Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes

Roel Vertegaal
Queen's University
Canada
roel@acm.org

Robert Slagter
Telematics Institute
The Netherlands
slagter@telin.nl

Gerrit van der Veer
Vrije Universiteit
The Netherlands
gerrit@acm.org

Anton Nijholt
Twente University
The Netherlands
anijholt@cs.utwente.nl

ABSTRACT
In multi-agent, multi-user environments, users as well as agents should have a means of establishing who is talking to whom. In this paper, we present an experiment aimed at evaluating whether gaze directional cues of users could be used for this purpose. Using an eye tracker, we measured subject gaze at the faces of conversational partners during four-person conversations. Results indicate that when someone is listening or speaking to individuals, there is indeed a high probability that the person looked at is the person listened (p=88%) or spoken to (p=77%). We conclude that gaze is an excellent predictor of conversational attention in multiparty conversations. As such, it may form a reliable source of input for conversational systems that need to establish whom the user is speaking or listening to. We implemented our findings in FRED, a multi-agent conversational system that uses eye input to gauge which agent the user is listening or speaking to.

KEYWORDS: Attention-based Interfaces, Multiparty Communication, Gaze, Conversational Attention, Attentive Agents, Eye Tracking.

INTRODUCTION
The ability to communicate without words plays an important part in our everyday use of language. Not only do conversational cues such as gestures, facial expressions, looks and tone of voice often determine the meaning of the words we use, such nonverbal expressions may also play an essential role in regulating the conversational process [3]. We are now beginning to see how the lack of support for nonverbal conversational cues may limit the usability of speech recognition and production systems [8]. In this paper, we focus on one particular problem: knowing when the system is being addressed or expected to speak. This problem becomes apparent particularly in multi-agent, multi-user environments, such as our Virtual Theatre [13]. The Virtual Theatre is an animated 3D VRML model of a theatre, in which users can see previews of shows and book tickets through conversational agents. Different agents are used for different queries: to ease contextual knowledge requirements for the system, the embodiment of each agent is used as a metaphor for its functionality. However, the ability to speak to multiple agents means users as well as agents should have a means of establishing who is talking to whom. It has long been presumed that gaze directional cues are an important source of such information in human conversation [16]. In order to verify how well the looking behavior of users predicts their conversational attention, we performed an experiment in which we evaluated where people looked in normal face-to-face group conversations. First, we will discuss previous empirical work, after which we will present hypotheses and details of our experiment. Finally, we will discuss our results and outline our current work towards a multi-agent conversational system that can observe and use gaze directional cues.

SIGCHI'01, March 31-April 4, 2001, Seattle, WA, USA.

PREVIOUS WORK
According to Kendon [10], in two-person (dyadic) conversations, seeking or avoiding to look at the face of the conversational partner (i.e., *gaze*) serves at least four functions [2, 3, 10]: (1) to provide visual feedback; (2) to regulate the flow of conversation; (3) to communicate emotions and relationships; and (4) to improve concentration by restriction of visual input. In the early seventies, Argyle [1, 3] estimated that when two people are talking, about 60% of conversation involves gaze, and about 30% involves mutual gaze (or eye contact). According to Argyle, people look nearly twice as much while listening (75%) as while speaking (41%). The amount of gaze is also subject to individual differences such as personality factors and cultural differences. For example, an extravert may gaze more than an introvert [12]. However, in general, gaze patterns seem closely linked with speech behavior. According to Kendon [10], person A tends to look away as she begins speaking, and starts to look more at her interlocutor B as the end of her utterance approaches. This pattern should be explained from two points of view. Firstly, in looking away at the beginning, person A may be withdrawing her attention from person B in order to concentrate on what she is going to say. When she approaches the end of her utterance, the subsequent action will depend largely upon how person B is behaving, necessitating person A to seek information about her interlocutor. Secondly, these changes in gaze function as signals to person B. In looking away at the beginning, person A signals that she is about to begin speaking, forestalling responses from person B. Similarly, in looking at person B towards the end of her utterance, she may signal that she is now ceasing to talk yet still has attention for him, effectively offering the floor to person B.

Argyle [1] argued that gaze functions only as a *minor* signal for taking turns. According to him, the first explanation is the most important: people look at each other to obtain visual feedback for subsequent actions. Although this argument may hold true for dyadic conversations, one should note that in a multiparty setting, *any* turn-taking cue may function effectively only if information about the addressee is included.

