

# Audiovisual cues to uncertainty

Swerts, M.G.J.; Krahmer, E.J.; Barkhuysen, P.; van de Laar, L.

*Published in:*
Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems

*Publication date:*
2003

# Audiovisual cues to uncertainty

*Marc Swerts, Emiel Krahmer, Pashiera Barkhuysen and Lennard van de Laar*

Tilburg University, Communication & Cognition, The Netherlands
{m.g.j.swerts,e.j.krahmer,p.n.barkhuysen,l.v.d.laar}@uvt.nl

## Abstract

This paper presents research on the use of audiovisual prosody to signal a speaker's level of uncertainty. The first study consists of an experiment, in which subjects are asked factual questions in a conversational setting, while they are being filmed. Statistical analyses bring to light that the speakers' Feeling-of-Knowing (FOK) correlate significantly with a number of visual and verbal properties. Interestingly, it appears that answers tend to have a higher number of marked feature settings (i.e., divergences of the neutral audiovisual expression) when the FOK score is low, while the reverse is true for non-answers. The second study is a perception experiment, in which a selection of the utterances from the first study is presented to subjects in one of three conditions: vision only, sound only or vision+sound. Results reveal that human observers can reliably distinguish HighFOK responses from LowFOK responses in all three conditions, be it that answers are easier than non-answers, and that a bimodal presentation of the stimuli is easier than their unimodal counterparts. Results of these two experiments are potentially relevant for improving the communication style in human-machine interaction.

## 1. Introduction

Uncertainty is an inherent element of human-machine communication, in that both the user and the system can be unsure whether the other understood them correctly. For instance, imagine a user who poses a question about the inventor of the telephone to a spoken question-answering (QA) system, which uses the web as its source of answers. For different reasons, the system may have difficulties to give an appropriate answer. First of all, given the status of current Automatic Speech Recognition (ASR) modules, the system may have problems to correctly decode the question (was it 'telephone' or 'telegraph'?) (see Hirschberg et al. 1999). But even if it has correctly understood the user input, the system may still not be sure about the correct answer (see Buchholz and Daelemans (2002) for discussion). For instance, besides Bell, you may get other names from the WWW as potential telephone inventors, like Reis, Bourseul, or Gray. Given such different sources of communication problems, it would seem user-friendly if a system would, besides giving an answer, also express the level of confidence it has in the information it returns. Indeed, a user's acceptance of incorrect system output might be higher if the system made it clear in its self-presentation that it is not sure about the correctness of its answer.

However, spoken dialogue systems are usually not cooperative in this sense. Yet, there is one kind of user interface that lends itself ideally for signaling level of uncertainty, and which has become increasingly more popular as a medium between the human and the machine, namely embodied conversational agents (ECA's). (Cassell et al. 2000; Gustafson et al. 1999). ECA's can be viewed as specific software components in an interface that appear to users as animated characters (e.g. a talking head or a complete figure). In principle, many such ECA's have ideal expressive devices to package the information they transmit to the user, as they can rely on various prosodic cues. In a broad sense, prosody can be defined as the whole gamut of features that do not determine *what* people are saying, but rather *how* they are saying it. Originally, the term was used to strictly refer to verbal prosody, i.e., the set of suprasegmental features such as intonation (speech melody), rhythm, tempo, loudness, voice quality and pausing that are encoded in the speech signal itself (Cruttenden, 1986; Ladd 1996). More recently, various researchers tend to broaden its definition to also include visual prosody, i.e., specific forms of body language that communication partners send to each other during the interaction, such as facial expressions, arm and body gestures and pointing (Cassell et al. 2000). Many studies have shown that such (audiovisual) prosody can provide utterances with 'extra' information that is often not explicitly contained in the lexical and syntactic make-up of a sentence. Therefore, the current paper investigates to what extent it is also used for signaling (degree of) uncertainty in spoken interactions.

