Eye-trackers and Multimodal Communication Studies

Kristiina Jokinen
Institute of Behavioural Sciences
University of Helsinki
kristiina.jokinen@helsinki.fi

Abstract

This article provides an overview of eye-tracking technology in multimodal communication studies. It presents a short review of the human visual perception system and the eye-tracking technology, and discusses two types of eye-gaze studies as examples of how eye-trackers can be used in interaction management: in turn-taking analysis and involvement in conversation.

1 Introduction

The basic function of the eye is to provide visual information from the environment to the perceiving agent. Eye-gaze indicates where the speaker's focus of attention is directed, and it is thus one of the important multimodal feedback signals in human communication. For instance, if the gaze is rapidly wondering around, the person is understood as surveying the environment and collecting information from various points of interest, whereas looking straight at the partner normally signals interest and presence in the interaction with the partner. Gazing also has culturally determined interpretations related to appropriate social behaviour: looking into the partner's eyes can signal the speaker's reliability and truthfulness, although staring at the partner in general can be intimidating for the partner being scrutinized. Looking down can be a sign of humbleness, whereas the gaze wondering around can be interpreted as the person being absent-minded or demonstrating lack of interest in what the partner is presenting which would be considered socially unacceptable behaviour.

Due to the signalling of one's focus of attention, gaze is a powerful indicator of one's cognitive processing. It gives feedback to the partner of the mental efforts and emotions of the speaker, and also indicates the speaker's attitudes towards the partner in a given situation (Cassel et al. 2001). Eye-gaze is also used to control the interaction, as well as to build trust and rapport. Early work by Argyle and Cook (1976) described the role of eye-gaze in turn-taking and introduced the notion of "mutual gaze" for the point in interaction when the partners gaze at each other for a short time to agree on the change of the speaker. Much work on describing the functions and use of eye-gaze in human interactions has been conducted, and the reader is directed to the work e.g. by Cassell et al.1999; Goodwin 1981; Kendon 1990; Streeck and Knapp 1992; Gullberg 1999, among others.

Gazing forms the basis for joint visual attention, and is important when learning social cooperative behaviour. For a child, learning to follow the care-taker's gaze and to understand where their focus of attention is directed to, are important steps in the child's social development and language learning: they enable the child to distinguish "self" from "other" as well as to learn references to salient entities in the shared visual field (Trevarthen, 1984).

Gaze also has a strong cueing effect. Gullberg and Holmqvist (1999) show how interlocutors usually focus their attention on the speaker's face and not on their hands even though hand gestures may be large and peripheral. However, if the speaker fixates their gaze on the hand first, also the partner's gaze follows the hand. The gaze signals to the partner that something has attracted the speaker's attention in the hand movement, and the gaze following is thus automatic in order to maintain joint attention.

Gazing in the partner's face and gaze following seem to be conditioned to social face-

to-face situations. In an experiment Gullberg and Holmqvist (2006) demonstrate how interlocutors who communicate via a small-screen video-conference do not exhibit similar kind of gaze behaviour as those who communicate via a large-screen video-conference or in the presence of a live partner. In particular, they do not look at the partner's face nor follow gaze as often as the participants in the other conditions.

Eye-gaze is also accompanied by other type of eye activity, such as blinking of the eyes or change in the size of the pupil. They are mostly involuntary physiological reactions, but can tacitly indicate the person's cognitive occupation or emotions. For instance, the blinking of the eyes can signal the person's emotional state, and the size of the pupil is related to cognitive load. Facial expressions and eye-muscle movements are also related to gaze quality, and in ordinary language, one often talks about "twinkling" of the eyes, caused by the contraction of eye-muscles when laughing, or describes lack of emotion by the phrase "cold eyes". Along these lines, Poggi (2001) talks about the "alphabets of the eyes", where the shape and form of the eyes and eye-brows function as iconic communicative signals related to the speaker's mental attitudes: squeezed eyes can indicate that the speaker finds the partner's presentation unbelievable or the overall situation difficult or strenuous, while wide eyes usually signal surprise or fear.

Previous work has been mainly descriptive and based on manual analysis of videos. In the past years, eye-trackers have been introduced in the field of human communication studies, and it has become possible to collect more accurate and objective data on the interlocutors' gaze behaviour. Eye-trackers have already been used in medical and clinical research for a long time, but due to developments in the technology, they have improved in robustness and reliability, and also become cheaper, so their use in interaction studies has become feasible and more common.

This article aims to give a short overview of the use of eye-tracking technology in interaction studies. It is not meant to be an exhaustive and systematic overview of the research conducted in the area, but to provide a review of the state-of-the-art in eye-tracking research for understanding interaction and human communication.

The article is structured as follows. In Section 2, the human visual perception system is briefly presented, and in Section 3 an overview of the eye-tracking technology is given. In Section 4, some specific issues related to eye-tracking in the research of human communicative behaviour are discussed, and in Section 5 two research directions of the field are presented: turn-taking analysis and involvement in conversation. Section 6 concludes the article with a discussion of challenges of eye-tracking research.

