Multimodal feedback cues in human-machine interactions

Björn Granström, David House* and Marc Swerts***

Names in alphabetical order, *CTT, KTH, Stockholm, Sweden, **CNTS, Antwerp University, Belgium and TU/e, Eindhoven, The Netherlands

{bjorn; davidh}@speech.kth.se, m.g.j.swerts@tue.nl

Abstract

This paper reports on an experiment, whose goal it was to explore the relevance of both acoustic and visual cues for signaling 'negative' or 'affirmative' feedback in a conversation. Using the WaveSurfer software developed at CTT, the stimuli were created by orthogonally varying 6 parameters (4 visual and 2 acoustic ones), which always had two settings: one which was hypothesised to lead to affirmative feedback responses, and one which was hypothesised to lead to negative responses. Listeners were told that they were going to see and hear a series of exchanges between a talking head, representing a travel agent, and a human who wants to make a booking with the agent. They had to imagine that they were standing beside the human, and they were witnessing a fragment of a longer dialogue exchange. Their task was to rate this fragment in terms of whether the agent signals that he understands and accepts the human utterance, or whether the agent signals that he is uncertain about the human utterance. Results show that listeners are sensitive to both the visual and acoustic features when judging the utterances in terms of their function as feedback signals. Four of the six parameters had significant influence on the judgements, with Smile and F0 as the most prominent, followed by Eyebrow and Head_movement. Eye_closure and Delay contributed only marginally to the judgements but the tendency was in the expected direction.

1. Introduction

One of the central claims in many theories of conversation is that dialogue partners seek and provide evidence about the success of their interaction (Clark and Schaeffer 1989; Traum, 1994; Brennan, 1990). That is, conversants tend to follow a proof procedure to check whether their utterances were understood correctly or not and constantly exchange specific forms of feedback that can be affirmative ('go on') or negative ('do not go on'). Previous research has brought to light that conversation partners can monitor the dialogue this way on the basis of at least two kinds of features not encoded in the lexico-syntactic structure of a sentence: namely, prosodic and visual features. First, utterances that function as negative signals appear to differ prosodically from affirmative ones in that they are produced with more 'marked' settings (e.g. higher, louder, slower) (Shimojima et al, 2002; Krahmer et al 2002). Second, other studies reveal that, in face-to-face interactions, people signal by means of facial expressions and specific body gestures whether or not an utterance was correctly understood (Satinder et al. 1999).

Given that current spoken dialogue systems are prone to error, mainly because of problems in the automatic speech recognition (ASR) engine of these systems, a sophisticated use of feedback cues from the system to the user is potentially very helpful to improve human-machine interactions as well (e.g. Hirschberg et al. 2000). There are currently a number of advanced multimodal user interfaces in the form of talking heads that can generate audio-visual speech along with different facial expressions (Beskow 1995, 1997; Beskow et al. 2000, 2001; Granström et al. 2001). However, while such interfaces can be accurately modified in terms of a number of prosodic and visual parameters, there are as yet no formal models that make explicit how exactly these need to be manipulated to synthesize convincing affirmative and negative cues.

One interesting question, for instance, is what the strength relation is between the potential prosodic and visual cues. The interaction between acoustic intonational gestures (F0) and eyebrow movements has been studied in production in e.g. Cavé et al (1996). A preliminary hypothesis is that a direct coupling is very unnatural, but that prominence and eyebrow movement may co-occur. In an experiment investigating the contribution of eyebrow movement to the perception of prominence in Swedish (Granström et al. 1999), words and syllables with concomitant eyebrow movement were perceived as more prominent than syllables without the movement. In addition, other research on multimodal cues for prominence (House et al. 2001; Krahmer et al., submitted) has shown that there may be subtle interactions between visual and prosodic modalities on subjects' perception of spoken stimuli, so that it may also be the case that prosodic and visual cues interact when used for backchanneling (see also Massaro et al. 1996).

The goal of the current paper is to gain more insight into the relative importance of specific prosodic and visual parameters for giving feedback on the success of the interaction. In the research presented below, use is made of a talking head whose prosodic and visual features are orthogonally varied in order to create stimuli that are presented to subjects who have to respond to these stimuli and judge them as affirmative or negative backchanneling signals.

