Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System[*]

Erwin Marsi, Ferdi van Rooden
Communication and Cognition
Tilburg University
The Netherlands
e.c.marsi@uvt.nl

Abstract

One of the strategies that question-answering (QA) systems may follow to retain users' trust is to express the level of uncertainty attached to answers they provide. Multimodal QA systems offer the opportunity to express this uncertainty through other than linguistic means. On the basis of evidence from the literature, it is argued that uncertainty is in fact better expressed by audiovisual than by verbal means. We summarize unpublished work on audiovisual expression of uncertainty in the context of QA systems which suggests that users prefer visual over linguistic signaling. Next, we describe a perception experiment showing that uncertainty can be reliably expressed by means of a talking head using a limited repertoire of animated facial expressions, i.e. only combinations of eyebrow and head movements. In addition, we discuss a number of open issues that need to be resolved before a talking head can really be employed for signaling uncertainty in multimodal human-computer interaction.

Keywords: certainty, confidence, trust, facial expression, facial animation, embodied conversational agents, talking heads, multimodal dialogue, question answering

1 Introduction

A commonly held opinion among researchers in the field on automatic question answering (QA) is that "incorrect answers are worse than no answers" (Burger et al., 2003). Incorrect answers evidently make the system look unreliable and undermine the user's trust in its capabilities. Since flawless QA systems are unlikely to appear soon, strategies are required to retain the user's trust. Recent QA tracks in the TREC evaluations have included questions that have no answers in the underlying data collection, forcing systems to 'know' that they are not certain of an answer (Voorhees, 2003). Other approaches include providing additional context so users can make their own judgments regarding the reliability of the answer's source (Lin et al., 2003), associating *trust values* to source documents and using these to calculate trust values for answers based on them (Zaihrayeu et al., 2005), or explaining how the answer was derived (Moldovan et al., 2003).

In this work, we explore yet another aspect of coping with uncertainty in QA systems (as a matter of fact, none of the approaches mentioned are mutually exclusive). It was carried out in the context of the IMIX project, which aims at building a multimodal QA system capable of answering questions in the medical domain, especially about Repetitive Strain Injury (RSI) (Boves and den Os, 2005; Theune et al., 2007). The IMIX demonstrator produces multimodal output in the form of text and pictures, as well as speech output and facial animation. The latter relies on the Nextens speech synthesizer for Dutch in cooperation with the RUTH talking head (DeCarlo

[*] This work was carried out within the IMIX-IMOGEN (Interactive Multimodal Output Generation) project, sponsored by the Netherlands Organization of Scientific Research (NWO). The authors like to thank Emiel Krahmer, Marc Swerts and colleagues for kindly permitting us to summarize their unpublished work on the preferred modality to express uncertainty in Section 2.2. We also like to acknowledge the useful comments and suggestions of three anonymous reviewers on our initial abstract.

and Stone, 2003; DeCarlo et al., 2004). The system incorporates multiple QA engines, some of which are capable of attaching confidence levels to their answers, albeit not always reliably. We are interested in the best way to convey uncertainty in the context of such a multimodal QA system, which offers the opportunity to exploit other communication channels besides text. In particular, the question addressed in this work is whether we can express uncertainty by means of talking head.

The remainder of the paper is organized as follows. In Section 2, we elaborate on the background and context of this work. We argue – on the basis of evidence from human-human dialogue studies – that uncertainty is better expressed by visual means than by text only. We summarize an unpublished study on audiovisual expression of uncertainty in the context of QA systems. We also discuss related work on *trust*. Section 3 reports on an experiment to test whether we can reliably express certainty or uncertainty by means of a limited repertoire of animated facial expressions, in particular, only combinations of eyebrow movements and head movements were considered. The results are in principle positive, but a number of remaining problems are discussed. In the final Section we summarize our findings and finish with a general discussion of open issues that need to be addressed before we can actually apply this approach in a multimodal QA system.