2 Visual perception system

2.1 Basic concepts

Visual perception includes constant eye activity with *saccades* and *fixations*. The saccades are rapid eye movements when the eyes move simultaneously to the same direction, and the fixations are stops when the gaze is maintained on a single location. The eyes can also follow small moving objects in a manner called *smooth pursuit*, where the fovea is kept steady on the moving object. Another type of eye movement is *vergence movements* when the eyes move to opposite direction so as to adjust the fovea of both eyes to a near object. Seeing occurs only during fixations, and the area of accurate vision (foveal area) is of the size of about 0.3°-2° visual angle. On average there are three fixations per second, their length varying from less than 100ms to about two seconds.

As demonstrated already in the early vision studies (Buswell, 1935), the length of fixation can vary greatly. It is assumed that durations are determined by information processing and cognitive processes that concern interpretation of visual information and recognition of particular objects (Groner and Groner 1989). In other words, seeing is not the same as visual perception: the latter is affected by processing limitations, attention selection, memory capacity etc. and thus always includes interpretation of the visual information. In fact, despite the visual perception being based on discrete and mostly inaccurate seeing, human experience of the surrounding world is continuous and vision accurate. This is because the brain processes visual information by categorizing perceptions of the environment into objects and areas of interest, in a manner which seems to include elements of "problem solving" (Pylyshyn, 1999).

2.2 Eye-gaze and focus of attention

Our experience of the world is based on the things we attend to. Human attention is attracted by various multimodal aspects and features of

the environment, usually related to contrast and change. For instance, unusual shape, unexpected configurations, moving things and surprise appearances catch attention, but also familiarity and conventional forms can be important, especially when learning new skills. It should be emphasised that the focus of attention is not the same as visual attention, and although eye-gaze is commonly used to indicate what the agent is attending to, this can be different from what they are overtly looking at. For instance, looking at the person talking does not necessarily mean that the listener's focus of attention is on the speech: the listener may be attending to something else while directing their overt visual attention on the speaker. Visual attention can also be changed if something suddenly catches attention (ringing of a bell, being called by name, feeling cold, etc.).

One method to estimate the agent's focus of attention is to calculate *saliency maps* (see e.g. Walther and Koch, 2006). Saliency maps use low-level image features such as colour, intensity and orientation to identify high-contrast edges of possible objects of interest, and the bottom-up feature predictions produce output that concerns contrasting areas, e.g. face and background areas. However, eye-tracking experiments show that eye fixations do not coincide with the salience maps as such, but rather with the areas *inside* the salient edges, i.e. humans focus their eyes on the objects which can be perceived by the salient contours. For instance, human faces, eyes in particular, always attract attention, as well as text in images. Visual saliency maps thus need to be augmented by semantic knowledge of the vision scene, while saliency estimation must take into account those objects and activities that are meaningful to the agent.

Information for cognitive processes is selected through visual attention. Human experience is based on attending to salient objects and events in the environment and combining expectations with the actual perceptions via selective attention. Selective attention is a mechanism that is used to serialise the perception of objects in a complex scene. For instance, changes in the world are perceived by attending salient objects and features of the environment. However, perception presupposes interpretation of the visual data, and includes processing limitations, so what is seen is not exactly what is perceived. Simons and Chabris (1999) demonstrate this via selective attention tests. Striking experimental data exists about people's "change blindness", i.e. failure to notice apparent changes in images that are identical but one feature of minor importance (Rensink et al. 1997).

2.3 Visual attention

The manner in which human visual system works is complicated, and matches expectations of salient objects with the actual perception of the world. Current theories of visual attention hypothesise that human vision operates via two visual mechanisms, the global and the local one. They serve different objectives in continuous visual tasks and consequently employ different distribution of saccades and fixations (Unema et al 2005; Pannasch et al. 2011).

The global processing system is in the service of the ambient attention mode, and aims at getting a cursory view of the main regions of interest in the visual scene, whereas the local processing system contributes to the focal mode of attention, and focuses on examining details of the interesting objects. Global processing appears early in the viewing, and is associated with short fixations and long saccades (larger than 5°), so as to scan a larger area with accurate vision, whereas local processing occurs later in the viewing and is characterised by long fixations and short saccades (smaller than 5°), so it is possible to get more information from a particular object of interest. Fixations can thus be classified on the basis of the preceding saccade amplitude: larger amplitudes belong to the ambient attention mode, and shorter ones to the focal attention mode.

2.4 Coherence theory of visual attention

Visual attention on various objects depends on the task that the attention is to serve. The higher-level plans and goals provide targets on which to focus one's attention. The famous work by Yarbus (1967) showed that the eyes move differently in picture inspection depending on the initial goal given to the subject, although there is variation in the individual eye-movements. For instance, compared with free examination of the picture, the task to estimate the age of the people appearing in the picture resulted in a fixation pattern where the subjects focused their attention on the people's faces rather on scanning the whole scene. Task-related cognitive processes seem to control visual exploration of the environment in a top-down fashion.