In order to gain insight into the expression of uncertainty by means of audiovisual prosody, we have opted to explore some human interactions, with the goal of implementing the results of these analyses in ECA's. To this end, the study presented below, consists of two parts, in which the research problem is tackled from both a production and a perception perspective. The motivation for doing so is that we believe that prosodic phenomena can only be convincingly shown to have functional validity if they can be proven to be relevant for both traditional participants in the communication chain, the speaker and the listener, whereas previous studies have often been limited to a purely speaker-oriented approach to prosodic function. In particular, we elicit natural speech data in such a way that they become useful as stimuli for perceptual evaluation. As will become clear below, our work largely builds on previous studies on the so-called Feeling-of-Knowing (FOK), which term refers to people's ability to assess and monitor their own knowledge (Hart, 1965). Whereas such earlier work has largely concentrated on lexical and speech cues, we also included the study of visual cues as part of our analyses.

## 2. Speakers' expression of uncertainty

### 2.1. Background

Smith and Clark (1993) explored how speakers handle problems of self-presentation when they are asked to respond to questions. They argue that there are at least two reasons why speakers may want to express their level of certainty about a given response to a question. First, they should express their uncertainty in order to avoid implying that they *are* certain. This follows

Figure 1: *Two stills taken at the end of the word 'Shakespeare' in two different realizations: on the left is a response where the speaker has a High FOK, on the right where he has a Low FOK.*

from the Gricean cooperativeness principle of Quality ("Do not state what you believe to be false"). Second, they should also indicate uncertainty to save face; if their answer turns out to be incorrect, they will look less foolish if they indicated little confidence in the response. Their experiment consisted of three parts. First, subjects were each asked 40 factual questions in a conversational setting. Then, they had to rate for each question their feeling that they would recognize the correct answer (their feeling-of-knowing (FOK)), and finally they had to do a recognition test, i.e., a multiple-choice test in which all the original questions were again presented. It was found that there was a significant difference between two types of responses, answers and non-answers ("I don't know"): the lower their FOK, the slower the answers, but the faster the non-answers. In addition, the lower the FOK, the more often people answered with rising intonation, added filled pauses such as "uh" and "uhm", and other face-saving comments. There also appeared to exist a significant difference in usage of filled pauses, in that "uh" was more often used to signal brief delays.

The different results regarding occurrence of filled pause and delay thus suggest that the correlates of LowFOK and HighFOK responses may be quite different for answers and nonanswers. LowFOK answers and HighFOK nonanswers appear to be alike in that they show similar correlates of a mental search procedure. Therefore, it is interesting to investigate whether these two cases share other properties as well. Besides additional verbal features, we are particularly interested in potential visual cues as well. Goodwin and Goodwin (1986) (see als Clark, 1996) already discussed the communicative relevance of the so-called "thinking face": it often happens that a respondent turns away from the addressee with a distant look in his eyes in a stereotyped facial gesture of someone thinking hard. Speakers appear to use the thinking face to signal that they are doing a word search and to account for why they aren't proceeding with their utterance. In addition, there have been some studies on prosodic correlates of negative feedback cues in spoken interactions (e.g. Krahmer et al. 2001; Granström et al. 2002), which may be similar to indicators of uncertainty, as they also follow problems of understanding.

### 2.2. Approach

#### *2.2.1. Data collection*

Following procedures outlined in Smith and Clark (1993), 20 subjects (11 male, 9 female), colleagues and students from Tilburg University, participated as speakers in the experiment on a voluntary basis. Subjects were unaware of the real purpose of the study, but were told that its goal was to learn more about the average complexity of a series of questions which could be used in future psycholinguistic research. They were warned beforehand that we expected that questions would vary in degree of difficulty. In order to encourage them to do their best and guess the correct answer in case of uncertainty, they were told that the winner of the game, the person with the highest number of correct answers, would get a small reward. The stimuli consisted of a series of factual questions of the Dutch version of the "Wechsler Adult Intelligence Scale" (NT-WAIS), an 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. The 40 questions in total covered a wide range of topics, like literature, sports, history etc. Subjects were presented with this list of questions in one of two random orders. Questions were posed by the experimentor whom the subjects could not see, and the responses by the subject were filmed (front view of head). The experimentor asked the series of questions one by one, and the pace of the experiment was determined by the subject. As an example, here are 5 responses (translated from Dutch) to the question about the name of the person who drew the pictures in "Jip en Janneke", a famous Dutch book for children:

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

| Experiment | Response | FOK |
|---|---|---|
| Open Question | All answers (n=704) | 6.32 |
| | Correct Answers (n=575) | 6.55 |
| | Incorrect Answers (n=129) | 5.29 |
| | All Nonanswers (n=96) | 3.03 |
| Multiple Choice | Correct Answers (n=717) | 6.17 |
| | Incorrect Answers (n=83) | 3.84 |

a. Fiep Westendorp
b. uh Fiep Westen-(short pause)-dorp
c. (short pause) Isn't that Annie M.G. Schmidt?
d. no idea
e. uh the writer is Fiep Westendorp, but the drawings? No, I don't know

The example shows cases of correct answers, which could be fluent (a) or hesitant (b), an incorrect answer (c), and a simple (d) and complex (e) case of a nonanswer.

After this test, the same sequence of questions was again presented to the same subjects, but now they had to express on a 7-point scale how sure they were that they would recognize the correct answer if they would have to find it in a multiple-choice test, with 7 meaning "definitely yes" and 1 "absolutely not". 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. The participants were instructed to answer every question, even if they had to guess.

#### *2.2.2. Labeling, annotation*

All utterances from the first test (800 in total) were transcribed orthographically and manually labelled regarding a number of auditive and visual features by four independent transcribers on the basis of an explicit labelling protocol, which included various double-checks. Regarding verbal cues, we labelled the presence or absence of the following features:

**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. Note that we did not attempt to isolate question intonation, as it turned out to be difficult to consistently differentiate 'real' question intonation from list intonation.

**Filled pause** Whether the utterance contained one or more filled pauses, or whether these were absent. We did not differentiate between 'uh', 'uhm' or 'mm'.

In addition to these categorical variables, we counted the number of words spoken in the utterance, where family names, like Elvis Presley, were considered to be one word. The reason for doing so, is that it was found earlier that utterances tended to be longer after communication problems (Krahmer et al. 2001).

As to the visual cues, we labelled the presence or absence of one of the following features:

**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.

Table 2: *Pearson correlation coefficients of FOK scores with number of words, gaze acts and marked features.*

| Correlations of | Response | |
|---|---|---|
| FOK scores with | Answers | Nonanswers |
| Words | –.343** | .401** |
| Gaze acts | –.309** | .347** |
| Marked features | –.422** | .462** |

$** p < .01$

**Low Gaze** Whether a speaker looked downward during the response.

**High Gaze** Whether a speaker looked upward during the response.

**Diverted Gaze** Whether a speaker looked away from the camera (to the left or the right) during the response

**Funny face** Whether the speaker produced a marked facial expression, which diverted from a neutral expression, during the response. A typical example is the right still in Figure 1.

In addition, we counted the number of different gaze acts, i.e, combinations of high, low or diverted gaze.

The labeling was divided among the collaborators of the project, whom each was given the task to individually label a particular set of features. The different gaze directions (low, high, diverted), which were more difficult and which consisted of combinations of eye and head movements, were annotated by consensus labeling of two annotators. All features were transcribed independently from the FOK scores in order to avoid circularity. Cues were only marked if they were clearly present, and only based on perceptual judgments.

### 2.3. Results

It appeared that the subjects found a majority of the questions very easy, as they gave a maximum FOK score of 7 to 61.1% of the questions, a score of 6 to 13.4% of the questions, and lower scores to the remaining 25.5%. In addition, it appeared that 71.9% of the questions of the first task were indeed answered correctly and 89.6% of the same list of questions in the multiple-choice test. Table 1 lists the average FOK scores as a function of Question Type (open question versus multiple choice), and the response categories (correct answers, incorrect answers, nonanswers). The table shows that there is a close correspondence between the FOK scores and the correctness or incorrectness of a response in both the open test and the multiple-choice.