2 Background

2.1 Uncertainty in Human-Human Dialogue

In human-human information seeking dialogue, the information exchange is usually not limited to facts, but includes all sorts of additional meta-information. This kind of meta-information is often expressed by non-verbal means such as speech prosody, facial expression or gesture (e.g. Burgoon, 1994). One important example of this is the level of confidence or certainty associated with a particular piece of information. A number of researchers have used the *Feeling of Knowing* (FOK) paradigm (Hart, 1965) to study production and perception of uncertainty in human question answering. Smith and Clark (1993) found that speakers signal uncertainty regarding the correctness of their answer by means of prosodic cues such as filled pauses, increased delays and rising intonation. Subsequently, Brennan and Williams (1995) showed that listeners use these prosodic cues to estimate the level of certainty of a speaker's answer, suggesting that Smith and Clark's uncertainty cues do indeed have communicative relevance. Recent work by Swerts, Krahmer and colleagues has extended this line of research to audio-visual prosody, and in particular facial cues to uncertainty (Swerts et al., 2003; Krahmer and Swerts, 2005; Swerts and Krahmer, 2005). They found that in addition to the auditory cues, there are a number of facial cues that speakers produce to signal their uncertainty about an answer, and that those same signals are perceived by listeners in order to reliably detect the level of certainty associated with answers. Furthermore, detecting uncertainty turned out to be easier with bimodal presentation (i.e. both speech and face) in comparison with unimodal presentation. It is suggested that these findings have potential for improving human-computer interaction.

2.2 Preference for Visual versus Linguistic Cues

In recent, hitherto unpublished work, Krahmer et al. studied the expression of uncertainty in the context of a QA system. Since to the best of our knowledge no other published work addresses this topic, we will summarize their work here. The main questions were whether users appreciate it at all when a QA system signals its level of confidence regarding the answer, and whether users prefer signaling by either linguistic or visual means. In an experiment subjects were shown screenshots of a fancy-looking – but non-existent – medical QA system ("MediQuest TM"), each one containing both a question and an answer. The questions (e.g., "What is anesthesia?") were intentionally not that hard, so subjects were expected to recognize correct answers ("The process of blocking the perception of pain and other sensations."). Of the 20 answers presented, 13 were in fact correct and 7 were incorrect. The 75 subjects were equally divided in three groups, the first of which received no signaling of uncertainty at all, the second received signaling by linguistic cues, and the third by visual cues. Signaling uncertainty by linguistic cues comprised the use of modal expressions (e.g. “I think it is the process of blocking the perception of pain and other sensations.”). For visual signaling of uncertainty, the equivalent of a thermometer was used to express the degree of certainty. The majority of the correct answers (11 out of 13) were signaled as *certain*, whereas the majority of the incorrect answers (5 out of 7) were signaled as *uncertain.*

Subjects were asked to judge

1. the *formulation* of the answer,
2. the *adequacy* of confidence signaling, and
3. overall *quality* of the answer

on a 7-point scale. The results showed that answers containing linguistic signaling of uncertainty scored significantly worse on *formulation* than their certain counterparts. No such effect was found in case of visual signaling of uncertainty. The ratings on *adequacy* showed a strong negative effect in case of an inconsistent visual signal, i.e. thermometer indicating low confidence for a correct answer or thermometer indicating high confidence for an incorrect answer. This negative effect was much smaller in case of inconsistent linguistic cues. The overall *quality* scores also showed that answers with linguistic cues for uncertainty were judged significantly worse than their counterparts with visual signaling of uncertainty.

Although the choice of domain in combination with the sometimes less subtle linguistic expression of uncertainty might have affected the results to a certain extent, a likely interpretation of these findings is that subjects disliked linguistic signaling of uncertainty and preferred visual cues instead.

2.3 Trust

A QA system that is able to indicate confidence levels for its answers is arguably perceived as more trustworthy than a system lacking this capability. In that sense, expressing uncertainty is related to trust. Interestingly, several studies suggest that audiovisual communication enhances trust in comparison with text-only communication. Riegelsberger et al. (2005) showed that humans tend to have a media bias towards audio and video advice rather than text-only advice while seeking expert advice. Work by Cassell et al, (e.g. Cassell and Bickmore, 2000)), points out that believable Embodied Conversational Agents are an important factor in building trust relations between humans and computers.

3 Experiment

3.1 Design

The goal of the experiment was to test whether we can reliably express certainty or uncertainty by means of a limited repertoire of animated facial expressions. Only combinations of eyebrow movements and head movements were considered. The experiment was designed to test three hypotheses:

1. humans notice a difference between certain and uncertain animated facial expressions;
2. humans correctly recognize animated facial expressions as certain or uncertain;
3. humans are more sensitive to eyebrow movements than to head movements as a cue for certainty.