The main problem in visual cognition is to account for the relation between higher-level

decisions that concern the attention of recognized objects and the visual perception itself: how the objects can be attended before they are recognized. The coherence theory (Rensink, 2000) proposes a solution which models bottom-up attention to salient objects, and uses the notion of *proto-object* to represent possible objects of attention in a salient region of the visual scene. The proto-objects are volatile structures based on bottom-up visual information processing, and they function as constantly re-generated units of visual information. It can be assumed that they function as expectations of the important events and objects in the environment, and they can be selected as the focus of attention depending on how they match with the cognitive requirements.

In computer vision, low-level categorisation of object features is used to produce salience maps within which the proto-objects can be accessed and be validated. These salient regions are used to restrict spatial locations which are likely to contain proto-objects, while the proto-objects can be validated as the actual objects of the scene by selective attention. The spatial location itself functions as an index that links the low-level features into proto-objects across space and time.

So far the visual attention studies have mostly dealt with static visual environments where the eye movement patters have been outlined with respect to a picture on a screen. Mobile eye-tracking technology has brought forward possibilities to study eye movements when the subject is in action, e.g. walking, typing, making tea, playing piano, etc. Two different principles have been identified in the eye-body movement correlation: fixations can focus exactly on the object the agent is engaged with, or they provide information of an action just before the particular action. The studies show that gaze is about one second ahead of the action start, see Land (2006) who gives an overview of the use of eye-gaze in action studies. In other type of tasks, for instance when driving a car, it has been observed that experts anticipate the route about 2-3 seconds ahead, while novices keep their eyes on the road just in front of the car (Sodhi et al., 2002).

In conclusion, we can say that human visual system is a complex mechanism which includes both bottom-up and top-down processes which function in integration. The system provides a means to attend the surrounding world, and maintain coherent experience of it.

3 Eye-tracking technology

As already mentioned, eye-trackers have long been used in medical diagnostic and clinical work. However, technology has developed much from the mechanical eye-trackers used by Huey (1898) to the present-day infra-red light reflection devices with advanced video image processing techniques. Eye-trackers have become more robust and practical, and available for interaction researchers in computer science, social and communication studies. In this article we focus on a short review of the technology only, and refer to Räihä and Majaranta (2007) for a more comprehensive overview of the development of eye-trackers and their use in human-computer interaction research.

Eye tracking refers to measuring where the agent is looking, i.e. their point of gaze. The eye tracker device measures gaze points and eye movement in real time, and reports gaze fixations as scan paths (gaze plots, Figure 1), or heatmaps.

Figure 1Eye fixations and a scan path.

The operation of modern eye-trackers is based on infra-red light reflection from the corneas of the user's eyes. The reflection patterns are collected by image sensors and image processing algorithms are used to identify relevant features, with the help of which gaze point coordinates on the screen can be calculated. Sampling rate is usually 50-120 Hz, which determines the relative accuracy between two consecutive gaze points.

In order to compensate for head movements, two reference points on the eye are needed and the difference in the reflection patterns account for head movements. Usually the pupil centre and the corneal reflection point are used as the reference points. The gaze points used to be measured with respect to head, which requires that the head has to be kept still with the help of a head rest. Modern computer vision-based eye-trackers can take head movements into account, although they still require that the subjects do not

turn their head sideways or move head up-down or back- and forward beyond certain limits. For most table-top trackers an optimum distance from the screen is 50-90 cm, while tolerance to side-turns is less than 20 degrees. Mobile head-mounted trackers or eye-tracking glasses allow subjects to move their head freely as well as walk around. The optics is similar to the table-top trackers except that it is in a miniature form. Measured accuracy is about 0.5 degree, and will always stay in order of 1° visual angle since the exact focus point can be determined only within the foveal area of about 2° visual angle.

Calibration of the tracker with respect to the user's gaze patterns is done before the tracking starts, and sometimes repeated also during long tests so as to compensate possible slight changes. This kind of calibration consists of recording the user's gaze when they are looking at the fixed points on the screen.

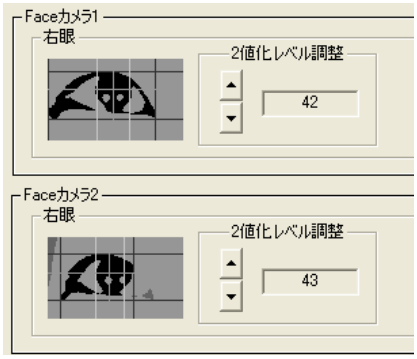

Figure 2 Calibrating the shape of the user's right eye.