Table 2 lists the correlation coefficients between the FOK scores and number of words, gaze acts and marked features, defined as the presence of features defined in section 2.2.2. It can be seen that there are negative correlation between the FOK scores and these variables for answers, and positive correlations for nonanswers. In other words, for answers, higher FOK scores correspond with a lower number of words, gaze acts and marked features, while an opposite relation holds for nonanswers. An analogous picture about opposite findings for answers and nonanswers emerges from tables 3 and 4, which give the average FOK scores for presence versus absence of audiovisual features for answers and nonanswers, respectively. Table 3 shows that the presence of a verbal or visual feature in answers always coincides with a significantly lower FOK score, whereas Table 4 shows that the presence of such a feature in non-answers

Table 3: *Average FOK scores for answers as a function of presence or absence of audiovisual features. Statistics are based on T-test analyses.*

| | Present (1) | Absent (2) | Diff. (1)-(2) |
|---|---|---|---|
| Filled pause | 6.50 | 5.79 | +0.71*** |
| Delay | 6.54 | 5.24 | +1.31*** |
| High Intonation | 6.43 | 6.08 | +0.35*** |
| Eyebrow | 6.46 | 5.74 | +0.72*** |
| Smile | 6.35 | 6.07 | +0.28* |
| Low gaze | 6.45 | 6.07 | +0.39*** |
| High gaze | 6.47 | 5.94 | +0.52*** |
| Diverted gaze | 6.64 | 5.98 | +0.66*** |
| Funny Face | 6.37 | 5.17 | +1.21** |

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

Table 4: *Average FOK scores for nonanswers as a function of presence or absence of a audiovisual features. Statistics are based on T-test analyses.*

| | Present (1) | Absent (2) | Diff. (1)-(2) |
|---|---|---|---|
| Filled pause | 2.64 | 5.00 | –2.36*** |
| Delay | 2.40 | 3.81 | –1.42*** |
| High Intonation | 2.85 | 4.00 | –1.15* |
| Eyebrow | 2.85 | 3.61 | –0.76 |
| Smile | 2.86 | 3.52 | –0.66 |
| Low gaze | 2.51 | 3.68 | –1.17** |
| High gaze | 2.41 | 3.92 | –1.51*** |
| Diverted gaze | 2.71 | 3.11 | –0.40 |
| Funny Face | 3.02 | 3.17 | –0.14 |

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

leads to higher FOK scores, be it that not all of the differences are significant, probably because of the limited number of data.

In order to learn more about the cue value of combinations of features, we also calculated, for answers and non-answers separately, the average FOK scores for responses that differ regarding the number of marked feature settings (minimum: 0, maximum: 7). The results of this are visualized in Figure 2, which again illustrates opposite trends for the two response categories: for answers, the average FOK score decreases with an increasing number of marked features, while the opposite is true for nonanswers.

### 2.4. Discussion

This first study has replicated some of the findings of the research by Smith and Clark (1993). It appears that our subjects' FOK scores correspond with their performance in the open question and multiple-choice test. In addition, particular audiovisual surface forms of the utterances produced by our speakers are indicative of the amount of confidence speakers have about the correctness of their response. For answers, lower scores correlate with occurrences of long delay, filled pause, question intonation, a number of gaze features, funny face and smile. In addition, speakers tend to use more words and more gaze acts, when they have a lower FOK. Interestingly, for nonanswers, the relationships between FOK scores and the different audiovisual features is the mirror image of the outcome with answers. In this way, the current outcome generalizes earlier findings of Smith and Clark that answers and nonanswers differ in speaker behaviour. The fact that the audiovisual properties of LowFOK answers resemble those of HighFOK nonanswers may be due to the fact that they reflect similar mental operations, in particular word search procedures of a speaker.

Obviously, the current study was limited in that we have not fully explored possible interactions between cues. This was not entirely possible, since one quickly runs into sparse data problems, as not every combination of features is well represented in the dataset. Yet, our intuition is that the combined use of particular features is very important. For instance, anecdotal evidence suggests that a person may smile for completely different reasons: because he is delighted that he knows the answer, or because he is ashamed that he does not know the answer to a seemingly easy question. Presumably, the difference between these two types of responses is reflected not only in the type of smile, but also in the combination with other feature settings. See the discussion of 'blends' by Ekman and Friesen (1978), where a face displays two different emotions.

