On the Importance of Eye Gaze in a Face-to-Face Collaborative Task

Sascha Fagel, Gérard Bailly, Frédéric Elisei, Amélie Lelong

On the Importance of Eye Gaze in a Face-to-Face Collaborative Task

Sascha Fagel Gérard Bailly Frédéric Elisei Amélie Lelong
GIPSA-lab, Speech & Cognition dpt., UMR 5216 CNRS/U. Grenoble
961 rue de la Houille Blanche
38402 Saint Martin d'Hères CEDEX
+33 4 76 57 47 11
sascha.fagel@tu-berlin.de,{gerard.bailly,frederic.elisei,amelie.lelong}@gipsa-lab.grenoble-inp.fr

ABSTRACT
In the present work we observe two subjects interacting in a collaborative task on a shared environment. One goal of the experiment is to measure the change in behavior with respect to gaze when one interactant is wearing dark glasses and hence his/her gaze is not visible by the other one. The results show that if one subject wears dark glasses while telling the other subject the position of a certain cube, the other subject needs significantly more time to locate and move this cube. Hence, eye gaze – when visible – of one subject looking at a certain cube speeds up the location of the cube by the other subject. The second goal of the currently ongoing work is to collect data on the multimodal behavior of one of the subjects by means of audio recording, eye gaze and head motion tracking in order to build a model that can be used to control a robot in a comparable scenario in future experiments.

Categories and Subject Descriptors
H.1.2 User/Machine Systems - Human factors; Human information processing; Software psychology

General Terms
Measurement, Experimentation, Human Factors.

Keywords
Eye Gaze, Eye Tracking, Head Motion, Motion Capture, Deixis, Collaborative Task.

1. INTRODUCTION
Speech is a natural and highly developed means of human information exchange between humans. However, while a person speaks there are more sources of information accessible for the listener than just the spoken words. Along with the linguistic content of the speech, para-linguistic and extra-linguistic cues contained in the speech signal are also interpreted by the listener. Moreover, when humans communicate not only through an acoustic channel, e.g. face-to-face, there are also non-verbal cues that accompany speech, appear simultaneously with speech or appear without the presence of speech at all. Aside from static features such as the shape of a person's face, clothing etc., non-verbal cues potentially arise from any movements of the body other than speech articulatory movements. The most obvious non-verbal cues during speech communication originate from movements of the body (Bull and Brown 1985), the face (Collier 1985), the eyes (Argyle and Cook 1976), the hands and arms (Kendon 1983), and the head (Hadar, Steiner et al. 1983). More recent reviews can be found in (Pelachaud, Badler et al. 1996; Beattie and Shovelton 1999; McClave 2000; Heath 2004; Maricchiolo, Bonaiuto et al. 2005; Heylen 2006).

Iconic gestures produced during speech such as nods and shakes for "yes" and "no" are rare. These cues mostly contribute to the conversational structure or add non-verbal information in form of visual prosody. Hence, head motion can be predicted or generated for the use in virtual agents by the use of acoustic prosodic features e.g. by mapping to head motion primitives (Graf, Cosatto et al. 2002; Hofer and Shimodaira 2007), or orientation angles (Busso, Deng et al. 2005; Sargin, Yemez et al. 2008). Eye gaze is linked to the cognitive and emotional state of a person and to the environment. Hence, approaches to eye gaze modeling have to deal with high level information about the communication process (Pelachaud, Badler et al. 1996; Lee, Marsella et al. 2007; Bailly, Raidt et al. 2010).

In scenarios where humans interact in a shared environment, head movements and eye gaze of a person can also relate to objects or locations in that environment and hence deliver information about the person's relation to that environment. While hand movements are often explicitly used to point to objects, head motion and eye gaze yield implicit cues to objects (multimodal deixis, see Bailly, Raidt et al. 2010) that are in a person's focus of attention as the person turns and looks towards the object of interest. The present paper describes an experimental scenario of a face-to-face task-oriented interaction in a shared environment. The accessibility of eye gaze information during the referencing of an object in the environment is manipulated and it is hypothesized that the task completion time decreases when eye gaze cues are prevented or not perceived.

AFFINE'10, October 29, 2010, Firenze, Italy.