The first hypothesis states that the difference between animations intended as certain or uncertain is at least perceivable, whereas the second hypothesis states that certain and uncertain animations are recognized as intended.

Animations with either certain or uncertain facial expressions were produced by means of three different combinations of cues: (1) primarily eyebrow movements; (2) primarily head movements;

(3) both eyebrow and head movements. This amounts to six different conditions. To minimize the effect of semantics and prosody, these conditions were tested with ten different sentences.

Animations were presented to human judges with the question *How certain do you think the speaker is of the provided answer?*. Judgments were recorded on a 5 point scale, ranging from *uncertain* (1) to *certain* (5).

3.2 Material

The text material consisted of ten question-answer pairs from the domain of Repetitive Strain Injury (RSI); see Table 3 for two examples. This choice was motivated by the desire to apply our findings in future versions of the IMIX demonstrator system. The questions were taken from the list of target questions occurring in the functional specifications of the first version of the IMIX system. The answers are full sentences which were manually extracted from the shared text material as available to the IMIX QA systems. Answers are always correct, but in some cases the formulation is suboptimal given that the original context is removed. The answers are nevertheless typical for real output of a multimodal QA system.

As our talking head, we used RUTH (Rutgers University Talking Head), a freely available cross-platform real-time facial animation system (DeCarlo and Stone, 2003; DeCarlo et al., 2004). RUTH allows one to markup text with synchronized annotations for intonation and facial movements, including eyebrow and head movements, eye blinks and smiles. It relies on the Festival text-to-speech system (Black et al., 2002) to produce the speech. We ported RUTH to Dutch, using the Festival-based Nextens TTS system to produce Dutch speech.[1]

Answers were first annotated for intonation. The original English version of RUTH relies on the ToBI (Tone and Break Indices) system, the de facto standard for annotating American-English intonation. However, as Dutch intonation is significantly different, we used the equivalent system for annotating Dutch intonation (Gussenhoven, 2005), known as ToDI (Transcription of Dutch Intonation), which is supported by the Nextens TTS system for Dutch. Two examples are given in Table 3. One of the main differences is that there is no notion of *phrasal tone* or *intermediary phrase* in ToDI; there are only pitch accents and intonational phrases. Suitable locations for pitch accents and intonational phrase boundaries were determined by the first author (who has significant experience with annotation and prediction of Dutch intonation). Non-final intonational phrases start with a low initial boundary tone (%L) and end in a high final boundary tone (H%), whereas final phrases also end in a low tone (L%). All pitch accents are realized as H*L; subsequent pitch accents within an intonational phrase are downstepped (!H*L). This annotation results in arguably the most default and unmarked pitch contour in Dutch.

Next, answers were annotated for facial expressions, which in our case was limited to the commands for eyebrow and head movements as presented in Table 1. These movements come in two types. *Batons* highlight a single word and are indicated by a final star symbol. For example, *4** signals a frown associated with a single word. *Underliners* accompany several successive words. Following the convention for intonational phrases, we use an initial and final percent symbol to signal the start and end of an underliner respectively. For instance, *%4* followed by *4%* signals a frown stretching over several words; cf. the examples in Table 3. Figure 1 provides some illustrations of RUTH head movements. For details on how these abstract specifications are realized as facial expressions in RUTH, see DeCarlo and Stone (2003).

In order to create (un)certain animations we adhered to the guidelines in Table 2 as derived from the literature (Chovil, 1991a,b; Poggi, 2002; McClave, 2000; Swerts et al., 2003; Krahmer and Swerts, 2005; Swerts and Krahmer, 2005). The notion of *new information* was in practice considered as information not previously mentioned in the question. Evidently, there is a substantial gap between these global trends and the detailed specifications required by RUTH, in particular with respect the number and alignment of movements. Our annotations are therefore to a certain extent the result of what looked right and natural to the authors within the limits of the above guidelines.