## 3. Observers' perception of uncertainty

### 3.1. Background

The first study was a speaker-oriented approach to gain insight into audiovisual correlates of FOK. While our analyses revealed that there was a statistical relationship between the two, this in itself does not prove that the audiovisual properties also have communicative relevance. In order to prove this, we performed a perception study, for which we used earlier work by Brennan and Williams (1995) as our main source of inspiration. They did a study which can be seen as a follow-up study to the research by Smith and Clark (1993). More specifically, they examined whether listeners are sensitive to a speakers' display of their metacognitive state, i.e., whether the manner in which an utterance is produced leads the listener to conclude that a speaker is confident or tentative. After an experiment which replicated Smith and Clark's earlier findings, a selection of the speakers' responses was presented to listeners, who were tested on their Feeling-of-Another's-Knowing (FOAK) to see if metacognitive information was reliably conveyed by the surface form of responses. It appeared that there was again a significant difference for answers and non-answers: the results for the former category showed that rising intonation and longer latencies led to lower FOAK scores, whereas for nonanswers longer latencies led to higher FOAK scores. Given that the study by Brennan and Williams (1995) focused on auditory cues alone, the goal of our second study is to explore whether observers of the speakers' answers of our first study are able to guess these speakers' FOK scores on the basis of visual cues as well. In particular, we are interested in whether a bimodal presentation of stimuli leads to better FOK predictions than the unimodal components in isolation. While we expect that we get the best performance for bimodal stimuli, it is an interesting empirical question whether the auditory or the visual features from the unimodal stimuli are more informative for FOK predictions.

### 3.2. Method

#### *3.2.1. Selection of data*

From the original 800 responses, we selected 60 utterances, with an equal amount of answers and non-answers, and an even distribution of high and low FOK scores. Only the answer of a question-answer pair was presented to subjects. The selection was based on written transcriptions of the responses by someone who had not heard or seen the original responses. Given the individual differences in the use of the FOK scale, we chose to

Table 5: *Experimental design of perception experiment. Number of stimuli per condition.*

| | Answer | | Nonanswer | |
|---|---|---|---|---|
| | highfok | lowfok | highfok | lowfok |
| Vision | 15 | 15 | 15 | 15 |
| Sound | 15 | 15 | 15 | 15 |
| Vision+Sound | 15 | 15 | 15 | 15 |

use -per speaker- his/her two highest scores as instantiations of HighFOK scores and the two lowest as LowFOK scores. 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 experimentor should not appear again in the subjects' response.
2. All the answers should be lexically distinct, and should thus not occur twice. This criterion was not applied to the non-answers as they were very similar.
3. The responses should be maximally distributed across speakers. There should be maximally two answers and two non-answers per speaker.

Having applied this procedure on the basis of written transcriptions of the data, we finally replaced a couple of stimuli by others, if the background noise made them unsuitable for the perception experiment. The design of the experiment is visualized in Table 5.

### *3.2.2. Procedure*

The selected stimuli were presented to subjects in three different conditions as a group experiment: one third of the subjects saw the original videoclips as they were recorded (vision+sound), another third saw the same videoclips but then without the sound (vision), whereas the last third could only hear the utterances without seeing the image (sound). In all three conditions, stimuli were presented on a screen where they first saw the stimulus ID (1 through 30) and then the actual stimulus. In case of the sound-only stimuli they saw a black screen instead of the original videoclip. The motivation to present sound-only stimuli also visually, was to make sure that subjects could "see" the start of the utterance, in case there was a silent pause in the beginning of the utterance. The interstimulus interval was 3 seconds. Subjects were 120 native speakers of Dutch, students