2. Method

2.1. Stimuli

The stimuli consisted of an exchange between a human, who was intended to represent a client, and the face, representing a travel agent. An observer of these stimuli could only hear the client's voice, but could both see and hear the face. The human utterance was a natural speech recording and was exactly the same in all exchanges, whereas the speech and the facial expressions of the travel agent were synthetic and variable. The fragment that was manipulated, always consisted of the following two utterances:

Human: *"Jag vill åka från Stockholm till Linköping."*
("I want to go from Stockholm to Linköping.")
Head: *"Linköping."*

Using the WaveSurfer software developed at CTT (Sjölander and Beskow, 2000), the stimuli were created by orthogonally varying 6 parameters (4 visual and 2 prosodic ones), using two possible settings for each parameter: one which was hypothesised to lead to affirmative feedback responses, and one which was hypothesised to lead to negative responses. For all stimuli, the head was given a neutral face during the time that the human was talking, with three eyeblinks at randomly chosen but natural intervals. The facial expressions changed during the head's response utterance, through modifications of the following parameters shown in Table 1:

Table 1: Different parameters and parameter settings used to create different stimuli

	Affirmative setting	Negative setting
Smile	Head smiles	Neutral expression
Head movement	Head nods	Head leans back
Eyebrows	Eyebrows rise	Eyebrows frown
Eye closure	Eyes narrow slightly	Eyes open widely
F0 contour	Declarative	Interrogative
Delay	Immediate reply	Delayed reply

The parameter settings were largely created by intuition and observing human productions. The smile was a gesture throughout the whole utterance, largely encompassing a widening of the mouth and a slight upwards movement of the mouth corners. The head movement for the affirmative setting was a short nod (300 ms) starting at the first vowel. The negative setting comprised a rise of the head throughout the whole utterance. The eyebrow rise for the affirmative setting was initiated at the start of the utterance, being at its maximum from the start of the second syllable to the end. The eyebrow frown for the negative setting was an immediate frown from the beginning of the utterance which extended throughout the utterance. The affirmative gesture for the eye closure was a short (250 ms) narrowing of the eyes starting in the middle of the first vowel. The negative gesture was a widening of the eyes during the entire utterance. The affirmative and negative F0 contours were based on two natural utterances (see Figure 1). The delay for the negative setting was one second longer (1150 ms) compared to the essentially immediate response (150 ms) for the affirmative setting.

All combinations of the two settings for the 6 parameters led to a total of 64 stimuli, which were presented to listeners in a perception experiment (see below). In principle, we could have included at least two additional prosodic parameters in our test, tempo and loudness, since these have also been shown to signal affirmative and negative feedback. However, apart from the fact that this would have increased the number of stimuli considerably so that it would be difficult to present all of them in a single experiment, we decided not to take these into account because temporal modifications did not easily fit in our orthogonal design, since just changing the tempo would basically have affected the speed of change in all other visual and prosodic parameters as well. Loudness was excluded since it was uncertain if loudness effects would be perceptible in a group experiment. Samples of the resulting stimuli are seen in Figure 2.

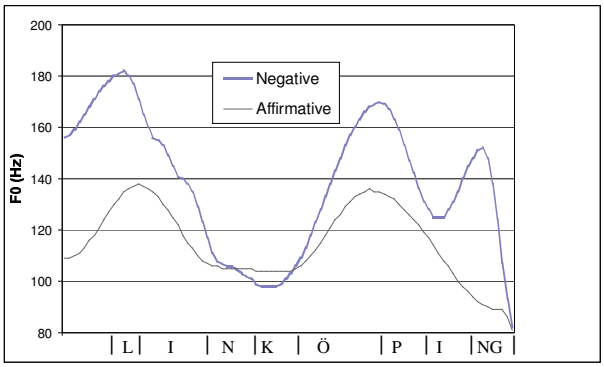

Figure 1: F0 contours of the test word *Linköping*

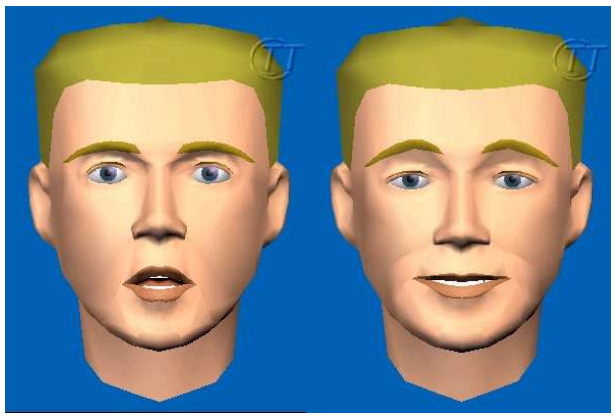

Figure 2: The all-negative and all-affirmative faces sampled in the end of the first syllable of *Linköping*

