Speech, gaze and head motion in a face-to-face collaborative task

Sascha Fagel, Gérard Bailly

SPEECH, GAZE AND HEAD MOTION IN A FACE-TO-FACE COLLABORATIVE TASK

Sascha Fagel and Gérard Bailly

GIPSA-lab Grenoble, France
sascha.fagel@tu-berlin.de

Abstract: In the present work we observe two subjects interacting in a collaborative task on a shared environment. During the experiment we measured the task completion time under two different conditions: either one interactant is wearing dark glasses and hence his/her gaze is not visible by the other one or no interactant is wearing dark glasses. The results show that if one subject wears dark glasses while telling the other subject the position of a certain object, the other subject needs significantly more time to locate and move this object. Hence, eye gaze – when visible – of one subject looking at a certain object speeds up the location of the cube by the other subject.

1 Introduction

Speech is a natural and highly developed means to transmit information 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 also para-linguistic and extra-linguistic cues contained in the speech signal are interpreted by the listener that together constitute the verbal information. 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 [5], the face [7], the eyes [1], the hands and arms [14], and the head [9]. More recent reviews can be found in [10], [18], [11], [16], and [17].

Iconic gestures produced during speech such as nods and shakes for “yes” and “no” are rare. Non-verbal cues mostly contribute to the conversational structure or add information to speech 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 [8] [12], or orientation angles [6] [19]. 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 [18] [15] [2].

All the aforementioned modalities of communication can be used for deixis, i.e. to point to specific objects or regions in space. In scenarios where humans interact in a shared environment, head movements and eye gaze of a person can relate to objects or locations in that environment and hence deliver information about the person’s relation to that environment. Contrastive focus in speech (e.g. blue ball followed by GREEN ball) was shown to provide a processing advantage in the localization of objects [13] and hence can be used for verbal deixis. While hand movements are often explicitly used to point to objects, head motion and eye gaze yield implicit cues to objects (multimodal deixis [2]) 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 while referring to an object in the environment is manipulated and it is hypothesized that the task performance decreases when eye gaze cues are prevented or not perceived.

2 Experiment

2.1 Procedure

Two subjects are seated on chairs at opposite sides of a table. The table contains two identical designated areas, one on either side of the middle line between the subjects. 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 informs 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 the cube. The informant searches among the cubes and tells the symbol and color of the slot where the requested cube is located. Then the manipulator takes the cube, places it on the opposite field in the area close to herself and completes the move by pressing a mouse button. See Figure 1 for a snapshot of the game during a move. 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).

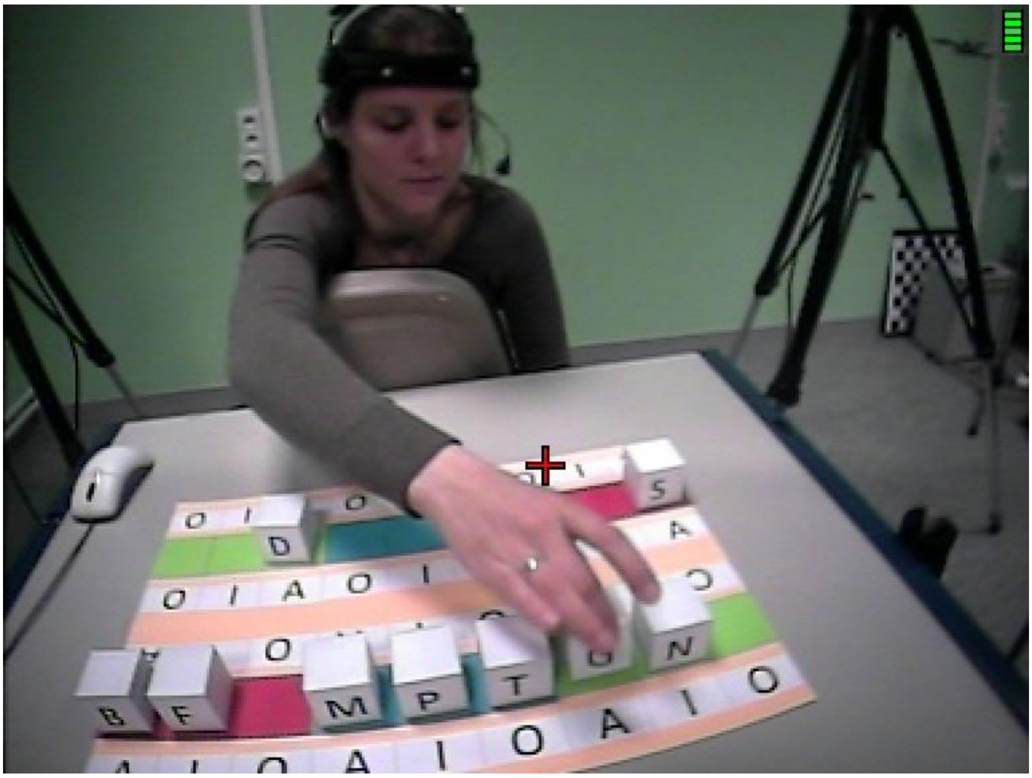

Figure 1 - One subject's view recorded by a head mounted scene camera. This subject's roll is informant: she sees the labels on the cubes and tells the position of the requested cube. The role of the opposite subject (shown in the figure) is manipulator: she requests the position of a cube by telling 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.