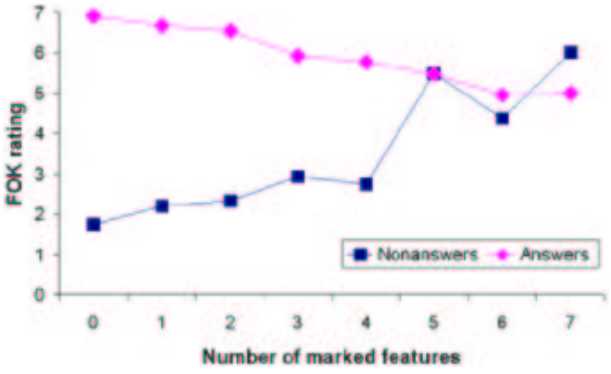


Figure 2: *Average FOK scores for answers and nonanswers as a function of the relative number of marked prosodic features.*

Table 6: *ANOVA results for perception experiment*

| Factor | Level | FOAK | F-stats |
|---|---|---|---|
| *Within Subjects* | | | |
| FOK | High | 4.792 | $F_{(1,117)}$=2229.886, |
| | Low | 2.646 | $p$ <.0001 |
| Response | Answer | 3.922 | $F_{(1,117)}$=90.477, |
| | Non-answer | 3.516 | $p$ <.0001 |
| *Between Subjects* | | | |
| Experiment | Vision | 3.779 | $F_{(2,117)}$=1.424, |
| | Sound | 3.669 | $p$ =.245 |
| | Vision+Sound | 3.709 | |

Table 7: *FOAK scores for HighFOK and LowFOK stimuli in different experimental conditions.*

| Condition | HighFOK (1) | LowFOK (2) | Diff. (1)-(2) |
|---|---|---|---|
| Vision | 4.434 | 2.903 | 1.531 |
| Sound | 4.890 | 2.668 | 2.222 |
| Vision+Sound | 5.052 | 2.367 | 2.685 |

from the University of Tilburg, none of whom had participated as speaker in the first test. Within a condition, subjects had to participate in two separate sessions, one with answers as stimuli and one with nonanswers. The question to the subjects about the answers was whether a person appeared very uncertain (1) or very certain (7) in his/her response. The question with the non-answer stimuli was whether subjects thought the person would recognize the correct answer in a multiple-choice test, with 1 meaning "definitely not" and 7 "definitely yes". Each part of the experiment was preceded by a short exercise session with 2 answers and 2 non-answers respectively to make subjects acquainted with the kinds of stimulus materials and the procedure.

## 3.3. Results

The subjects' responses were statistically tested with an analysis of variance with the FOAK scores as dependent variable, with original FOK scores and response type as within-subject factors, and condition (vision, sound, vision+sound) as between-subject factor.

Table 6 shows that there were significant effects on the subjects' FOAK scores of original FOK status of the utterance and of response category, while there was no main effect of the condition. However, there were significant 2-way interactions between FOK and condition ($F_{(2,117)}$=54.451, $p$ <.0001) and between response and condition ($F_{(1,117)}$=241.597, $p$ <.0001), and a significant 3-way interaction between FOK, condition and response ($F_{(2,117)}$=3.291, $p$ <.05). The 2-way interactions can easily be understood when we look at the average scores for combinations of FOK and condition and response and condition (see Tables 7 and 8, respectively). The first table shows that the difference in scores for low and high FOK scores is more extreme in the sound+vision condition, than in the unimodal conditions, meaning that the subjects' ratings were more accurate when subjects had access to both sound and vision. Notice that this explains why no main effect of experimental condition was found: the differences in FOAK scores between HighFOK and LowFOK stimuli change, while the overall FOAK averages stay the same (see Table 6). The second table shows that the difference between high and low FOK scores is -as expected- easier to perceive in answers than in non-answers.

Table 8: *FOAK scores for HighFOK and LowFOK stimuli for answers and nonanswers.*

| Response | HighFOK (1) | LowFOK (2) | Diff. (1)-(2) |
|---|---|---|---|
| Answer | 5.231 | 2.614 | 2.617 |
| Nonanswer | 4.353 | 2.678 | 1.675 |

### 3.4. Discussion

The results of the second perception test are consistent with the findings of the first analysis of speaker's expression of uncertainty. It appears that subjects are able to differentiate LowFOK responses from HighFOK responses in the unimodal experimental conditions, but they clearly performed most accurate in a bimodal condition. This suggests that the addition of visual information, which the aforementioned FOK and FOAK studies did not consider, is beneficial for detecting uncertainty. While we had seen that answers and nonanswers exhibit completely opposite audiovisual features, human subjects are able to adapt their judgments: they are able to tell the difference between Low and High FOK for both reponse categories, be it that the performance for nonanswers drops compared to answers, in line with previous observations of Brennan and Williams (1995). In conclusion, this study brought to light that the audiovisual features of our original utterances have communicative relevance as they can be interpreted by listeners as cues of a speakers' level of confidence.