Gaze Patterns in Multiparty Situations

When it comes to managing conversations, multiparty conversational structure is much more complicated than its dyadic equivalent. As soon as a third speaker is introduced, the next turn is no longer guaranteed to be the non-speaker. When the number of participants rises beyond three, it becomes possible to have side conversations between subgroups of people. This can pose problems for the regulation of turn taking. When we consider the above example of a speaker yielding the floor in a multiparty situation, the question arises to whom he or she would like to yield the floor. Although one could explicitly indicate the addressee by naming or pointing to the person, a more nonverbal way of coding such attention-related information would interfere less with verbal communication. As such, we believe gaze behavior would provide an ideal candidate for conveying addressee information. Vertegaal et al. [19] investigated the effect of gaze on triadic mediated conversations. They found a significant positive correlation between the percentage of gaze conveyed and the number of turns taken during conversations. Subjects indicated they were better able to perceive who was talking to whom when the percentage of gaze was higher, i.e., closer to normal. In one of the few (unpublished) studies on gaze patterns in group conversations, Weisbrod [21] found that subjects gazed over 70% of their speaking time, but only 47% of their listening time. Indeed, Kendon [10] attributed this reversal of the pattern observed in dyadic studies to the need to make clear to whom one is speaking. Due to divided visual attention, when addressing a group, a speaker cannot look at all individuals simultaneously. An increase in the overall percentage of gaze would be required to maintain the quality of the speaker's attentive signal.

HYPOTHESES AND PREDICTIONS

In order to verify whether gaze could function as an indicator of conversational attention in multiparty conversations, one would need to compare gaze patterns of an onlooker at conversational partners that are in and out of the focus of his conversational attention. We therefore performed an experiment in which we compared time spent gazing at individuals spoken or listened to with time spent gazing at *others* in four-person group conversations. Our first hypothesis was:

H1*"On average, significantly more time is spent gazing at the individual one listens or speaks to, than at others"*

To make sure Hypothesis 1 would still hold in cases where visual attention is divided, we added a second hypothesis:

H2*"On average, significantly more time is spent gazing at each person when addressing a group of three, than at others when addressing a single individual"*

Given the evidence presented in the previous paragraphs, we predicted that Hypotheses 1 and 2 should hold true. However, this would not provide evidence as to whether gaze is actually *used* as an attentive signal. If speakers use gaze as an attentive signal, one would expect them to maintain the quality of that signal when speaking to larger groups, by compensating for divided visual attention. Weisbrod's finding of an increase in the amount of gaze with group size was confounded by a comparison across listening and speaking behavior. To avoid this we formulated two separate hypotheses, one aimed at evaluating the effect of group size, and the other aimed at evaluating the effect of listening or speaking behavior on gaze:

H3 *"On average, time spent gazing at each individual when addressing a group of three is significantly more than one third of the time spent gazing at a single addressed individual"*

H4*"On average, significantly more time is spent gazing at the individual one listens to, than at the individual one speaks to"*

Given the evidence presented in the previous paragraphs, we predicted that Hypotheses 3 and 4 should hold true.

METHODS

Our experiment applied a within-subjects design in which all variables were measured, rather than controlled. 7 four-person groups discussed current-affairs topics in face-to-face meetings. Subjects participated in four 8-minute sessions: one in which we recorded where they looked using a desk-mounted eye tracker [11], and three in which they were conversational partners only. In each session we registered the mean location of the each conversational partner's face, as outlined below. During analysis, this allowed us to establish, for any moment in time, whether the tracked subject was looking at the face of a conversational partner. We also registered speech activity of all discussants using microphone headsets, and asked the tracked subject to score his conversational attention while watching a video registration after the session. During analysis, we combined speech data with the tracked subject's conversational attention scores. This allowed us to establish not just whether the subject was listening or speaking, but also *whom* she was listening or speaking to at any given moment in time. Finally, by combining gaze analysis data and conversational analysis data, we could calculate the percentage of time spent by tracked subjects gazing at each partner while speaking or listening to that partner, while speaking or listening to other partners, and while speaking to everyone.

Figure 1. Conversational partners as seen from a camera located above the subject's head.