Following the design of the game board, the position of a certain cube is verbally specified by e.g. “A rouge” (“A red”); due to the particular structure of the language – the experiment was carried out in French – the noun (label) naturally precedes the adjective (color). Therefore, the first part of a verbalization specifies the exact position inside one of the three colored areas where the rough position is given last. This is important to note as implicit deixis by head and eye gaze can be assumed to be not completely unambiguous but to convey information of the approximate position which thus precedes the accordant verbal information.

We monitored four interactions of one person (reference subject) with four different interactants to study the influence of social rapport and interpersonal attunement. All rounds with the same role assignment and condition are grouped to a block. The order of these blocks is counterbalanced across the 4 recordings so that in recordings 1 and 2 the reference subject is manipulator first and in recording 3 and 4 second, and in recordings 1 and 3 the dark glasses are 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 as fast as possible but not overhastily.

2.2 Technical Setup

During the interaction we recorded the subjects’ heads 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, and the eye gaze of the reference subject and a video of what she sees by a head mounted eye tracker. 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 shutter 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 2 shows an overview of the technical setup.

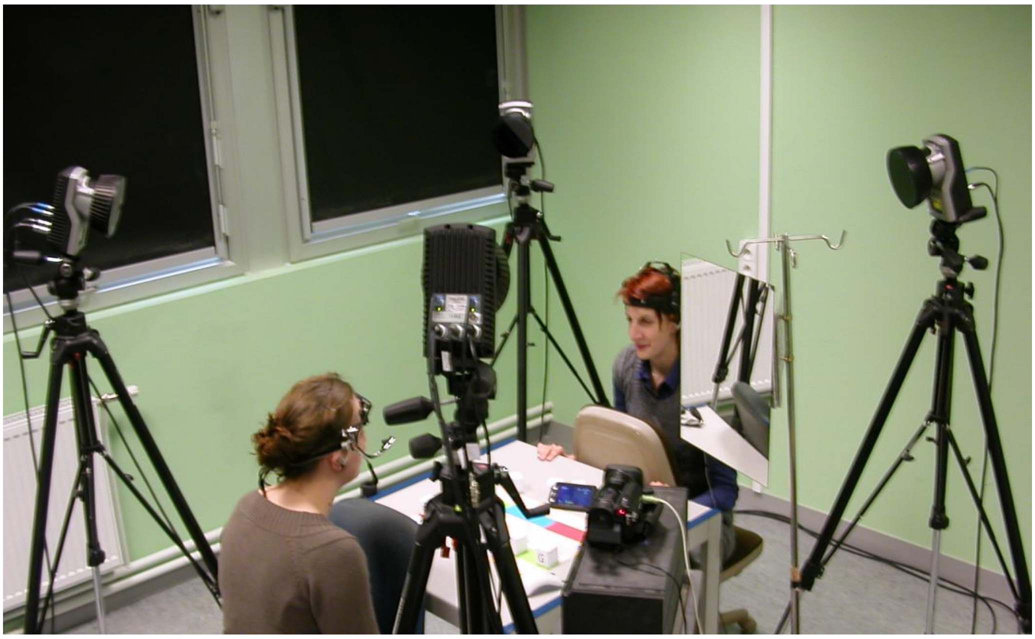

Figure 2 - Technical setup of the experiment comprising head mounted eye tracking, head mounted microphones, video recording, and motion capture of head movements.

3 Results

3.1 Analyses

The speech was annotated on utterance level with Praat [4], head orientations are computed and then refined and labeled in ELAN [3]. 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 its relocation.

The duration from the end of the confidential playback of the instruction to the completion of a move 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 place 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.

The rigid motion of the heads, i.e. translation and rotation, is extracted from the captured motion data of the markers on the head mounts. The rotation data is then analyzed with respect to the angular path the informant's head covered during the search for the right cube. The frame-by-frame angular velocity (Euclidean distance between two subsequent rotation angles) is accumulated over of each of these periods and correlations of this measure of quantity of movement to other measures of the search are computed.

3.2 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.

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**		

* p<.05, ** p<.001, p>.05 otherwise

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

3.3 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.

	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**		

* 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).

3.4 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 3). Strong correlation is found between search time and number of wrong cubes gazed at.

3.5 Head Motion

An ANOVA shows no significant differences between the conditions with and without dark glasses concerning the quantity of movement during the search. However, there is a tendency for the reference subject to more head motion without dark glasses.

Correlation coefficients are calculated between the quantity of movement during the search and the number of wrong cubes gazed at, the number of cubes left and the search time. The quantity of movement is only marginally correlated to the number of cubes left to search. The quantity of movement is strongly correlated to the number of wrong cubes gazed at and moderately correlated to search 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*

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

	No. of wrong cubes gazed (subj A only)	*No. of cubes left*	*search time*
quantity of movement	**0.76**	0.19	0.41

Table 4 - Correlation coefficients between the quantity of movement during the search and the number of wrong cubes gazed at, the number of cubes left and the search time. The number of wrong cubes gazed at is only determined for the reference subject.