2.2. Testing

The actual testing was done via a group experiment using a projected image on a large screen. Listeners were told that they were going to see and hear a series of exchanges between a talking head, representing a travel agent, and a human who wants to make a booking with the agent (see example above). They had to imagine that they were standing beside the human, and they were witnessing a fragment of a larger dialogue exchange. Subjects were told that they could both see and hear the talking head, but only hear the human, and they were informed that the visual expression of the head and the pronunciation of *'Linköping'* by the head varied, whereas the human utterance was the same in all conditions. Their task was to respond to this dialogue exchange in terms of whether the head signals that he understands and accepts the human utterance, or rather signals that the head is uncertain about the human utterance. In addition, they needed to express on a 5-point scale how confident they were about their response. They were asked to always give an answer, even if they did not have an intuition as to what the head was signalling. No feedback was given on the 'correctness' of the responses; the stimuli were presented in a randomized order. Each stimulus was presented only once. The silent interval between two consecutive stimuli was 4.5 sec. The interval between the onset of each stimulus was either about 7 or 8 seconds depending on the delay parameter. Both the first three and the

final two utterances were dummies, which were excluded from the analyses afterwards, to make sure that the stimuli were not biased by unwanted 'list' effects. All subjects were volunteers, recruited from KTH personnel. They were not paid for their contribution, but were given coffee and cake after the experiment. After excluding the responses from three subjects who made some unrecoverable errors on their answer sheets, the responses from 17 subjects could be retained for further analyses.

3. Results

As can be seen in Figure 3 there is only a weak tendency for extreme responses to obtain a higher confidence rating than the more ambiguous ones. In this figure the affirmative response is given the value +1 and the negative is –1. The numbers plotted in the figure are mean confidence rating versus mean response for each individual stimulus. There is a tendency for the stimuli to be judged as being more affirmative than negative, with four different stimuli receiving unanimously positive responses (mean value +1 across subjects) but with no stimuli receiving unanimously negative responses (mean value –1 across subjects). Six of the subjects gave more than two-thirds negative responses, while one subject gave only affirmative responses. All subjects, however, used the full confidence scale from 1 to 5, and all results were thus retained in the analysis below.

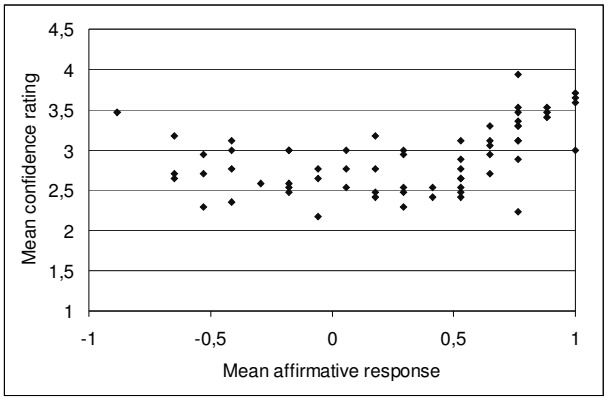

Figure 3. Mean confidence rating for the different stimuli.

The analyses presented are based on numbers that combine the different scores of the subjects, i.e. their yes/no reponses and the confidence rating, in the following way: the responses to a stimulus as a negative or an affirmative cue were first reinterpreted as -1 or 1, respectively, and then multiplied by the confidence rating to obtain a score on a scale between -5 (very negative) and +5 (very affirmative).

These latter numbers were analysed statistically via repeated measurements ANOVA's run on each of the six parameters of our experimental design. Table 2 gives the mean values for each affirmative and negative setting, the value difference between the affirmative and negative settings, and the corresponding F-statistics. This table shows that 4 of the 6 parameters (Smile, F0_contour, Eyebrow and Head_movement) have a significant effect on subjects' responses, with affirmative settings leading to higher, positive values than the negative settings. The effects of Eye_closure and Delay are not significant, but the trends observed in the means are clearly in the expected direction.