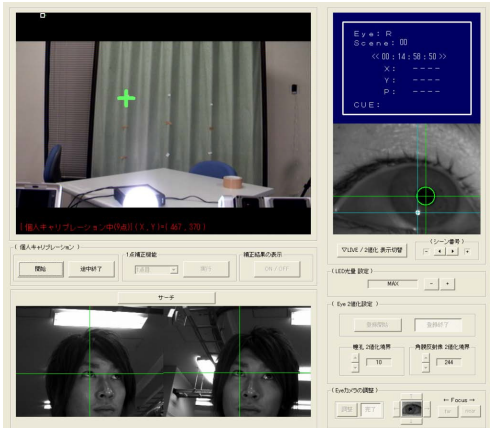

Figure 3 Control panel of an eye-tracker.

Visual information of the user is usually also included into calculations, e.g. facial features as well as eye shape and relative portion of white in the eye. Figure 2 shows the right eye of a user and calibration of the amount of light and dark areas in the eye shape.

Figure 3 shows a control panel of an eye-tracker, featuring the camera views of the visual scene (top left), the user's right and left eyes, and the reflection of the right eye.

4 Eye-tracking and interaction studies

In human communication and interaction research eye-tracking is a useful method as it adds objective information to descriptive observations. It supports analytical approach to estimate where the person is looking at, what they might have perceived, and what has drawn their attention. Moreover, it enables quantitative measures on gaze position, gazing time, and gaze plots, and thus support experimental studies and hypothesis testing on human cognitive processes and communication, e.g. studies on how the task affects gaze behaviour, or how gaze patterns indicate turn-taking. Eye-tracking experiments are also used for usability testing, cognitive load measurements and user evaluation of computer applications. Furthermore, gaze monitoring by an eye-tracker allows development of applications that make use of visual attention: human-computer interfaces for special user groups, computer-mediated communication, and controlling home appliances.

4.1 Metrics in eye-tracking studies

Common metrics used in eye-tracking studies deal with the number and length of fixations, gaze (cumulative duration of consecutive fixations on a particular spatial location), and scan paths (spatial arrangement of a sequence of fixations). Time to the first fixation on the target area of interest can also be useful. Fixations are defined as relatively stable eye positions with some threshold of spatial variation over a minimum duration (typically 100-200 ms). A set of several fixations on the area of interest together with short saccades between these fixations is referred to as "gaze". Jakob and Karn (2010) suggest that gazes are often more meaningful than counting the number of individual fixations. According to them, some authors have also used the term "dwell" in this meaning, although it has not yet become a common term.

Usual measurements include mean and overall number and duration of fixations, commonly measured with respect to particular areas of interest. The areas of interest are defined in advance by the researcher, and in human interaction studies they can include face, certain

(task-related) objects, background, etc., as well as temporal events like turn changes or gesturing.

The number of overall fixations is assumed to have negative correlation with search efficiency (more fixations tell about poor arrangement of the screen or visual scene) but the number should be normalised with respect to the task time, since more complicated tasks are longer and require more fixations. On the other hand, a large number of fixations on a particular object or an area of interest can also signal importance of the object; interpretation depends on the task. Overall fixation rate (the number of fixations in a time unit) is also used; it can signal about the person's emotional state, or about time pressure to learn about the important objects in the scene as quickly as possible.

Longer fixations are usually associated with problems on the particular object (unknown word in reading, complex object on a display), but can also indicate the importance of the object to the user. It has been pointed out that length and frequency may server different purposes: while duration of fixation may reflect difficulty of extraction information, frequency may reflect the importance of object or the area of interest.

4.2 Eye-tracker research

Visual attention studies with the help of eye-trackers started in the 1970's, focussing on experimental investigations on visual perception and cognitive processes, on tasks such as reading texts, searching information, or evaluating image quality. From early on, the typical use of eye-trackers has concerned medical diagnostics and clinical research, while in human-computer interaction and ergonomics, eye-trackers have been used, together with various other bio-physiological devices, to study human reaction, perception, and cognitive load, on complex practical tasks. By measuring the user's gaze patterns and how these differ depending on the user's experience as well as the task and overall layout of the environment, it is possible to get information about the human factors, i.e. about the user needs, skills, and processing constraints, which can help in the design and development of appropriate, efficient, and user-friendly applications. Research on human behaviour in complex tasks often use fairly sophisticated simulation environments, e.g. flight or car simulators, which enable observations in realistic but safe situations, about how various human factors affect the user's control and operation of an application at a particular point in time.

The "applied eye-tracking" deals with interface design and application development where eye-trackers are used to monitor the user's visual attention. This kind of information can then be used to infer the user's intention so as to adapt the application to the needs of the user's and thus serve the user better. For instance, eye-typing interfaces (Hyrskykari et al. 2005) allow the user to input text by focussing on particular letters on the screen, whereas gaze-aware systems aim to anticipate the user's mouse clicks by moving the mouse close to the point where the use's visual attention is located (MIDAS). The European network COGAIN maintains the research activities in this respect, while Räihä and Majaranta (2007) provide an overview of the issues related to gaze-based interfaces. Several workshops and conferences are also associated with the growing interest in eye-gaze studies. For instance ETRA and the series of GAZE-IN workshops (Gaze in Interaction) at ICMI provide annual meetings for studies on gaze and interaction.