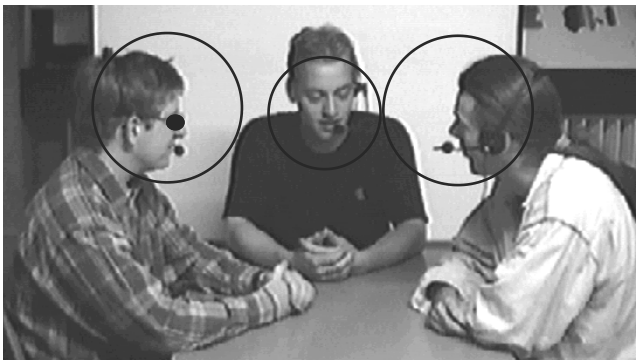

Figure 2. When subject fixations (the black dot) hit one of the circles, gaze was registered for that person.

Experimental Details

All subjects were paid volunteers, mostly university students from a variety of technical and social disciplines. Prior to the experiment, we tested all subjects on eyesight, personality, and their ability to operate the eye tracking system. We allocated each subject to a discussion group in a way that matched groups on average of personality score (extraversion), age, and sex composition. The tracked subject was seated behind a table in a chair with a very comfortable neck support, which effectively removed the need for head movement. Conversational partners were seated around the same table at distances of 1 to 2 meters from each other (see Figure 1). Care was taken there were no potential objects of interest on the table.

Next, we will discuss the details of our measurement methodology, and the subsequent analysis of our data.

ANALYSIS

Results were calculated over 24 sessions, with 5 female and 19 male subjects. Only the last 5 minutes of each session were analyzed. We used automated analysis only, verified by human observers. All measurements were corrected for system lag and synchronized using time code signals.

Analyzing Eye Movement Recordings for Gaze

We analyzed eye movement registrations of tracked subjects for gaze at the facial region of conversational partners. At the start of each session, we determined the mean center of gravity of the conversational partners' faces by tracking subject fixations at the eyes of each partner. During this procedure, we asked partners to successively look at: (1) the person on her right-hand side; (2) the person in front of her; (3) the person on her left-hand side; (4) the table and ceiling. For each orientation of a partner's head, we measured approximately 25 fixation position samples, yielding a minimum of 100 samples per center of gravity. We then fitted the largest possible non-overlapping circles around the mean eye location of each partner to constitute the facial region boundaries (see Figure 2). The radii of these circles were calibrated for the distance at which the corresponding partner was seated. Our subsequent automated analysis procedure was straightforward: it registered *gaze* for a conversational partner whenever the tracked subject fixated within the circle around the facial region of that partner (see Figure 2). In the example in Figure 2, gaze is registered for the left conversational partner, since the black dot (indicating the location of a subject fixation) is within the left circle. During analysis, saccades were skipped, except when they occurred in between fixations within the same facial region, in which case they registered as gaze. This way, we compensated for eye movements during iterative fixations at the left eye, right eye and mouth of the same partner.

Accuracy of Eye Movement Registration

Before each session, we calibrated the eye tracker by asking the tracked subject to fixate on nine pre-determined positions, successively projected as dots on a video screen behind the discussion table. After calibration, the system calculated the match between the grid of fixation points and the grid of pre-determined positions as a weighted error. The calibration procedure was repeated until the error leveled below .45°. The virtual plane in which fixations were measured was 2.6 m wide and 1.68 m tall, with a spatial measurement resolution of 4.9 pixels/cm^2. The mean bias error of fixation points was 2.4 cm in this plane, less than the distance between the eyes of the conversational partner seated furthest away. Fixations were recorded with a temporal resolution of 120 ms, roughly the minimum human fixation time and well within the range of the eye tracker [11, 15]. The mean latency of fixation registration was .46 s. This latency was subtracted during analysis. The eye tracker lost track of the eye approximately 1% of total measurement time. Analysis showed that most of these cases were due to blinks and extreme looking behavior of the subject. We therefore treated these cases as not looking at a conversational partner, with one exception. When the system lost track in between fixations within the same facial region, this was due to a blink and treated as gaze.

Accuracy of Automated Gaze Analysis

A human observer checked the results from the automated gaze analysis procedure by reviewing the video of each session with the fixation position of the subject, the circles around the facial regions of partners, and the results from the gaze analysis algorithm superimposed. An error could be due to one of two reasons: the subject fixated within the circle but outside the actual facial region of a partner, or the partner had moved his facial region outside the circle. No errors were detected.