One possible confounding factor in the perception study, however, is that we did not clearly distinguish stimuli with a considerable initial delay from utterances with a shorter delay. While such pause is, strictly speaking, a speech feature, it can obviously also be observed from the visual information alone, since one can see from a speakers' face whether or not a person is talking. Therefore, to see to what extent this factor has influenced the judgments, also of our vision-only stimuli, we intend to redo the perception test with stimuli whose initial pauses (if any) have been cut from the signal.

## 4. General discussion and perspectives

The current study has reported a functional investigation of audiovisual prosody: we showed that it can be used to signal a speakers' level of confidence in the answers he or she returns to an addressee. A perception study revealed that such audiovisual correlates of (un)certainty also have real cue value. This result is potentially relevant for improving ECA's, which, as stated in the introduction, become increasingly more popular as computer interfaces, like a virtual presenter who helps the user navigating through a website or who presents information through various media, such as computer graphics, non-speech audio, text and speech. To make these agents 'believable' and 'communicative', it is important to know in full detail how the specific auditive and visual parameters of such characters contribute to speech communication. Given that many spoken dialogue systems are relatively uncertain, especially with question-answering or information-giving systems that take speech as input, it is interesting to investigate whether the interaction between the human and the machine would improve if the system would convincingly express its level of confidence through an embodied conversational agent. Therefore, we are currently planning to redo our perception study, but now using a synthetic face in which some audiovisual cue combinations of our human subjects are implemented. As a first step, we intend to model the expression of our talking head on the basis of copy synthesis, and we may later try a rule-based implementation. It will be interesting to see whether we can replicate our earlier findings on level of uncertainty with the synthetic head.

## 5. Acknowledgements

This research was conducted as part of the VIDI-project "Functions of audiovisual prosody (FOAP)", sponsored by the Dutch NSF (NWO). Swerts is also affiliated with the Flemish Fund for Scientific Research (FWO-Flanders) and Antwerp University. Thanks to Judith Schrier (Antwerp) and Jorien Scholze (Tilburg University) for their help in carrying out the experiments.

## 6. 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, pp. 383-398.

[2] Buchholz, S. and Daelemans, W. (2001), "Complex answers: a case study using a WWW question answering system", Natural language engineering, 7, pp. 301-323

[3] Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. (Eds.) (2002). Embodied Conversational Agents. Cambridge: MIT Press.

[4] Clark, H.H. (1996). Using Language. Cambridge: Cambridge University press.

[5] Cruttenden, A.(1986), Intonation. Cambridge: Cambridge University press.

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

[7] Goodwin, M.H. and Goodwin, C. (1986), "Gesture and coparticipation in the activity of searching for a word", Semiotica, 62, pp. 51-75.

[8] Granström, B., House, D. and Swerts, M. (2002). "Multimodal feedback cues in human-machine interactions", Proc. Speech Prosody 2002, 11-13 April, Aix-en-Provence, pp. 347-350.

[9] Gustafson, J., Lindberg, N. and Lundeberg, M. (1999). "The August spoken dialogue system", Proc. of Eurospeech'99, Budapest, pp. 1151-1154.

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

[11] Hirschberg J., Litman D.J., and Swerts M. (1999). "Prosodic cues to recognition errors", Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), December 12-15 1999, Keystone, Colorado.

[12] Krahmer, E., Swerts, M., Theune, M. and Weegels, M. (2001), "Error Detection in Spoken Human-Machine Interaction" International Journal of Speech Technology, 4, pp. 19-30.

[13] Ladd, D.R. (1996). Intonational Phonology. Cambridge: Cambridge University Press.

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