5 Eye-gaze in human communication

In human communication and social studies, gaze has been extensively studied, although quantitative measures with eye-trackers are only recently being used in this context. In this paper we will not give an exhaustive and systematic overview of how eye-trackers have been used in human communication studies, but review two directions, where eye-trackers have been used to provide an objective basis for certain human behaviours: the coordination of turn taking and the effect of silent partners in multiparty communication situations.

5.1 Eye-gaze in turn-taking

Gaze is an effective means to coordinate turn-taking and to organize talk: by eye-gaze, the interlocutor can indicate which participant they are addressing their speech to or whether they have understood the speaker's utterance. Besides conversational feedback, eye-gaze is also used to coordinate and control turn-taking: looking at the conversational partner or looking away from the partner provides cues of the agent's willingness to continue interaction (Kendon 1967, Argyle

and Cook 1976, Nakano et al. 2007, Edlund et al. 2004, 2005, 2009).

Earlier studies on spoken interaction have identified several turn taking signals. Acoustic correlates related to distinctive intonation patters have been confirmed in many languages: low or low and falling intonation patterns are associated with turn-yielding and thus suitable turn-taking places, while mid- and high-level intonation patterns indicate turn-keeping and are inappropriate places for turn-taking (Koiso et al. 1998, Noguchi and Den 1998, Edlund et al. 2009). In the absence of boundary tones, also pauses play a role (Wennerström and Siegel 2003). Listeners are likely to wait longer before taking the turn, but the speaker is likely to continue speaking if the pause is longer than 0.3 seconds. After 0.5 seconds, the current speaker was likely to resume talking.

Gaze is a convenient way to convey meaning as it can occur simultaneously with speaking. Simultaneous gazing, or the mutual gaze by the speakers, is important when agreeing on the speaker change (Kendon 1967, Argyle and Cook 1976, Novick et al. 1996, Bavelas 2005, Jokinen et al. 2009, 2010). As described above, gazing at particular elements in the vision field can tell where the speaker's focus of attention is, and this is used in manage turn taking: the speaker who wants to yield the turn, signals this by directing their attention to a potential next speaker, while the partner who is willing to take the turn, focuses their attention onto the current speaker. If this happens simultaneously, the partners can thus synchronise their intentions, and turn taking is possible. Once the partners have visually shared and agreed on the speaker change, the next speaker will start their turn, and also break the mutual gaze by looking away. In fact, in casual conversations, the pressure on the next speaker to speak is so high that uttering nothing is considered extremely rude or it requires an explanation why the listener is not able to react as expected.

Coordination of turn taking in dialogues is often unproblematic since the two partners can fairly easily manage their intentions by gaze in fact-to-face situation. However, in multiparty conversations, gaze is not so reliable since the participants can focus their attention on other than the speaker, and also the speaker need not look at the partner who is willing to talk next. In these cases, head movement functions as an important signal: it is more visible than eye-gaze but still associated with visual attention and fairly reliable in group configurations where the participants need to turn their head to have a straight look at the partner.

In the series of studies by Jokinen et al. (2009, 2010a, 2010b, 2010c), the research centered on questions how eye-gaze affects turn-taking coordination in multiparty conversations, and if eye-gaze can help in predicting turn-taking possibilities. The work is based on the Doshisha Conversational Eye-gaze Data (Jokinen 2010b) which consists of 28 three-party conversations on free topics of interests, among participants who either know each other or are unfamiliar with each other. The corpus was collected using the NAC EMR-AT VOXER eye-tracker, and each conversation is about 10 minutes long, totalling almost 5 hours of data. Figure 1 shows a screen shot of the data.

The study found out that eye-gaze improves the prediction of turn-taking possibilities in spoken conversations, and used together with speech features, it is effectively used to distinguish between two different types of long pauses: those associated with turn-holds and those with turn-change. Long pauses and focussing of gaze on the partner indicate that the partner wants to give the turn to the partner, while gaze aversion during long pauses indicates turn-holds: because of hesitation or need to plan their utterance, the speaker does not focus on the partner and there is no possibility to yield the turn due to the lack of mutual gaze.

When studying the mean and standard deviation of gaze offset related to speech, Levitski et al., (2012) noticed that more gaze activity takes place in the beginning of the utterance than in the middle or at the end of the utterance, the times measured as one second before and after the start or the end of the utterance. The eyes are fixated significantly more often and longer in the beginning than at end of one's utterance, which corroborates with the notion of mutual gaze: the next speaker needs to make sure that the previous speaker indeed agrees to yield the turn and only then can break the mutual gaze, while the previous speaker only needs to scan if the partner is willing to accept the speaker change.