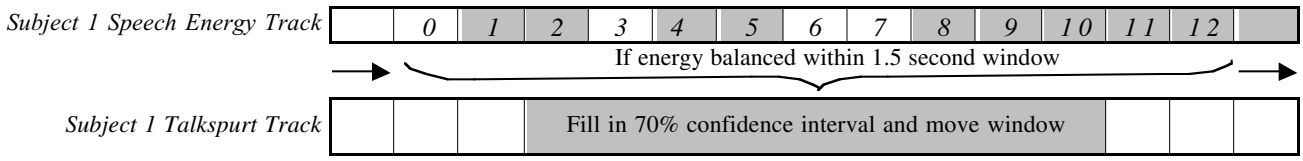

Figure 3. *A graphical illustration of talkspurt analysis of one subject's speech energy recordings. The procedure counts the number of samples with a value true (indicated as gray boxes) in a 13-sample window (1.5 seconds) of the speech energy track. If this total is less than 7, the procedure does nothing and shifts one position ahead in time. If it is greater, the mean position of samples with a value true within the window is calculated. In the above example, this amounts to: (1+2+4+5+8+9+10+11+12)/9 = 6.89. If the mean value is between 5 and 7 inclusive, it decides these samples are evenly spread over the window. It now sets all samples in the talkspurt track between 2 and 10 to true (as indicated by the large gray box). Then, the window shifts one position ahead in time.*

Analyzing Speech Recordings for Utterances

The speech energy produced by each discussant was registered with a temporal resolution of 120 ms. We analyzed the speech energy recordings to determine, for any moment in time, whether a discussant was speaking or not. It is not evident that a registration of the energy produced during speech activity is a good indicator of what we consider to be an *utterance*: a string of words produced by the same speaker. This is because throughout the articulation of speech, a speaker may introduce various moments of silence: between phonemes, between words, and between strings of words. During analysis, such silences had to be treated as part of the utterance.

We designed a fuzzy algorithm which analyzed the speech activity of all discussants simultaneously, designating moments of silence and moments of utterance activity by a single speaker holding the floor. First, our algorithm filled in 240 ms pauses to account for stop consonants, effectively removing pauses within words [7]. Then, the algorithm removed pauses between consecutively spoken words to identify *talkspurts*, series of phonemic clauses uttered by the same speaker. The phonemic clause is regarded as a basic syntactic unit of speech. On average, it consists of 2-10 words with a duration of approximately 1.5 s, providing an estimate for finding the shortest uninterrupted vocalizations (see [9, 14] for a discussion). To identify talkspurts, a 13-sample (1.5 s) window moved over the speech data, filling samples within a 70% confidence interval around its mean position with speech energy if more than half of the samples in the window indicated speech activity, and if this speech activity was balanced within the window (see Figure 3). Finally, an utterance was assigned if one of the speakers had a talkspurt longer than an average phonemic clause (i.e., 1.5 s), with everybody else being silent for the same length of time. This utterance would end with a speaker switch, i.e., when the current speaker would fall silent and a new speaker would produce an utterance. Note that although we used the pause that occurs before a speaker switch to identify utterance boundaries, we did not consider this pause as part of any utterance. Also note that our algorithm treats side conversations erroneously. The few errors this produced were marked by hand and skipped during verification.

Reliability and Validity of Utterance Analysis

The results of the analysis process were evaluated by the experimenter by superimposing the results of the utterance analysis algorithm for each partner onto the real-time video image of each session. The experimenter could indicate an error by pressing a key for the duration of that error. Errors were typically due to speakers retaining the floor during joint laughter or side conversations. Of each session, an average of two 5-second periods were skipped from analysis as a result of this review process (3.3% of session time). We also checked the validity of the utterance analysis algorithm by calculating the correlation over time between an utterance classification produced by the algorithm and one produced by a trained linguist (see [17] for details). With a correlation of $r=.64$ ($p<.001$, 2-tailed) between classification methods the algorithm, which identified phonemic clauses simply by checking the duration of consecutive speech, did well against the human expert, who used intonation and semantics of speech to identify phonemic clauses.