There appears to be a strength order with Smile being the most important factor, followed by F0_contour, Eyebrow, Head_movement, Eye_closure and Delay. In Figure 4, the mean response value difference (from Table 2) for stimuli with the indicated cues set to their hypothesised affirmative setting and their negative setting is shown.

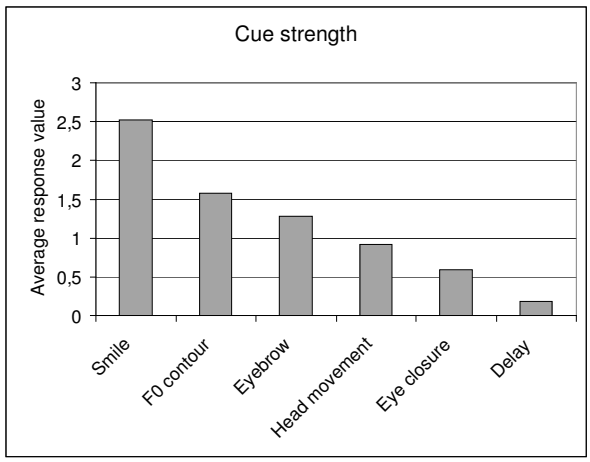

Figure 4. The mean response value difference for stimuli with the indicated cues set to their affirmative and negative value.

The combined effect of cues is visualized in Figure 5. From left to right, the figure shows a monotone increase in affirmative judgments from stimuli that have only negative settings to stimuli that have only affirmative settings. Also in this case a bias towards affirmative responses can be observed. It is obviously not one single factor which has a predominant effect on subjects' responses, but rather it is the case that subjects attend to combinations of features. A further examination of the data did not, however, reveal any specific interaction between the different features, rather the combinations tend to have an additive effect on the responses.

Table 2: Mean value for affirmative and negative settings of different parameters, mean difference value and corresponding F-statistics.

	Affirmative	Negative	Diff. value	F(1,62)	p	η^2
Smile	2.19	-0.33	2.52	61.18	<.001	.50
F0 contour	1.72	0.14	1.58	15.07	<.001	.20
Eyebrow	1.57	0.29	1.28	9.06	<.005	.13
Head movement	1.39	0.47	0.92	4.33	<.05	.07
Eye closure	1.23	0.64	0.59	1.74	n.s.	-
Delay	1.02	0.84	0.18	< 1	n.s.	-

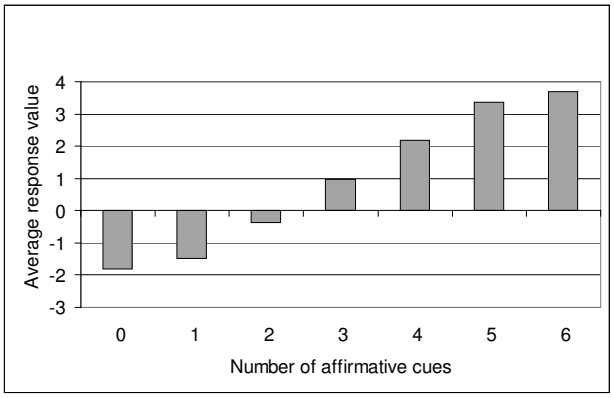

Figure 5. The average response value for stimuli with different number of affirmative cues

4. Discussion and conclusion

Our research has shown that subjects are sensitive to both acoustic and visual parameters when they have to judge utterances as affirmative or negative feedback signals. Although the results of the experiment do not indicate any unexpected cue interactions, the differences between cue strengths can be of interest when implementing feedback signals in animated agents. It is noteworthy that the smile cue (a visual cue) contributed the most to the perception of affirmative feedback. Of all the visual cues used in the experiment, the smile is the one least likely to be associated with a prosodic function other than feedback, such as prominence. The other cues, especially brow raising and nodding, can potentially be associated with a prominence function as well as signalling feedback (House, et al. 2001). The fact that the brow frown functions as a negative cue is not surprising as the frown can signal confusion or disconcernment. Brow rise as an affirmative cue is more surprising in that a question or surprise can be accompanied by raised eyebrows. In this experiment, however, the brow rise was quite subtle. A larger raising movement is likely to be interpreted as surprise. The fact that F0 was the second strongest cue demonstrates the importance of acoustic parameters for feedback in the multimodal environment. The relative importance of F0 may also have been enhanced by the shortness of the utterance.