[1] `http://nextens.uvt.nl`

Value:	Effect:
1+2	raises brows
4	frowns
D	nods downward
U	nods upward
F	brings the whole head forward
B	brings the whole head backwards
L	turns to model's left
R	turns to model's right
J	tilts the whole head clockwise
C	tilts the whole head counterclockwise
DR	nods downward with some rightward motion
UR	nods upwards with some rightward motion
DL	nods downward with some leftward motion
UL	nods upwards with some leftward motion
TL	tilts clockwise with downward nodding
CL	tilts counterclockwise with downward nodding

Table 1: RUTH commands for controlling eyebrow and head movements

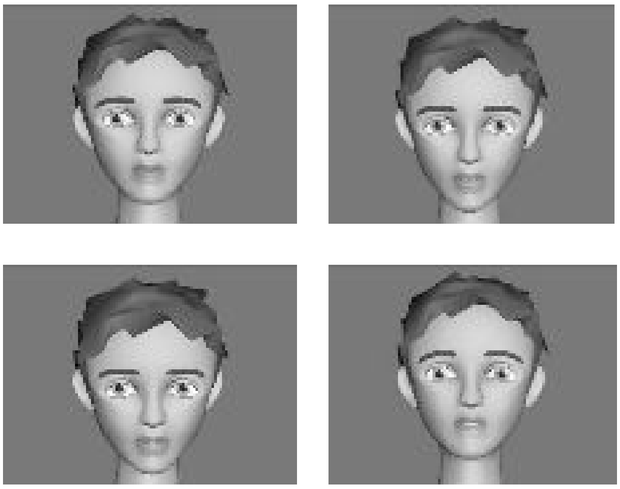

Figure 1: Illustration of several of RUTH's head movements: neutral (top left), downward nod (top right), downward nod with some leftward movement (bottom left), and upward nod (bottom left)

	Eyebrows:	Head:
Certain:	– few movements – frown with new information	– few movements – nodding with new information
Uncertain:	– many (unnecessary) movements – raising eyebrows with new information	– many (unnecessary) movements – sideward movement (shaking) with new information

Table 2: Guidelines for expressing (un)certainty through eyebrow and head movement

A related issue is that we found that animations lacking any eyebrow or head movements are almost as strange and artificial as animations without lip and jaw movements. We therefore avoided creating animations with only eyebrow movements or only head movements. Instead, all animations have at least some eyebrow and head movements, roughly corresponding to what are called *conversational facial signals* in DeCarlo et al. (2004). We used the following rules of thumb:

- Movements frequently occur with focused information – which is accented as well – and less frequently with unfocused information – which is unlikely to carry pitch accent.
- Syntactic connectives (e.g. *and, or, because*) may trigger movement, in particular when they are contrastive (e.g. *however, but, on the other hand*).
- Elements of a list may be indicated by sideward movement of the head, alternating leftward and rightward movements.
- Punctuation symbols like comma's and colons are often accompanied by a slow movement; periods often trigger a frown and/or nod; questions marks are associated with upward movement of the head and raising of the eyebrows.

The resulting RUTH animations were checked by the authors. Animations that were for some reason unnatural (e.g. suboptimal synchronization between speech and movements) were adapted. Pronunciation errors were fixed by adding words to the user lexicon.

Finally, the animations were saved as sequences of TIFF image files. The aligned synthetic speech was saved as an audio file and converted to MP3 format. Next, Adobe Premiere video editing software was used to convert images and sound to an AVI movie compressed with a standard MS Windows codec.

Question 1:	Words:	Wat	is	RSI?								
	Gloss:	what	is	RSI								
Answer 1:	Words:	RSI	is	een	**beroepsziekte**	**bij**	**mensen**	**die**	**steeds**	**dezelfde**	**beweging**	**uitvoeren**
	Gloss:	RSI	is	a	professional-disorder	of	people	who	repeatedly	the-same	movement	perform
Intonation:	Accents:				H*L					!H*L	!H*L	
	Boundaries:	%L										L%
Brows uncertain:	Brows:	1+2*			1+2*				1+2*	%1+2		1+2%
	Head:	DR*		%DL	DL%		U*		J*		%U	U%
Brows certain:	Brows:			%4	4%				4*		%4	4%
	Head:	DR*		%DL	DL%		U*		J*		%U	U%
Head uncertain:	Brows:				1+2*					%1+2		1+2%
	Head:	TL*			%TR		TR%		%L	L%		R*
Head certain:	Brows:				1+2*					%1+2		1+2%
	Head:	B*		%DL	DL%		D*		%DR			DR%
Brows & head uncertain:	Brows:	1+2*			1+2*				1+2*	%1+2		1+2%
	Head:	TL*			%TR		TR%		%L	L%		R*
Brows & head certain:	Brows:			%4	4%				4*		%4	4%
	Head:	B*		%DL	DL%		D*		%DR			DR%