Measuring Conversational Attention

After each session, the tracked subject was asked to watch a video of the last five session minutes in order to score whom she had spoken or listened to at the time. By combining these scores with utterance data, we could reconstruct which partner the subject was listening or speaking to at any given moment of time. Subjects were seated behind a screen in the control room showing a video registered from the approximate point and angle of view of the subject (see Figure 1). When the subject thought she had been listening or speaking, she would press one or more keys of an accord keyboard to indicate which partner(s) she had listened or spoken to.

Reliability and Validity of Conversational Attention Scores

Before scoring, subjects were trained and carefully instructed on the task. During this training, subjects scored a video of a session in which four actors played the role of subject and partners according to a script that specified the conversational attention of each actor. Training scores were compared with the actual enactment of this script as scored by the experimenter. The agreement of scores was calculated as a percentage of time in which scores overlapped exactly. Before being admitted to the scoring procedure, each subject had to reach a 60% agreement with the pre-specified score for at least 45 consecutive seconds.

After scoring, we also asked subjects to indicate any mistakes while reviewing the video with their scores superimposed.

Variable	*Listening to individual*	*Addressing individual*	*Addressing all three*
Gaze at individual	62.4	39.7	19.7
	(3.8)	*(4.7)*	*(1.8)*
Gaze at others	8.5	11.9	
	(1.2)	*(2.4)*	
Gaze at all three			59.0
			(5.4)

Table 1. *Means and std. errors for percentages of time spent by subjects gazing at partners in the last 5 session minutes.*

All data for which subjects scored a mistake was skipped from analysis. Finally, we asked subjects to perform a simple stimulus-response test that was an abstraction of the scoring task. During analysis, we corrected timing of all scores for the mean subject response time.

RESULTS

For each session and for each conversational partner, we calculated the mean percentage of time in which the tracked subject fixated his gaze within the facial region of that partner while:

1) *Speaking to that partner*, which was true when the subject scored conversational attention for that partner while the subject had an utterance.
2) *Listening to that partner*, which was true when the subject scored conversational attention for that partner while that partner had an utterance.
3) *Speaking to others*, which was true when the subject had an utterance but scored conversational attention for someone else.
4) *Listening to others*, which was true when another partner had an utterance for which the subject scored conversational attention.
5) *Speaking to all three*, which was true when the subject scored conversational attention for all partners while the subject had an utterance.

We averaged percentages across partners and subjects to compile the resulting statistics, presented in Table 1. All data was normally distributed (Kolmogorov-Smirnov test, $p>.05$). All planned comparisons were carried out using 1-tailed paired t-tests, evaluated at $\alpha=.05$.

Subjects gazed approximately 7.3 times more at the individual listened to (62.4%), than at others (8.5%) ($t(23)=12.92$, $p<.001$, 1-tailed). They also gazed 3.3 times more at an addressed individual (39.7%), than at others (11.9%) ($t(23)=5.2$, $p<.001$, 1-tailed), thus confirming Hypothesis 1. Subjects gazed approximately 1.7 times more at an individual when addressing all three (19.7%), than at others when addressing a single individual (11.9%) ($t(22)=2.71$, $p<.01$, 1-tailed), thus confirming Hypothesis 2.

Time spent gazing at an individual when addressing all conversational partners (19.7%) was approximately 1.5 times more than time spent gazing at a single addressed individual divided by three (39.7 / 3 = 13.2%) ($t(22)=-4.47$, $p<.001$, 1-tailed), thus confirming Hypothesis 3. Note that this is equivalent to comparing gaze at all three while addressing all three (59%) with gaze at an individual while addressing an individual (39.7%). Subjects gazed approximately 1.6 times more at an individual listened to (62.4%), than at an addressed individual (39.7%) ($t(23)=5.49$, $p<.001$, 1-tailed), confirming Hypothesis 4.

DISCUSSION

The main objective of our study was to verify whether looking behavior of users predicts whom they are speaking or listening to. We will now discuss the results from our experiment for each of our hypotheses.

Hypothesis 1: People Look More at the Person They Speak or Listen to Than at Others

To verify our first hypothesis, we compared the time spent gazing at individuals spoken or listened to with the time spent gazing at others during group conversations. We found that on average, subjects indeed gazed significantly more at individuals for which they had conversational attention. They did this both while listening as well as while speaking to individuals. Our results indicate that when someone is listening to an individual, there is an 88% chance ($\approx$ 7:1 ratio) that the person gazed at is the person listened to. When someone is speaking to an individual, there is a 77% chance ($\approx$ 3:1 ratio) that the person gazed at is the addressed individual. As such, we can consider gaze behavior to be an excellent predictor of the user's conversational attention.