The analysis method advocated in the research is called Multi-level Hybrid Method, and it contains both top-down and bottom-up

techniques. The top-down approach refers to human observation and analysis of conversational phenomena, and uses annotation of the dialogues at the dialogue meaning level. The top-down approach reflects the observer's theoretical view-point and understanding of the phenomena in questions, and is to be validated by the inter-coder agreement calculation so as to reach more objective view or "gold standards".

The bottom-up approach refers to the analysis of the data at the signal level, and uses statistical and machine learning techniques to produce meaning correlations among the data. This can include common data mining techniques related to segmentation and clustering of the data, and can be used to produce meaningful relations. For instance, eye-trackers can be used to provide quantitative data about eye-gaze in various interactive situations which can be automatically analysed.

5.2 Conversational activity

Since turn taking coordination is a matter of the participants' engagement in the conversation, it is interesting to study how the interlocutors' engagement, as measured via their non-verbal activity, influences the other participants' gaze behaviour, and especially how the participants' focus of attention is changed in conversational situations when some of the partners are more active than the others?

In the study of Levitski et al. (2012), conversational activity refers to the interlocutor's general activeness in the interaction. It is defined as an individual speaker's intentional state characterised by energy and liveliness that produces expressive behaviour by speech, gaze, and body. Engagement, on the other hand, refers to the participant's presence in the interaction, and it is measured through their gaze activity. The definition coincides with the notion of entrainment, but is related to gaze.

The experiment used video data where three people discuss about their favourite films, with the subject being eye-tracked and one of the two discussants being naturally active in speaking and making many questions, while the other being naturally more quiet and passive. The Tobii X120 free standing eye tracking device was used in the experiments. Figure 4 is a screen shot of the experimental situation showing the eye-tracked person's eyes fixated at the person on the right.

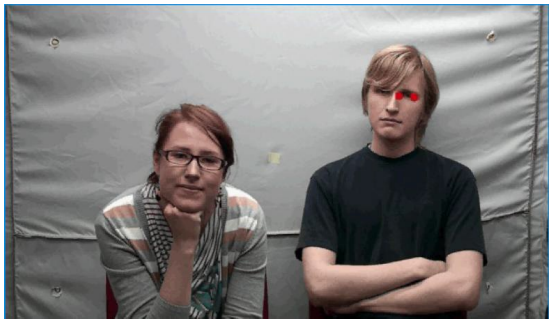

Figure 4 Experimental setup for the conversational activity studies: the active partner is on the left and the silent partner on the right.

In the study, the measurement is gaze activity rather than fixations. Gaze activity is defined as uninterrupted focussing on a particular target, so one token of gaze activity may contain many fixations. As is expected, more gazing is directed to the active partner than to the silent one, and also, the subjects had more gaze activity to both partners when speaking than when listening or backchannelling. When speaking, the subject directs gaze at the active partner, but when they are listening, gaze is divided between the two partners. This confirms the general observation that the participant's own gaze activity is related to speaking, i.e. to a more energetic (active) situation, and that speaking and showing active engagement also attracts the participants' focus of attention.

The experiment also suggests that the silent partner influences the subject's gaze behaviour, and indicates their awareness of the other partner. As expected, the subject's fixation targets and the silent partner's engagement (measured in number of overlapping segments) are correlated: there are more fixations on the silent partner if this is engaged, and if the silent partner is passive, there are more fixations on the other, active partner. However, if the silent partner is passive rather than engaged, the subject gazes at the partner's face less often, but twice as long.

Moreover, it also appears that there are more fixations on the active partner's face when the silent partner is engaged. This gives rise to the hypothesis that the silent partner's activity increases the subject's activity level, since the subject now needs to check the other partner's reaction, too: the increased engagement by the silent partner may cause a reaction in the other

partner as well. The subject is aware of both participants, so in order to keep up-to-date with the whole situation, the subject needs to quickly focus their attention to the other partner as well.

These observations and experimental results confirm the fact that turn-taking is a highly regulated event in the conversations, and that interactions involve social issues that need accurate gaze activity and rapid change of focus of attention so as to be able to manage smooth turn taking.

6 Visual Interaction Management

When looking at images of a person, people look at their faces, especially the area of eyes and mouth, if the faces are big. People are trying to see what social messages there are in the image, and gaze is mainly used for looking and retrieving information. However, in interactive situations, visual attention does not only function as a means to get information about the environment, but it is also a strong signal about communication. As discussed above, gaze can be used to direct the partner's attention to some important aspect of the environment (or distract them from something), and it is also effectively used to coordinate and control the interaction, i.e. there are social rules that regulate attention allocation (see also e.g. Skarratt et al. 2012).

Eye-gaze functions like other multimodal means, such as head nods, hand gestures, and body movement, in enabling the construction of shared understanding among the interlocutors. These means allow unobtrusive signalling of the speaker's conversational status simultaneously with their speaking, and are important in providing feedback about the basic enablements of the communication: whether the partner is willing to be in contact, if they are able to perceive and understand the partner's message, and consequently willing and capable to produce relevant continuation to the interaction. They can all signal the participants' engagement in the interaction.