2. EXPERIMENTAL METHOD

Procedure

Two subjects are seated on chairs at opposite sides of a table. The table contains two identical designated areas, one in front of each subject. An area consists of 9 slots in a row: each slot has one of the symbols {A,I,O}, each three slots (A,I,O) have the same color {red, green, or blue}. 9 cubes are placed in the slots of one of the two areas. Each cube shows a label, a letter from {P,T,B,D,G,M,N,F,S}, on that side facing the subject (informant) who is sitting close to the cubes. The informant has access to the labels of the cubes but only the other subject (manipulator) is allowed to modify the environment, i.e. to move cubes. Each move starts with a quasi-random pause of the control script that aims to establish mutual attention. Then the computer tells the manipulator confidentially by earphones about the label of one cube to be moved. Then the manipulator tells the label to the informant in order to request the position of that cube. The informant searches among the cubes and informs the manipulator about the position of the requested cube by telling the symbol and color of the slot where it is located. Then the manipulator moves the cube to the opposite field in the area close to herself. See Figure 1 for a snapshot of the game during a move. Figure 2 is a time flow diagram of one such move with the states of the subjects during the interaction and the observed behavior of one of the subjects (more details in section 3.1). 72 of these moves are completed, arranged in 12 rounds of 6 moves. The role assignment (who is informant and who is manipulator) is changed during the experiment as well as the condition (with or without dark glasses).

Figure 1: One subject's view recorded by a head mounted scene camera. Here, it is the informant's view: she sees the labels on the cubes and tells the position of the requested cube. The role of the opposite subject (visible in the figure) is manipulator: she requests the position of a cube by telling its label, moves the cube (here the third from six cubes to be moved) and ends the move by a click on the mouse button.

We monitored four interactions of one person (our reference subject) with four different interactants. During the interaction the reference subject acted as manipulator in 6 rounds and as informant in another six rounds, she wore dark glasses in half of the rounds and did not wear dark glasses in the other half. All rounds are grouped to a block with the same role assignment and condition (dark glasses or not). The order of these blocks was counterbalanced across the 4 recordings so that in recordings 1 and 2 the reference subject was manipulator first and in recording 3 and 4 second, and in recordings 1 and 3 the dark glasses were worn first and in recording 2 and 4 second. Two training rounds of three moves were played before the recording (one for each role assignment) and subjects were instructed to play fast but accurately.

Technical Setup

During the interaction we recorded the subjects' head motions with an HD video camera (both subjects at a time by using a mirror), the subjects' head movements by a motion capture system, the subjects' speech by head mounted microphones, the eye gaze of the reference subject and a video of what she sees by a head mounted eye tracker. Unlike human interlocutors, the eye tracker works on infrared light: it was not affected by the dark glasses. We also monitored the timing of the moves by the log file of the script that controls the experiment. The different data streams are post-synchronized by recording the sync signal of the motion capture cameras as an audio track along with the microphone signals as well as the audio track of the HD video camera, and by a clapper board that is recorded by the microphones, the scene camera of the eye tracker and the motion capture system simultaneously. Figure 3 shows an overview of the technical setup.

Figure 3: Technical setup of the experiment comprising head mounted eye tracking, head mounted microphones, video recording, and motion capture of head movements.

3. RESULTS

Analyses

The speech was annotated on utterance level with Praat (Boersma and Weenink 1996), head orientations are computed and then refined and labeled in ELAN (Hellwig and Uytvanck 2004; Berezm 2007). The timings of the confidential playbacks and mouse clicks that end the moves are imported from the log file of the control script. The timings of the phases of the interaction are inferred from these data, i.e. for the manipulator: wait for confidential instruction, listen to confidential instruction, verbalization of cube request, wait for information, move the cube and complete the move; for the informant: wait for cube request, search the cube, verbalization of its position, observe the move (see Figure 2).

