Human-Computer Interaction - INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 701-704

Facial Orientation During Multi-party Interaction with Information Kiosks

Ilse Bakx, Koen van Turnhout & Jacques Terken

Dept. of Industrial Design, Eindhoven University of Technology, The Netherlands

I.H.M.Bakx, K.G.v.Turnhout, J.M.B.Terken@tue.nl

Abstract: We hypothesize that the performance of multimodal perceptive user interfaces during multi-party interaction may be improved by using facial orientation of users as a cue for identifying the addressee of a user utterance. Multi-party interactions were collected in a user test where one participant would both interact with an information kiosk and negotiate with another person about the information to be obtained. It was found that users indeed look at the system when they speak to the system, but that they also look at the system most of the time when they negotiate with the other person. It is concluded that facial orientation by itself does not fully identify the addressee of a user utterance, but there are promising results for a combination of facial orientation and utterance length.

Keywords: Multimodal interaction, perceptive user interfaces, facial orientation

1 Introduction

The current research links to two research trends. In the field of Multimodal Perceptive User Interfaces researchers aim to create systems that exploit the natural expressiveness of the users, by allowing them to interact with the system in ways that go beyond those supported by conventional Graphical User Interfaces, e.g. by speech and gesture. In addition, multimodal-perceptive user interfaces are sensitive to aspects of behaviour that are not as such consciously controlled, but nevertheless provide useful information to the system, such as eye gaze.

A second trend is a move away from single user-single system interaction to the field of ubiquitous computing or aware environments. Researchers in this field have come to realize that their systems need to deal with situations where the interaction with the system is interleaved with other activities, including human-human interaction. In such situations, individual systems need to develop a notion of "addressee" and need to be aware of the fact that a particular user action may be addressed to someone else or to another system in the environment, the more so if the user is using speech to address other people and systems.

One possible cue that a perceptive system may have in order to determine whether it is being addressed by the user is eye gaze, or its concomitant feature facial orientation (Morency et al., 2002; Stiefelhagen & Zhu, 2002), and this is the focus of the current paper. Concretely, we investigate whether facial orientation can be used for differentiating between utterances directed to the system and utterances directed to other people in the environment.

Maglio et al. (2000) and Brumitt & Cadiz (2000) show that people look at the device that they talk to when there are multiple systems to address in a room. Vertegaal et al. (2001), confirming findings from conversation analysis, show that in multiparty conversation it is much more likely that people are looking at the person they talk to than at someone else. It enables them to monitor whether the speaker is capturing the attention of the addressee and whether the message gets across.

From these considerations we derive the hypothesis that, in a situation where a user alternates between talking to a system and talking to other people in the environment, they look at the system when addressing the system and look at the other people when addressing those other people. If so, facial orientation can be used to discard utterances that are not meant for the system. In the following we call this the Conversational hypothesis.

Alternatively, we derive a hypothesis from Argyle & Graham (1977) that the above-mentioned

forces driving gaze behaviour may be overruled by the presence of "situational attractors", i.e. objects or situations in the environment that attract people's eye gaze when they are talking to each other. From everyday experience we may think of a television set that draws our attention when being engaged in a conversation.

According to the Conversational hypothesis we predict that, in a situation where a user is interacting with a system and in the meantime having a conversation with another person in the environment, facial orientation indicates the addressee of an utterance and can be used to discard utterances that are not addressed to the system. According to the Situational Attractor hypothesis we predict that facial orientation does not indicate the addressee of a user utterance, because the face will be directed to the system both when the user is interacting with the system and when s/he is talking to another person in the environment.

2 Methods

A laboratory experiment was conducted in which pairs of subjects had to plan trips. The subjects negotiated particular aspects of the trips and in the meantime one of the subjects interacted with a multimodal information system to obtain train timetable information relevant to the trip.

2.1 System

The multimodal information system supported both speech input and gesture input (tapping the screen). With respect to the speech input, the interaction was user-initiated (tap-and-talk): the user could enter values for the fields in a form by tapping the corresponding fields on the screen, thereby activating the microphone for speech input (see Figure 1). Station names, times and other dates than today or tomorrow required speech input. The remaining values could be selected by direct manipulation, using the buttons on the screen. The screen would show the results of the speech recognition module, and activated buttons for the remaining values would change their state, so that the screen always gave an up-to-date view of the current state of the dialogue. Further information about the input facilities and how they can be used are given in Sturm et al. (2002).

The visual part of the interface was implemented as a Java-applet on a desktop computer with a touch screen and no keyboard. To interact with the system the users called the system with an ordinary telephone, using a headset with close talking microphone in order to keep both hands free for interaction with the touch screen. The touch screen was mounted at eye level so that the users stood in front of the screen, as if they were standing in front of an information kiosk.