It is also necessary that the interlocutors are familiar with the non-verbal means and have a similar set of interpretations so that they can be interpreted in the intended way. There are differences in the interpretation of a particular behaviour in different (cultural) contexts, and thus the interlocutors must learn the necessary and important gaze signals in order for the communication to be smooth and efficient. Through gaze studies one can also increase this kind of the awareness in the communication: although gaze patterns often are unconscious and unintentional, the speakers can learn to control them intentionally.

Acknowledgments

Thanks to Seiichi Yamamoto and his staff and students at Doshisha University for collaborating on the research on turn-taking, and to Jenni Radun and students at the University of Helsinki for research on interaction engagement.

References

Argyle, M. and Cook, M. 1976. *Gaze and mutual gaze.* Oxford, England: Cambridge U Press.

Bavelas, J. B. 2005. The two solitudes: Reconciling Social Psychology and Language and Social Interaction. In K. Fitch & R. Sanders (Eds.), *Handbook of Language and Social Interaction* (pp. 179-200). Mahwah, NJ: Erlbaum.

Buswell, G. T. 1935. *How people look at pictures*. University of Chicago Press, Chicago.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C. and Rich, C. 2001. "Non-Verbal Cues for Discourse Structure." *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics,* pp. 106-115. July 17-19, Toulouse, France.

Cassell, J., Vilhjálmsson, H. and Bickmore, T. 2001 "BEAT: the Behavior Expression Animation Toolkit." *Proceedings of SIGGRAPH '01*, pp. 477-486. August 12-17, Los Angeles, CA.

Cassell, J. and Ryokai, K. 2001. "Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling." *Personal Technologies* 5(3): 203-224.

Cassell, J., Bickmore, T., Vilhjalmsson, H. and Yan, H. 2001. "More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment." *Knowledge-Based Systems* 14: 55-64.

Cassell, J. and Bickmore, T. 2001 "A Relational Agent: A Model and Implementation of Building User Trust." *Proceedings of the CHI'01 Conference*, pp. 396-403. March 31-April 5, Seattle, Washington.

Cassell, J., D. McNeill, and K. E. McCullough. 1999. Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition* 7(1):1–34.

Cogain Network for Gaze and interaction studies: http://www.cogain.org/wiki/Bibliography_Gaze_Interaction

Duchowski, A.T. 2003. *Eye-tracking Methodology: Theory and Practice*. Springer

Edlund, J., Skantze, G. and Carlson, R. 2004. Higgins - a spoken dialogue system for investigating error handling techniques- In Proceedings of ICSLP , 2004

Edlund, J., House, D. and Skantze, G. 2005. The effects of prosodic features on the interpretation of clarification ellipses- In *Proceedings of Interspeech* 2005.

Edlund, J., Heldner, M. and Hirschberg, J. 2009a. Pause and gap length in face-to-face interaction. In *Proceedings. of Interspeech 2009*, Brighton.

Edlund, J., Heldner, M. and Pelcé, A. 2009b. Prosodic features of very short utterances in dialogue. In *Proceedings of the Nordic Prosody* 2008, pp. 57-68. Frankfurt am Main.

Goodwin, C. 1981. *Conversational Organization: Interaction Between Speakers and Hearers.* New York, NY: Academic Press.

Groner, R. and Groner, M. T. 1989. Attention and eye movement control: An overview. *European Archives of Psychiatry and Clinical Neuroscience*, 239, 9–16.

Gullberg, M. and Holmqvist, K. 1999. Keeping an Eye on Gestures: Visual Perception of Gestures in Face-to-Face Communication. *Pragmatics and Cognition* 7 (1):35-63.

Gullberg, M. and Holmqvist, K. 2006. What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition,* 14(1), 53-82

Huey, E. B. 1898. Preliminary experiments in the physiology and psychology of reading. *American Journal of Psychology, 9*, 575-586.

Hyrskykari, A., Majaranta, P. and Räihä, K.-J. 2005. From gaze control to attentive interfaces. *Proceedings of HCII 2005*, Las Vegas, NV.

Jakob, R.J.K. and Karn, K.S. 2010. Commentary on Section 4. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises.

Jokinen, K., Nishida, M. and Yamamoto, S. 2009. Eye-gaze Experiments for Conversation Monitoring. *The 3rd International Universal Communication Symposium*, Tokyo, Japan.

Jokinen, K. and M. McTear 2009. *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Jokinen, K. and F. Cheng 2010. *New Trends in Speech-based Interactive Systems*. Springer Publishers.

Jokinen, K. and J. Allwood 2010. Hesitation in Intercultural Communication: Some Observations and Analyses on Interpreting Shoulder Shrugging. In: T. Ishida (Ed.): *Culture and Computing*, LNCS 6259, pp. 55--70. Springer, Heidelberg.