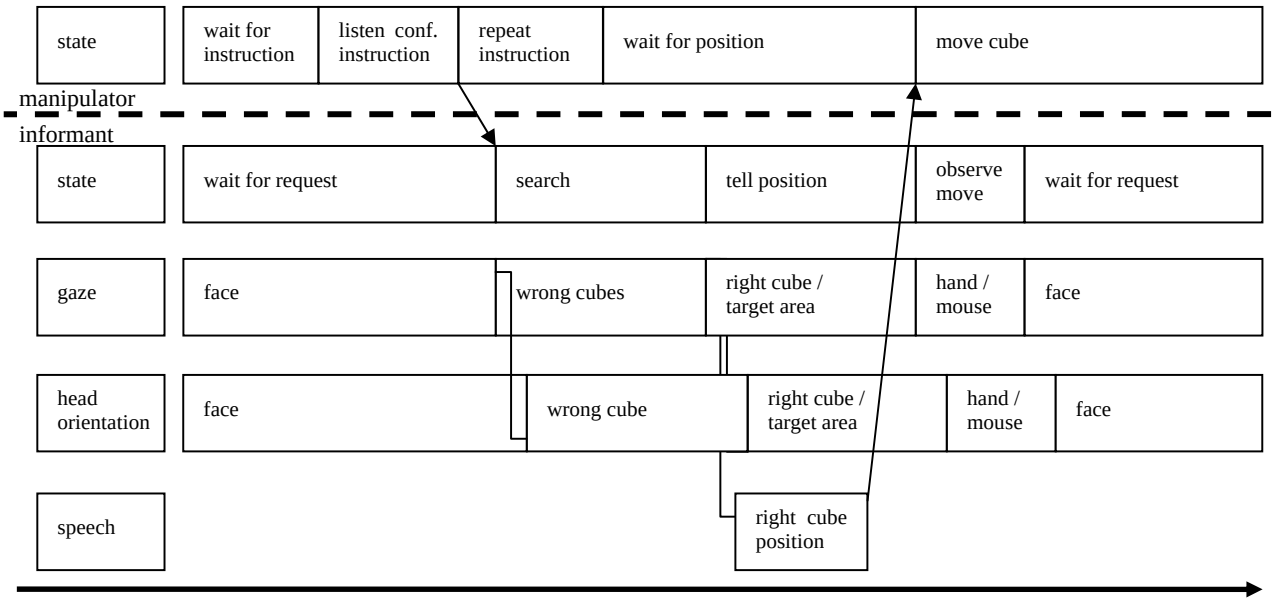

Figure 2: Time flow diagram of the interaction with reference subject in the roll of informant.

The duration from the end of the confidential playback of the instruction to the completion of a move triggered by pressing the mouse button (completion time) was calculated for each move. Additionally, this duration was split at the start of the verbalization of the position of right cube by the informant, which provides the time needed by the informant to search the cube (search time) and the time needed by the manipulator to locate and move the cube (location time).

The number of (wrong) cubes gazed by the reference subject during the search before she finally gazes at the requested cube is determined by visual inspection of the eye tracking data that is superimposed on the video of the eye tracker's scene camera (see Figure 1: the red cross marks the current gaze, here at the target slot where the cube has to be placed). Correlations between the number of wrong cubes, the number of cubes left in the source area (starting with 9 down to 4 in the 6th move of a round), the search time and the location time are calculated.

Completion Time

Task completion time is significantly increased ($p<.001$) when the informant wears dark glasses compared to not wearing glasses. No significantly different completion times were observed for one subject in recording 1 and for both subjects in recording 2. In all other cases completion times are significantly increased. See Table 1 for details.

Search Time and Location Time

Over all recordings the search time was not significantly different between with and without dark glasses. This indicates that the dark glasses did not perturb the search of the cube. Location times, however, i.e. the duration from hearing the position of the cube to its completed relocation, are significantly increased in five of eight cases (the two role assignments in each of the four recordings). No significant differences were observed for the same cases where no different completion times were found. Across all four recordings separately for both role assignments as well as across both role assignments the search times are not significantly differing where the location times are significantly increased. See Table 2 for details.

Number of Cubes

The total number of wrong cubes gazed before the requested cube is found is exactly the same in both conditions. Table 3 shows the correlation between number of wrong cubes gazed at and the number of cubes left as well as their correlation to the search and location times. The location time is negligibly correlated to the number of cubes left and weakly correlated to the number of wrong cubes gazed at. The number of wrong cubes gazed at is moderately correlated to the number of cubes left: there is a tendency to shorter search times when fewer cubes are left (Figure 4). Strong correlation is found between search time and number of wrong cubes gazed at.

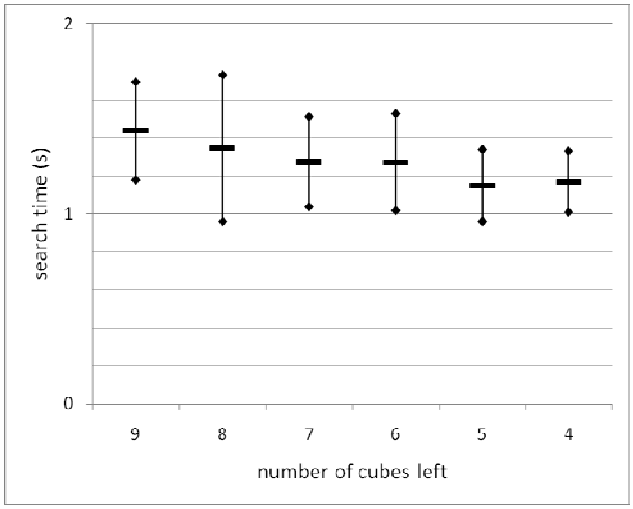

Figure 4. Mean and standard deviation of search time over the number of cubes left.