Question 2:	Words:	Wie	kan	RSI	krijgen?							
	Gloss:	who	can	RSI	get							
Answer 2:	Words:	Het	is	bekend	dat	RSI	**vaker**	**voorkomt**	**bij**	**vrouwen**	**en**	**jongeren**
	Gloss:	it	is	known	that	RSI	more-often	occurs	with	women	and	youths
Intonation:	Accents:			H*L			H*L			!H*L		!H*L
	Boundaries:	%L		H%	%L							L%
Brows uncertain:	Brows:			1+2*			%1+2	1+2%		1+2*		1+2*
	Head:	DR*			D*		%TL	TL%			TR*	
Brows certain:	Brows:						%4	4%		%4		4%
	Head:	DR*			D*		%TL	TL%			TR*	
Head uncertain:	Brows:			%4	4%		%1+2	1+2%			1+2*	
	Head:	DR*					U*			L*		R*
Head certain:	Brows:			%4	4%		%1+2	1+2%			1+2*	
	Head:			D*			U*			%D		D%
Brows & head uncertain:	Brows:			1+2*			%1+2	1+2%		1+2*		1+2*
	Head:	DR*					U*			L*		R*
Brows & head certain:	Brows:						%4	4%		%4		4%
	Head			D*			U*			%D		D%

Table 3: Specifications for two of the stimuli (see text for explanation)

3.3 Procedure

A pilot experiment made clear that presenting 60 animations to a single subject takes too much time and is not feasible. The material was therefore split into six different parts, each part presenting all conditions for two different sentences.

The experiment was presented as a sequence of web pages and ran through the internet, allowing subjects to use a standard computer with a broadband internet connection and a current web browser. Subjects were automatically assigned to one of the six parts of the experiment. The introduction page explained the purpose and procedure of the experiment, and asked for some personal information (gender, age, etc.). Another page played a test animation to check that sound and image were correctly received.

Next, the stimuli were presented in random order, each one on a separate web page containing four elements. At the top of the page, there was an embedded movie player for rendering the RUTH animation. Subjects could replay this animation as many times a they liked. Below it was a plain text version of the answer to make sure that subjects understood the answer, even in case the speech synthesis was imperfect. We decided not to show the original question to prevent subjects from focusing to much on the factual content instead of on the visual presentation. At the bottom of the page was a 5 point scale (in the form of radio buttons), ranging from *sure* to *unsure*, through which subjects could respond to the question *How certain do you think the speaker is of the provided answer?*. Finally, there was a button for going to the next page. Returning to previous pages was impossible.

The closing pages offered space to provide general comments, and thanked subjects for their time.

3.4 Subjects

The online experiment was visited by 77 people, of which 58 completed a valid run. To keep the number of participants per part evenly balanced, only 50 results were used for analysis. Subjects' age ranged from 20 to 70 years old ($x = 30.6$, $SD = 11.0$); 31 were male and 19 were female. All were native speakers of Dutch without hearing impairments.

3.5 Results

The results are summarized in Table 4. Testing deviation from the expected mean (middle of the scale, i.e. 3) with a one-tailed t-test revealed that the average score on certain animations (3.63) is significantly different from the expected mean score ($p < 0.001$). This is not the case for the average score on the uncertain animations (2.85). The difference between the two scores (0.78) is again significant ($p < 0.001$). These findings confirm that overall the difference between certain and uncertain animations is at least noticeable, and that overall certain animations are recognized as intended.

Looking at nonverbal cues, we can observe that both eyebrow and head movements on their own, as well as the combination of the two, are sufficient to signal certainty (all $p < 0.001$). As far as uncertainty is concerned, however, only head movements ($p < 0.025$) and combined movements ($p < 0.01$) are close to significance. The effect of eyebrow movements is in fact opposite to the one intended. That is, eyebrow movements intended to signal *uncertainty* are actually perceived as signaling *certainty.* Thus contrary to our initial hypothesis, humans appear to be more sensitive to head movements than to eyebrow movements as far as the perception of uncertainty is concerned.

3.6 Discussion

Given the often subtle differences between the stimuli, we did not expect the differences to be significant (if noticed at all), so we think this is a rather promising result. Still, there are several issues that deserve discussion.