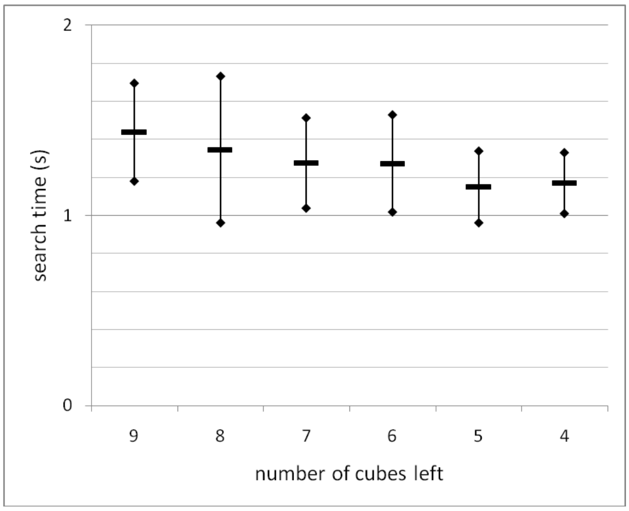

Figure 3 - Mean and standard deviation of search time over the number of cubes left.

4 Conclusions and Discussion

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. The task completion included the localization of an object, a labeled cube, that was explicitly referenced by speech and implicitly by head motion and eye gaze. This deixis by head and eye gaze can be assumed to be ambiguous to some extend but to provide information – on the approximate position of the object of interest – that is redundant to the final word of the verbalization of the object's position. Hence, the implicit deixis preceded the verbal information and potentially speeds up task completion in case the deictic information is used by the subject to locate the object. It was hypothesized that the task performance will be decreased when dark glasses block the visibility of the eyes of the subject that informs the other one about the position of a certain cube. 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 place 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 quantity of movement shows a non-significant tendency to more head motion without dark glasses for the reference subject. This indicates that the reference subject does not use explicit head deixis or implicitly exaggerated movements to overcome the reduced amount of information caused by invisibility of eye gaze. This matches the subject's informal report after the experiment that she did not feel to behave differently with or without wearing dark glasses.

Visual inspection of the recorded video data suggest that the head is not oriented separately to every object that is gazed by the eyes as the head moves somewhat slower than the eyes and the eyes can subsequently fixate two close objects without a change of the head orientation. Hence, the head orientation provides less accurate as well as delayed information about the position of the object of interest compared to eye gaze. This leads to increased task completion time.

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.

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.

6 Acknowledgments

This work was financed by the ANR Amorces. Thanks to Amélie Lelong and Frédéric Elisei for their help during the experiment.

7 References

[1] Argyle, M. and Cook, M. 1976. *Gaze and Mutual gaze*. Cambridge University Press.

[2] Bailly, G., Raidt. S., Elisei, F. 2010. Gaze, conversational agents and face-to-face communication. *Speech Communication* 52, 3, 598-612.

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

[4] Boersma, P. and Weenink, D. 2009. Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/

[5] Bull, P.E. and Brown, R. 1977. Body movement and emphasis in speech. *Journal of Nonverbal Behavior* 16.

[6] Busso, C., Deng, Z., Neumann, U., Narayanan, S.S. 2005. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation and Virtual Worlds* 16, 3-4, 283-290.

[7] Collier, G. 1985. *Emotional Expression*. Lawrence Erlbaum Associates.

[8] Graf, H.P., Cosatto, E., Strom, V., Huang, F.J. 2002. Visual prosody: Facial movements accompanying speech. *Proceedings of Automatic Face and Gesture Recognition*, 396-401.

[9] Hadar, U., Steiner, T.J., Grant, E.C., Clifford Rose, F. 1983. Kinematics of head movements accompanying speech during conversation. *Human Movement Science* 2, 35-46.

[10] Heath, C. 2004. *Body Movement and Speech in Medical Interaction.* Cambridge University Press.

[11] Heylen, D. 2006. Head gestures. gaze and the principles of conversational structure. *Journal of Humanoid Robotics* 3, 3, 241-267.

[12] Hofer, G. and Shimodaira, H. 2007. Automatic head motion prediction from speech data. *Proceedings of Interspeech*.

[13] Ito, K. and Speer, S.R. 2008 . Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language* 58, 2, 541-573.

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

[15] Lee, J., Marsella, S., Traum, D., Gratch, J., Lance, B. 2007. The Rickel Gaze Model: A window on the mind of a virtual human. *Lecture Notes in Artificial Intelligence,* 4722.

[16] Maricchiolo, F., Bonaiuto, M., Gnisci, A. 2005. Hand gestures in speech: Studies of their roles in social interaction. *Proceedings of the Conference of the International Society for Gesture Studies*.

[17] McClave, E.Z. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32, 855-878.

[18] Pelachaud, C., Badler, N.I., Steedman, M. 1969. Generating facial expressions for speech. *Cognitive Science* 20, 1, 1-46.

[19] Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M. 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.