Table 1. Completion times with and without dark glasses and the significance level of differences.

recording	A manipulator. B no glasses		A manipulator. B with glasses	B manipulator. A no glasses		B manipulator. B with glasses
1	4.39	*	4.74	4.33		4.10
2	3.87		3.75	3.96		4.08
3	3.22	**	3.80	3.82	*	4.40
4	3.47	*	3.77	3.52	*	4.04
all	3.73	*	4.02	3.91	*	4.16
both roles	3.82	**	4.09			

one factor ANOVA: * p<.05, ** p<.001, p>.05 otherwise

Table 2. Search time and location time with and without dark glasses for each recording and across all recordings (same conventions as above).

	search time				location time			
recording	A. no glasses	A. w/ glasses	B. no glasses	B. w/ glasses	A. no glasses	A. w/ glasses	B. no glasses	B. w/ glasses
1	1.36	1.27	1.49	1.32	3.03 **	3.47	2.84	2.78
2	1.31	1.27	1.23	1.34	2.55	2.48	2.72	2.74
3	1.15	1.33	1.17	1.22	2.06 **	2.47	2.65 **	3.18
4	1.49	1.47	1.18	1.24	1.98 **	2.30	2.34 *	2.80
all	1.33	1.33	1.27	1.28	2.41 *	2.68	2.64 **	2.87
both roles	1.30	1.31			2.52 **	2.78		

one factor ANOVA: * p<.05, ** p<.001, p>.05 otherwise

Table 3. Correlations between number of wrong cubes gazed at, number of cubes left, search time, and location time.

	No of wrong cubes	*No of cubes left*
cubes left	0.45	
search time	**0.74**	0.35
location time	0.26	*0.12*

4. CONCLUSIONS

The present experiment investigates the impact of the visibility versus invisibility of one subjects eye gaze on the performance in a task-oriented human-human interaction. As the task completion included the localization of an object, a cube. that was explicitly referenced by speech and implicitly by head motion and eye gaze, it was hypothesized that the task completion time will be decreased when dark glasses degrades the visibility of the subject that informs the other one about the position of the requested cube (Kobayashi and Kohshima 2001; Tomasello, Harea et al. 2007). Task completion time is in fact significantly increased when the informant wears dark glasses compared to not wearing dark glasses. Hence, invisibility of eye gaze decreases task performance measured by completion time. Where the time needed by the informant to find the right cube does not differ with or without dark glasses, the time for the other subject (manipulator) to locate and move the cube is significantly increased generally over the whole experiment and more specifically in all cases where the task completion times were increased. Furthermore, the total number of wrong cubes gazed at before the requested cube is found is exactly the same in both conditions. Consequently, dark glasses did not make the search for the cube by its label more difficult for the informant but only blocked the visibility of the eye gaze to the opposite subject that leads to degraded information for the manipulator to locate the cube. Or put the other way round: visible eye gaze provides an important cue for the location of the object of interest. The availability of the gaze path of the informant through the shared environment is crucial to trigger grasping: the resonance of motor activities during joint visual attention, the mirroring of the quest, favors synchronized analysis and decision.

The rounds played in the present experiment comprise an inherent decline of difficulty due to the decreasing number of alternatives left as possible object of interest. However, this difficulty was most obviously existent for the subject that has to find the object of interest by searching among the labels of the cubes left. The time needed by the opposite subject to locate the object of interest – referred to explicitly by speech and implicitly by head motion and, if visible, eye gaze – only marginally depends on the number of alternatives if not at all. Thus the referencing of the object in space can be assumed as nearly optimal and eye gaze is an integral part of the transmitted information.

The main result of the experiment is that visible eye gaze yields important information about the location of objects of joint interest in a face-to-face collaborative task between two humans. This is evident from the present work. It can be assumed that proper display of eye gaze might be an important aspect in human-robot interaction as well as in mediated task-oriented interaction.

An accompanying paper by Boucher et al provides the first results of our attempts to transfer human behaviors to robots.

5. FUTURE WORK

One of the recordings where the reference subject acts as informant and does not wear dark glasses was analyzed in detail. The timing of the reference subject's behavioral cues regarding gaze, head orientation, and verbalization was extracted. Both the timing of the interaction and the reference subject's behavior will be modeled by a probabilistic finite-state machine that is capable to control a robot in a comparable scenario. Further analyses of the behavior and a second experiment where a robot will act as informant on the basis of the state machine will follow (see preliminary results in the accompanying paper by Boucher, Ventre-Dominey et al. 2010).