Figure 1: Screen shot of the fill-in form

2.2 Subjects and Tasks

Twelve subjects (five female and seven male) participated in the test, making up six pairs. They were all employees of Eindhoven University of Technology. After a short introduction there was some time for individual subjects to explore the system. They were told that they had to plan two trips on the basis of scenarios, one to a museum and one to a zoo, together with the other member of the pair. The participants read the scenarios and the information of the museums or zoos. Members of a pair received different versions of the scenarios, giving different opportunities and priorities for the trips, in order to stimulate discussion and negotiation. One of the members of a pair, wearing the headset with the close-talking microphone (to be referred to as User in the following), interacted with the system. The other member of the pair, to be referred to as Partner, did not interact with the system during that scenario but engaged in the discussion/negotiation with the User. Hence, the User's utterances could be directed either to the system or to the Partner. The Partner's utterances could be directed only to the User. Each participant acted as User during one trip and as Partner during the other trip. Each scenario involved two dialogues with the system: one for the outward trip, one for the return trip. Participants were not told that their facial orientation was analysed afterwards.

2.3 Data Capture and Evaluation Measures

Both participants were recorded with a camera during all dialogues. Detailed analyses were conducted of the spoken utterances and the facial orientations. An utterance was defined in terms of a sequence of spoken words not interrupted by a break of more than 0.5 seconds. A change of facial orientation was only scored when people actually moved their head.

3 Results and Discussion

3.1 Facial Orientation as Cue

22 dialogues were recorded (2 other dialogues were not recorded due to technical failures). We analysed the facial orientation of both participants during the utterances. Table 1 shows the results of the analyses of facial orientation, which include Looking at screen, Looking at other (this can be the other person or something else) or switching facial orientation during the utterance (Switch).

Speaker	Addressee	Look at screen	Look at other	Switch
User	System	115 (94%)	5 (4%)	3 (2%)
	Partner	294 (57%)	64 (12%)	158 (31%)
Partner	User	292 (60%)	54 (11%)	138 (29%)

Table 1: Facial orientation during utterances in 22 two-person interactions.

We see that Users looked at the screen while they were talking to the system in 94% of the cases. This is not surprising given the fact that users had to tap the screen before starting to talk to the system. However, the User also looked at the screen in 57 % of the cases where s/he was talking to the Partner. Similarly, the Partner looked at the screen in 60% of the cases when talking to the User. These results are better predicted by Argyle's "Situational Attractor" hypothesis than by the Conversational hypothesis.

3.2 Design perspective

From the point of view of design, let us assume that we use facial orientation as a cue to identify the addressee of the utterance, taking utterances where the speaker looks at the screen as addressed to the system and utterances where the speaker looks elsewhere as utterances meant for someone else (the speaker here can be either the User or the Partner, since the system cannot distinguish between the two). The results show that such a strategy would produce many False Alarms in the current situation, i.e. cases where the system thought it was being addressed whereas in fact the speaker was talking to someone else. Applying metrics from Information Retrieval, that are commonly used to evaluate the effectiveness of cues, this strategy gives high Recall [hits/(hits+misses)] of .93, but a very low Precision [hits/(hits+false alarms)] of .16. This is a bad result because we want to get high Precision, corresponding to discarding as many irrelevant utterances as possible.

Therefore we looked at other cues that might help differentiating between utterances directed to the system and utterances directed to the partner. It should be noted that we might have used tapping behaviour rather than gaze behaviour, and this would have given us almost perfect Precision and Recall. However, the Tap-and-Talk interface was used mostly for sake of convenience; in fact we aim at an interface that supports interaction without requiring the user to indicate by tapping when s/he wants to address the system. From the recordings of the experimental sessions we found that utterances directed to the partner tended to last longer (mean = 1.7 s) than utterances directed to the system (mean = 1.1 s). Figure 2 shows that the difference in the mean is due to a much longer tail for utterances directed to the Partner.

