40 children, all different from the speakers that participated in the production studies. The children came from two different group 4 classes from the St. Jozef School, all were 7 or 8 years old. The adults were all students from Tilburg University, between 18 and 25 years.

3.1.3. Stimuli From the corpus of answers collected in the first study, we selected 30 adult utterances and 30 child utterances, both with an equal distribution of high FOK and low FOK scores. Given the individual differences in the use of the FOK scale, we chose to use —per speaker— the highest score as an instantiation of high FOK and the two lowest as representations of low FOK. The original selection of stimuli was random, but utterances were iteratively replaced until the following criteria were met: (1) the original question posed by the experimenter should not re-appear in the speakers' response; (2) all the answers should be lexically distinct; and (3) there should be maximally two answers per speaker. Having applied this procedure on the basis of written transcriptions of the data, we finally replaced some stimuli, if the background noise made them unsuitable for the experiment. Initially we planned to include non-answers in the perception experiment as well, but since we could not find sufficiently many high FOK non-answers meeting the criteria among the children's responses we had to give up this intention.

3.1.4. Experimental design The experiment had a 2 x 2 balanced Latin square design with original FOK score as a within subjects factor and speaker and judge as between subjects factors. Thus, of the 40 adult judges, 20 judged stimuli from adult speakers, and 20 from child speakers, and likewise for the 40 child judges.

3.1.5. Experimental procedure Stimuli were presented on a screen where the judges first saw the stimulus ID (1 through 30) and then

[3] Both the analyses of adult and of child speakers' data is limited in that we have not fully explored possible interactions between cues. Unfortunately, this was not possible, since one quickly runs into sparse data problems, as not every combination of features is well represented in the dataset.

Table 6: *Average FOAK scores for adult and child judges as a function of speaker type and FOK score.*

			FOAK scores of	
		FOK	Adult judges	Child judges
Speaker	Adult	High	0.79	0.66
		Low	0.32	0.47
	Child	High	0.70	0.70
		Low	0.44	0.53
	Average		0.56	0.59

Table 7: *Average FOAK difference scores for adult and child judges as a function of speaker type. Differences are computed via FOAK for high FOK stimuli minus FOAK for low FOK stimuli.*

		FOAK difference for	
		Adult judges	Child judges
Speaker	Adult	0.47	0.26
	Child	0.19	0.17

the actual stimulus. The inter-stimulus interval was 3 seconds for adult judges and 6 seconds for child judges. For all four groups, the experiment was preceded by a short exercise session to make judges acquainted with the kinds of stimulus materials and the procedure. Judges were instructed to estimate whether speakers were uncertain about their answers or not. Adult judges scored this on a 7-point Likert scale, child judges on a 5-point Likert scale (using the same representation as above). For presentation purposes, both scales are recoded to the interval [0, 1], with 0 = "very uncertain" and 1 = "very certain". These scores are referred to as the Feeling Of Another's Knowing (FOAK) scores.

3.2. Results

The overall results are summarized in Table 6. There was a small but significant main effect of speaker ($F(1, 76) = 6.574, p < .05$) (all tests for significance done with ANOVA). This means that overall child speakers received slightly higher FOAK scores than adults (0.59 and 0.56 respectively). Additionally, a main effect of FOK score was found ($F(1, 76) = 601.987, p < .001$). As one would expect, stimuli with a high FOK score get overall higher FOAK scores than stimuli with a low FOK.

To see whether there are differences between adults and children, we have to look at the interactions. A significant two-way interaction was found between speaker and FOK score ($F(1, 76) = 26.281, p < .01$). Inspection of Table 6 reveals that stimuli from adult speakers receive more diverging FOAK scores than stimuli from child speakers, irrespective of who the judges are. There was also a significant interaction between judge and FOK score ($F(1, 76) = 62.726, p < .01$); differences between FOAK scores for high and low FOK stimuli are larger for adult than for child judges (cf. Table 7). This same table can also be used to interpret the significant 3-way interaction between judge, speaker and FOK ($F(1, 76) = 19.419, p < .01$), it shows that the FOAK scores for high FOK and low FOK stimuli differ most for adults judging adults and least for children judging children.

3.3. Discussion

The experiment revealed a number of clear differences between adult and children, both as speakers and as judges. Overall, we found that FOAK scores assigned to stimuli from adult speakers offer a more accurate reflection of the original FOK score than stimuli from child speakers. This is not unexpected: what is not clearly signalled, can not be detected, and the results from the first experiment already revealed that adult speakers are more systematic in cuing their (un)certainty. The results of the current experiment also indicate that adults are "better" judges of uncertainty than children. Perhaps this can be explained along similar lines: what you do not signal yourself, is probably also more difficult to detect in others.

4. General Discussion and Perspectives

We have described two experiments on signaling and detecting uncertainty in audiovisual speech by adults and children. In the first study, we found that adult speakers signal their level of uncertainty in a more consistent and more clear way than our child speakers. The results for the child speakers display similar trends as those for the adults, but in general the results are much less pronounced and only in a few cases did we find significant differences in FOK scores when a particular audiovisual feature is present or absent. From the second study, we may conclude that both child and adult judges give more accurate FOAK scores (i.e., make better estimations of a speaker's level of certainty) for answers from adult than from child speakers. This is in line with the findings of the first study ("what is not signalled cannot be detected"). The second study also revealed that child judges overall provide less accurate FOAK scores than adult judges. Arguably, it is difficult to correctly interpret behavior that a person does not use himself.

These outcomes raise an interesting question: why is it that children between 7 and 8 years old, who are capable of forming metacognitive judgements about knowledge, do not signal their uncertainty as adults do? We conjecture that this has something to do with self-presentation. Adult speakers clearly indicate while answering, whether they are certain about the correctness of their answer or not. In doing so, they can save face when an answer turns out to be incorrect. Children in the age group 7-8 are apparently less worried about this aspect of question-answering. In this respect it is interesting to observe that delay is a significant cue for child uncertainty while filled pauses are not. A child that does not know the answer immediately cannot avoid a longer memory search, but does not signal this to the listener via a filled pause as an adult would do. It would be interesting to redo the current experiments with a slightly older age group to see if and when the notion of self-presentation is more fully developed.

5. References

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

[2] Flavell, J. (1999), Cognitive development: Childrens' knowledge about the mind, *Annual Review of Psychology* **50**: 21-40.

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

[4] Read, J., MacFarlane, S. and Casey, C. (2002), Endurability, engagement and expectations, in: *Interaction Design and Children*, M. Bekker, et al. (eds.), Shaker Publishing.

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

[6] Swerts, M., Krahmer, E., Barkhuysen, P., van de Laar, L. (2003), Audiovisual cues to uncertainty, in: *ISCA Workshop on Error Handling in Spoken Dialog Systems*, Switzerland.