Of particular interest is the rate of mutual adaptation if any. We are still seeking for the reasons for adaptive behavior in particular for speech. Such goal-direction collaborative interactions offer unique ways to characterize the impact of mutual adaptation on performance and smoothness of turn-taking.

6. ACKNOWLEDGMENTS

This work was financed by the ANR project AMORCES.

7. REFERENCES

Argyle, M. and M. Cook (1976). Gaze and mutual gaze. London, Cambridge University Press.

Bailly, G., S. Raidt and F. Elisei (2010). "Gaze, conversational agents and face-to-face communication." Speech Communication - special issue on Speech and Face-to-Face Communication **52**(3): 598–612.

Beattie, G. and H. Shovelton (1999). "Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech." Journal of Language and Social Psychology **18**: 438-462.

Berezm, A. L. (2007). "Review of EUDICO Linguistic Annotator (ELAN)." Language Documentation & Conservation **1**(2).

Boersma, P. and D. Weenink (1996). Praat, a System for doing Phonetics by Computer, version 3.4. Institute of Phonetic Sciences of the University of Amsterdam, Report 132. 182 pages.

Boucher, J.-D., J. Ventre-Dominey, P. F. Dominey, G. Bailly and S. Fagel (2010). Facilitative effects of communicative gaze and speech in human-robot cooperation. ACM Workshop on Affective Interaction in Natural Environments (AFFINE). Firenze, Italy.

Bull, P. E. and R. Brown (1985). "Body movement and emphasis in speech." Journal of Nonverbal Behavior **9**(3): 169-187.

Busso, C., Z. Deng, U. Neumann and S. S. Narayanan (2005). "Natural head motion synthesis driven by acoustic prosodic features." Journal of Computer Animation and Virtual Worlds **16**(3-4): 283-290.

Collier, G. (1985). Emotional Expression. Hillsdale, NJ, Lawrence Erlbaum Associates.

Graf, H. P., E. Cosatto, V. Strom and F. J. Huang (2002). Visual prosody: Facial movements accompanying speech. Automatic Face and Gesture Recognition (FGR). Washington, DC, pp. 396-401.

Hadar, U., T. J. Steiner, E. C. Grant and F. C. Rose (1983). "Kinematics of head movements accompanying speech during conversation." Human Movement Science **2**(1-2): 35-46.

Heath, C. (2004). Body Movement and Speech in Medical Interaction. Cambridge, UK, Cambridge University Press.

Hellwig, B. and D. Uytvanck (2004). EUDICO Linguistic Annotator (ELAN) Version 2.0.2 manual. Nijmegen - NL, Max Planck Institute for Psycholinguistics.

Heylen, D. (2006). "Head gestures. gaze and the principles of conversational structure." Journal of Humanoid Robotics **3**(3): 241-267.

Hofer, G. and H. Shimodaira (2007). Automatic head motion prediction from speech data. Interspeech. Antwerp, Belgium, pp. 722-725.

Kendon, A. (1983). Gesture and speech: How they interact. Nonverbal Interaction. J. M. Wiemann and R. P. Harrison. Beverly Hills, CA, Sage Publications**:** 13-45.

Kobayashi, H. and S. Kohshima (2001). "Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye." Journal of Human Evolution **40**: 419-435.

Lee, J., S. Marsella, D. Traum, J. Gratch and B. Lance (2007). The Rickel Gaze Model: A window on the mind of a virtual human. International Conference on Autonomous Agents and Multiagent Systems (AMAAS). Budapest, Hungary, pp. 321-328.

Maricchiolo, F., M. Bonaiuto and A. Gnisci (2005). Hand gestures in speech: studies of their roles in social interaction. Conference of the International Society for Gesture Studies. Lyon.

McClave, E. Z. (2000). "Linguistic functions of head movements in the context of speech." Journal of Pragmatics **32**: 855-878.

Pelachaud, C., N. I. Badler and M. Steedman (1996). "Generating facial expressions for speech." Cognitive Science **20**(1): 1-46.

Sargin, M. E., Y. Yemez, E. Erzin and A. M. Tekalp (2008). "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation." IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(8): 1330-1345.

Tomasello, M., B. Harea, H. Lehmanna and J. Call (2007). "Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis " Journal of Human Evolution **52**(3): 314-320.