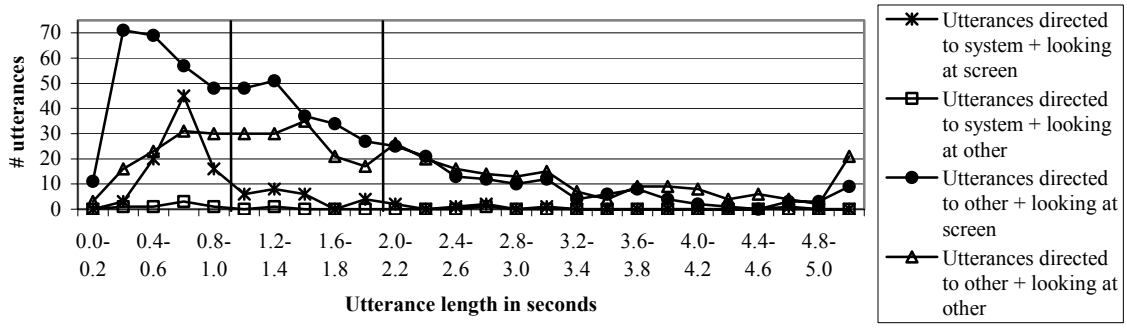

Figure 2: Frequencies of utterance lengths combined with facial orientation

Figure 2 shows the effects of combining facial orientation and utterance length. If we put a threshold at an utterance length of 1 second, we exclude extra 330 utterances, but also falsely reject extra 31 utterances meant for the system (giving Higher Precision, .25, at the expense of lower Recall, .68). Setting the threshold at 2 seconds gives Precision and Recall of .19 and .88, respectively.

Conversation Analysis has shown that eye gaze (and facial orientation) is not only governed by "looking at the addressee" but also by "looking at the speaker". If we include the facial orientation of the Non-speaking participant as an additional cue, we get the results in Table 2, showing the effect of combining gaze behaviour of both participants and a threshold for utterance length of 2 seconds. Precision and Recall for Table 2 are .21 and .80, respectively, giving better Precision (although still quite low) with reasonable Recall.

	Both look at screen and < 2 sec	Not both look at screen or > 2 sec
Utterance for system	98 (80%)	25 (20%)
Utterance for partner	362 (36%)	638 (64%)

Table 2: Combination of facial orientation of both participants and a 2 second threshold for utterance length.

4 Conclusion

We find that facial orientation in situations where users interact with a multimodal system and interchange this activity with talking to another person standing next to him/her look at the screen in the majority of cases, both when talking to the system and when talking to the other person. This result is better predicted by the Situational Attractor hypothesis, holding that rules for gaze behaviour may be overruled by the presence of interesting objects and situations in the environment, than by the Conversational hypothesis, which holds that a speaker looks at the person s/he is addressing. From a design point of view, the results mean that facial orientation can be used as a cue to identify the addressee of an utterance only asymmetrically: if the speaker is looking away from the system when talking, we can be pretty sure that the utterance is not directed to the system [P(*addressing system*|*looking away from system*) = 0.02, collapsing across User and Partner]. But if the speaker is looking at the system when talking, we can derive little positive information from this: still the speaker might be talking to someone else [P(*addressing system*|*looking at system*) = 0.16].

Other cues such as utterance length provide additional information for discarding irrelevant utterances, but at the expense of rejecting more utterances that are in fact addressed to the system. Further experiments need to be conducted in order to determine optimal levels of Precision and Recall. We don't know whether discarding utterances that are in fact addressed to the system should be avoided or whether modest levels of discarding relevant utterances are in fact acceptable if it prevents the system from reacting to many irrelevant utterances. Furthermore, we need to experiment with a different architecture, where the system would not simply decide whether an utterance is addressed to the system or not, but where the system weighs gaze behaviour against the confidence levels that are computed in the speech recognition module.

Acknowledgements

The current research is part of the CRIMI project (Creating Robustness in Multimodal Interaction) (http://www.industrialdesign.tue.nl/research/uceGroup/crimi), and is funded by the Dutch Ministry of Economic Affairs through the Innovation Oriented Programme Man-Machine Interaction (IOP-MMI).

References

Argyle, M. & Graham, J. (1977) The Central Europe Experiment - looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour,* 1, pp. 6-16.

Brumitt, B. & Cadiz, J.J. (2001) "Let there be light!"- Examining interfaces for homes of the future. *Proceeding of INTERACT 2001,* 375-382

Maglio, P.P., Matlock, T., Campbell, C.S., Zhai, S. & Smith, B.A. (2000) Gaze and speech in attentive user interfaces. *Proceedings of The third International Conference on Multimodal Interfaces ICMI*, Oct 14-16, 2000, Beijing, China. Springer. pp 1-7.

Morency, L.P., Rahimi, A., Checka, N. & Darrell, T. (2002) Fast Stereo-Based Head Tracking for Interactive Environment, *Proceedings of the Int. Conference on Automatic Face and Gesture Recognition, 2002*

Stiefelhagen, R. & Zhu, J. (2002) Head orientation and gaze direction in meetings. *Proceedings of ACM CHI 2002.* Minneapolis: ACM

Sturm, J., Bakx, I., Cranen, B. & Terken, J. (2002) The effect of prolonged use on multimodal interaction. *Proceedings of the ISCA Workshop on Multi-modal Dialogue in Mobile Environments*, Kloster Irsee, Germany

Vertegaal, R., Slagter, R., Van der Veer, G.C. & Nijholt, A. (2001) Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. *Proceedings of ACM CHI 2001.* Seattle: ACM.