To begin with, we can think of alternative explanations of these results. One simple hypothesis is that more movement is perceived as more less certain, and conversely, less movement as more certain. This is not directly compatible with our results however. If we compare the total number

Cue:	Certain				Uncertain:			
	n:	av:	SD:	p <:	n:	av:	SD:	p <:
Eyebrow movements	10	3.49	0.73	.0001	10	3.26	0.82	.05
Head movements	10	3.54	0.81	.0001	10	2.65	0.95	.025
Eyebrow & head movements	10	3.91	0.77	.0001	10	2.62	0.94	.01
Overall:	30	3.63	0.58	.0001	30	2.85	0.64	n.s.

Table 4: Average scores of perceived certainty on a five point scale (uncertain=1, certain=5) over all subjects (N=50), split according to non-verbal cues used and animation's intended meaning (certain vs. uncertain); p-scores indicate significant difference from the expected mean score (3) according to a one-tailed t-test

of head movements – both batons and underliners – in the *uncertain* animations (55) to the total number of head movements in the *certain* animations (43), the difference is relatively small (12), but nevertheless sufficient to be perceived as significantly different. In contrast, the difference between the total number of eyebrow movements in *uncertain* animations (46) versus in *certain* animations (29) is slightly larger (17), yet insufficient to cause a similar significant difference in perception.

Perhaps then eyebrow movements are irrelevant for expressing uncertainty, and the results depend solely on head movements. This would explain the outlier in the case of uncertainty expressed by eyebrow movements, and is also compatible with the fact that there is hardly any difference between uncertainty expressed by head movements only versus by both head and eyebrow movements. On the other hand, it contradicts the findings in the case of certainty, where certainty expressed by eyebrows was found to be effective, and even more so in combination with head movements. To sum up, there seem to be no straightforward alternative hypotheses.

With hindsight, the experimental setup has a number of weaknesses that should be properly addressed in future work. One of these is the simplifying assumption that the expected mean score is equal to the mid of the scale (3). However, answers may be inherently more certain or uncertain because of their semantic content. This inherent bias can be measured by running a separate experiment in which subjects are asked to rate certainty on the basis of the text only. This bias can then be taken into account during analysis and statistical testing.

Another issue is that the question *How certain do you think the speaker is of the provided answer?* severely constraints the range of responses. Without this strong bias, subjects might prefer to interpret the facial expressions along other, unintended dimensions such as *surprise* or *agitation*, rather than *certainty*. One possible method to reduce this bias is to ask subjects to score on other scales besides the one for certainty.

We found there is a tension between RUTH's (theoretical) requirement that batons should be time aligned with accented words and that of a natural rendering of facial movements. Our animations frequently had batons at unaccented words. Moreover, the recommendation that underliners should be aligned with the phrasal tones of intermediary phrases is even impossible in Dutch, as there is no such thing in descriptions of Dutch intonation. This suggests more research is needed on the topic of alignment between intonational and facial movements.

In order to keep the experiment manageable, we limited ourselves to eyebrow and head movement. However, RUTH supports at least two other movements: smiles and blinks. It would be interesting to run a similar experiment using these cues. At the same time, the repertoire of current talking heads is much more constrained than that of real humans. For instance, Swerts and Krahmer (2005) mention a complex expression they labeled *funny face*, which their subjects often used to express uncertainty.

4 General Discussion and Conclusion

In order to retain a user's trust, QA systems need to express the level of uncertainty attached to their answers. Multimodal QA systems offer the opportunity to express uncertainty through other than verbal means. On the basis of evidence from studies how uncertainty is expressed in human-human dialogue, it was argued that uncertainty is better expressed by audiovisual than by verbal means. Moreover, we summarized (unpublished) work on visual expression of uncertainty in the context of QA systems suggesting that humans dislike linguistic signaling of uncertainty and prefer visual signaling instead. Circumstantial evidence comes from general work on trust and ECA's.

An experiment was described to test whether we can reliably express certainty or uncertainty by means of a limited repertoire of animated facial expressions, in particular, only combinations of eyebrow movements and head movements. The results suggest that humans can correctly recognize animated facial expressions as certain, but that only head movements are a consistent cue. We discussed a number of issues with the experimental setup which preclude definite conclusions.

In addition, there are a number of open issues that need to be resolved before a talking head like RUTH can be employed for signaling uncertainty in multimodal human-computer interaction. If we take the IMIX multimodal QA system as a case in point, it is assumed that its QA engines can provide reliable confidence scores. In practice, however, it turns out that it is hard for a system to know that is does not know the answer, let alone how certain it is of a particular answer. Future development in QA is likely to improve this (Burger et al., 2003).