One obvious next step is to test whether the fluency of human-machine interactions is helped by the inclusion of such feedback cues in the dialogue management component of a system.

5. Acknowledgments

Marc Swerts is also affiliated with the Flemish Fund for Scientific Research (FWO-Flanders). The research reported upon here was mainly conducted during a stay at KTH by Swerts during April 2001. This trip was partly sponsored by CLIF (Computational Linguistics in Flanders). Thanks are due to Carel van Wijk (KUB, Tilburg) for help with statistics and to Jens Edlund (KTH, Stockholm) for invaluable technical assistance.

6. References

[1] Beskow, J. (1995). Rule-based Visual Speech Synthesis. In *Proceedings of Eurospeech '95,* 299-302. Madrid.

[2] Beskow, J. (1997). Animation of Talking Agents. In *Proceedings of AVSP'97, ESCA Workshop on Audio-Visual Speech Processing,* 149-152. Rhodes, Greece.

[3] Beskow, J., B. Granström and D. House. (2001) 'A Multimodal Speech Synthesis Tool Applied to Audio-Visual Prosody.' In E. Keller, G. Bailly, A. Monaghan, J. Terken, & M. Huckvale (eds.) *Improvements in Speech Synthesis,* 372-382. John Wiley & Sons, Inc. New York, New York.

[4] Beskow, J., Granström, B., House, D. and Lundeberg, M. (2000). Experiments with verbal and visual conversational signals for an automatic language tutor. In *Proceedings of InSTiL 2000,* 138-142. Dundee, Scotland.

[5] Brennan S.E. (1990). *Seeking and providing evidence for mutual understanding.* Unpublished doctoral dissertation, Stanford University, Stanford, CA.

[6] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In Bunnell, H.T. and W. Idsardi (eds.), *Proceedings ICSLP 96*, 2175-2178, Philadelphia, PA, USA.

[7] Clark H.H. and Schaeffer E.F. (1989). Contributing to discourse. *Cognitive Science 13*, 259-294.

[8] Granström, B., House, D. and Lundeberg, M. (1999). Prosodic Cues in Multimodal Speech Perception, In *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)*, 655-658, San Francisco.

[9] Granström B, House D, Beskow J, & Lundeberg M (2001) Verbal and visual prosody in multimodal speech perception. In W. van Dommelen and T. Fretheim (eds.) *Nordic Prosody VIII,* 77-87. Peter Lang. Frankfurt.

[10] Hirschberg, J., Litman D. and Swerts, M. (2001) Identifying user corrections automatically in spoken dialogue systems. *Proc. NAACL 2001* Pittsburg.

[11] House, D., Beskow, J. and Granstrom, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. *Proc. Eurospeech 2001*, 387-390, Aalborg, Denmark.

[12] Krahmer, E. Ruttkay, Z., Swerts, M. Wesselink, W. (subm.) Pitch, eyebrows and the perception of focus. Submitted to *Prosody 2002*.

[13] Krahmer, E. Swerts, M., Theune M. and Weegels M. (2002). The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates, *Speech Communication, 36* (1-2), 133-145.

[14] Massaro, D. W., Cohen, M. M. and Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am. 100*. 1777-1786.

[15] Satinder P. Gill, Kawamori, M. Katagiri, Y. and Shimojima. A. (1999) "Pragmatics of Body Moves," *The Proceedings of 3rd International Cognitive Technology Conference,* 345-358.

[16] Shimojima, Y. Katagiri, H. Koiso, and M. Swerts (2002), Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses," *Speech Communication, 36* (1-2), 113-132.

[17] Sjölander, K and Beskow, J. (2000). WaveSurfer - an open source speech tool. In *Proceedings of ICSLP 2000*, vol. 4, 464-467. Beijing, China.

[18] Traum, D.R. (1994), *A computational theory of grounding in natural language conversation*. PhD thesis, Rochester.