Jokinen, K., K. Harada, M. Nishida and S. Yamamoto 2010a. Turn-alignment using eye-gaze and speech in conversational interaction. *Proceedings of Interspeech 2010*. Makuhari, Japan.

Jokinen, K., M. Nishida and S. Yamamoto 2010b. Collecting and Annotating Conversational Eye-Gaze Data. *Proceedings of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*, Language Resources and Evaluation Conference (LREC-2010). Valetta, Malta

Jokinen, K., Nishida, M. and Yamamoto, S. 2010c. On Eye-gaze and Turn-taking. Proceedings of the Workshop on *Eye-gaze in Intelligent Human-Machine Interaction. International Conference on Intelligent User Interfaces*.

Kendon, A. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*,*26*, 22–63.

Kendon, A, 1990. Signs in the cloister and elsewhere. *Semiotica. 79, 307-29.*

Koiso, H., Horiuchi, Y., Tutiya, S. Ichikawa, A. and Den, Y. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech,* 41(3-4):295–321.

Land, M. F. 2006. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, *25*(3), 296–324.

Land, M. F. 2009. Vision, eye movements, and natural behavior. *Visual neuroscience*, *26*(1), 51–62.

Levitski, A., Radun, J. and Jokinen, K. 2012. Visual interaction and conversational activity. *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality.* Santa Monica, USA

Majaranta, P. and Räihä, K.-J. 2007. Text entry by gaze: Utilizing eye-tracking. In MacKenzie, I. S., and Tanaka-Ishii, K. (Eds.), *Text entry systems: Mobility, accessibility, universality,* pp. 175-187. Morgan Kaufmann

Noguchi, H. and Den, Y. 1998. Prosody-Based Detection of the Context of Backchannel Responses. In Fifth International Conference on Spoken Language Processing.

Nakano, Y. and Nishida, T. 2007. Attentional behaviours as nonverbal communicative signals in situated interactions with conversational agents. In Nishida, T. (Ed.), *Engineering approaches to conversational informatics,* pp. 85-102. John Wiley & Sons

Novick, D., Walton, L. and Ward, K. 1996. Contribution graphs in multiparty conversations, *Proceedings of the International Symposium on Spoken Dialogue* (ISSD-96), Philadelphia, PA, October, 1996, 53-56.

Pannasch, S., Schulz, J. and Velichkovsky, B. M. 2011. On the control of visual fixation durations in free viewing of complex images. *Attention, Perception, & Psychophysics*, Psychnomic Society, Inc. DOI 10.3758/s13414-011-0090-1

Poggi I. 2001. The lexicon and the Alphabet of Gesture, Gaze, and Touch. *Proceedings of the Third International Workshop on Intelligent Virtual Agents (IVA),* p. 235-236. http://dx.doi.org/10.1007/3-540-44812-8_20

Pylyshyn, Z. 1999. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception *Behavioral and Brain Sciences* 22:341–423. Cambridge University Press.

Rensink, R.A., O'Regan, J., Kevin and Clark, J. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* 8 (5): 368–373.

Rensink, R. A. 2000. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3), 17–42.

Simons and Chabris 1999. In youtube: http://www.youtube.com/watch?v=vJG698U2Mvo

Skarratt, P.A., Cole, G.G. and Kuhn, G. 2012. Visual cognition during real social interaction. *Frontiers in human neuroscience*, *6*, 196.

Sondhi, A., O'Shea, J. and Williams, T. 2002 *Arrest Referral: emerging findings from the national monitoring and evaluation programme.* DPAS paper 18. London: Home Office.

Streek, J. and Knapp, M. L. 1992. The interaction of visual and verbal features in human communication. In F. Poyatos (Ed.), *Advances in Nonverbal Communication; sociocultural, Clinical, Esthetic and Literary Perspectives. pp. 3-23.* Amasterdam and Philadelphia: John Benjamins.

Trewarthen, C. 1984. "Emotions in Infancy: Regulators of Contact and Relationships with Persons." Pp. sivunumero, ei löydy!! in *Approaches to Emotion,* edited by K. R. Sherer and P. Ekman. Hillsdale, NJ: Lawrence Erlbaum.

Unema, P. J. A., Pannasch, S., Joos, M. and Velichkovsky, B. M. 2005. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12, 473–494.

Walther, D. and Koch, C. 2006. Modeling attention to salient proto-objects. *Neural Networks* **19**, 1395-1407.

Wennerstrom, A. and Siegel, A. F. 2003. Keeping the Floor in Multiparty Conversations: Intonation, Syntax, and Pause. *Discourse Processes* 36, 77-107.

Yarbus, A. 1967. *Eye Movements and Vision*, Plenum Press, New York.

Yonezawa, T., Yamazoe, H., Utsumi, A. and Abe, S. 2007. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. *Proceedings of the 9th International Conference on Multimodal Interfaces* (ICMI'07), pp. 140-145. New York, NY: ACM.