It should also be noted that our results only concern two extremes, i.e. certainty versus uncertainty. In a practical system, a more likely setting is to express a *degree* of certainty. Our results in part suggest that a combination of cues gives a stronger effect, but more research is definitely required.

Another open issue is how to obtain the specifications for facial expressions. So far our annotations were produced manually, but a dialogue system should of course be able to predict them automatically. For some limited domains the use of templates may be sufficient, but in QA systems like the IMIX system, where text variation is unpredictable, such an approach is unlikely to succeed. Given the similarity to the problem of predicting prosodic markup in speech synthesis, and the successful application of machine learning techniques in that area (e.g. Marsi et al., 2003), a data-driven approach seems most promising. For training and evaluation purposes, we would then need a substantial corpus of annotated examples – of either human speakers or ECA's – and select informative (linguistic) features. One of our own topics for future research is data-driven prediction of annotations to appropriately express uncertainty.

References

Black, A. W., Taylor, P., and Caley, R. (2002). *The Festival Speech Synthesis System, System documentation*. Centre for Speech Technology Research University of Edinburgh.

Boves, L. and den Os, E. (2005). Interactivity and multimodality in the IMIX demonstrator. In *International Conference on Multimedia and Expo*, pages 1578–1581.

Brennan, S. 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(3):383–398.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C., Maiorano, S., Miller, G., et al. (2003). Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). Technical report.

Burgoon, J. (1994). Nonverbal signals. In Knapp, M. L. and Miller, G. R., editors, *Handbook of Interpersonal Communication*, volume 2, pages 229–285. Sage.

Cassell, J. and Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of ACM*, 43(12):50–56.

Chovil, N. (1991a). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194.

Chovil, N. (1991b). Social determinants of facial displays. *Journal of Nonverbal Behaviour*, 15(3):141–154.

DeCarlo, D. and Stone, M. (2003). The Rutgers University Talking Head: RUTH. Technical report, Rutgers University.

DeCarlo, D., Stone, M., Revilla, C., and Venditti, J. (2004). Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38.

Gussenhoven, C. (2005). Transcription of Dutch intonation. In Jun, S.-A., editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, Oxford.

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

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

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. (2003). What makes a good answer? The role of context in question answering. In *Human-Computer Interaction (INTERACT 2003)*.

Marsi, E., Busser, G., Daelemans, W., Hoste, V., Reynaert, M., and van den Bosch, A. (2003). Learning to predict pitch accents and prosodic boundaries in Dutch. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 489–496, Sapporo, Japan.

McClave, E. (2000). Linguistic functions of head movement in the context of speech. *Journal of Pragmatics*, 32:855–878.

Moldovan, D., Clark, C., and Harabagiu, S. (2003). COGEX: a logic prover for question answering. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 87–93. Association for Computational Linguistics Morristown, NJ, USA.

Poggi, I. (2002). Towards the alphabet and lexicon of gesture, gaze and touch. In Bouissac, P., editor, *Virtual Symposium on Multimodality of Human Communication.* http://www.semioticon.com/virtuals/index.html.

Riegelsberger, J., Sasse, M., and McCarthy, J. (2005). Do people trust their eyes more than their ears?: Media bias in detecting cues of expertise. In *Conference on Human Factors in Computing Systems*, pages 1745–1748. ACM Press New York, NY, USA.

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

Swerts, M. and Krahmer, E. (2005). Audiovisual prosody and Feeling of Knowing. *Journal of Memory and Language*, 53(1):81–94.

Swerts, M., Krahmer, E., Barkhuysen, P., and van de Laar, L. (2003). Audiovisual cues to uncertainty. In *Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 14–3, Chateau-D'Oex, Switzerland.

Theune, M., van Schooten, B., op den Akker, R., Bosma, W., Hofs, D., Nijholt, A., Krahmer, E., van Hooijdonk, C., and Marsi, E. (to appear 2007). Questions, pictures, answers: Introducing pictures in question-answering systems. In *Proceedings of the Tenth International Symposium on Social Communication*, Santiago de Cuba, Cuba.

Voorhees, E. (2003). Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC2003).*

Zaihrayeu, I., da Silva, P., and McGuinness, D. (2005). IWTrust: Improving User Trust in Answers from the Web. In *Proceedings of 3rd International Conference on Trust Management (iTrust2005)*, Rocquencourt, France. Springer.