Hypothesis 2: Listeners in a Group Can Still See They Are Being Addressed

With regards to our second hypothesis, we verified whether users can still see they are being addressed in cases where visual attention of the speaker is divided amongst three listeners. For this, we compared the time spent gazing at individuals while speaking to all with the time spent gazing at others while speaking to a single individual. Although levels of gaze per individual drop when addressing larger groups of three, each person still receives 1.7 times more gaze than could be expected had he not been addressed. This means the predictive function of gaze seems to be preserved when visual attention is divided. The reason for this is that speakers start looking more at their listeners in such cases, leading us to confirm our third hypothesis.

Hypothesis 3: Speakers Compensate for Divided Visual Attention

On average, time spent gazing at each individual when addressing a group of three is indeed almost 1.5 times more than would be the case if visual attention of the speaker would simply be divided by three. In fact, the total amount of gaze rises dramatically when addressing a triad to about the level (59%) of gaze while listening (62%).

Figure 4. *User interacting with FRED. An eye tracker camera mounted below the computer screen is used to determine which agent the user gazes at.*

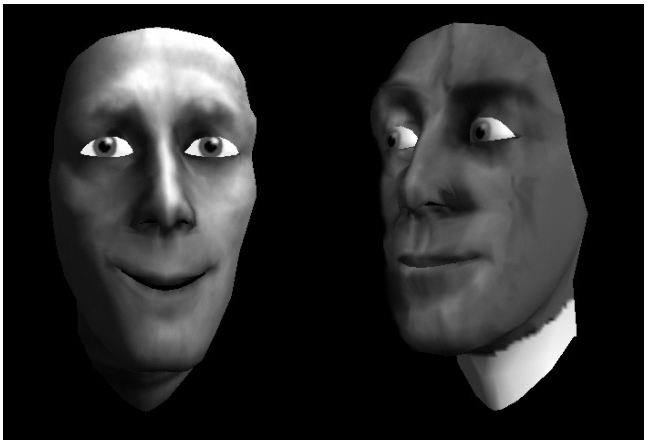

Figure 5. *FRED prototype with two conversational agents.*

Literature suggests three reasons why speakers look more when addressing larger groups:

1) *Visual Feedback.* It is evident that visual feedback of the listeners' responses is a predominant reason to gaze [6]. Since, when addressing triads, there is less time to collect feedback on each addressed individual, speakers would need to gaze more.
2) *Communication of Conversational Attention.* In multiparty situations, speakers may use gaze to signal whom they are addressing. When speaking to larger groups, they would need to gaze more to maintain this signal. This rationale is consistent with Kendon's explanation of Weisbrod's findings [6, 10, 21].
3) *Regulation of Arousal.* Argyle and Cook [3] suggested that mutual gaze (eye contact) has an effect on arousal. By avoiding or seeking gaze, speakers may attempt to maintain arousal of themselves and the addressed individuals at a mutually satisfactory level [4]. When addressing triads, speakers would need to gaze more to maintain sufficient eye contact.

The above discussion means we cannot consider the increase in gaze by speakers as conclusive evidence that gaze is in fact purposefully used to communicate conversational attention. With regards to this discussion, our final hypothesis provides some interesting insights.

Hypothesis 4: Listeners Gaze More Than Speakers

To verify our fourth hypothesis, we compared the time spent gazing at an individual listened to with time spent gazing at an addressed individual. We found that on average, subjects indeed gaze 1.6 times more while listening than while speaking, which is consistent with Argyle's findings in dyadic situations [1]. According to Argyle and Cook [3], when preparing their utterances, speakers need to look away to avoid being distracted by visual input (such as prolonged eye contact with a listener). This corresponds to our finding that the gaze of speakers is somewhat less predictive of conversational attention than the gaze of listeners. However, our results also show this is only the case when speaking to a single individual. Qualitative observations of video registrations showed that when addressing a group, speakers used an iterative pattern of alternating fixations on listeners. We believe that speakers need not look away when addressing a group because they can easily avoid prolonged eye contact by looking at other people. This, in turn, is evidence for the signaling function of gaze: when addressing a single person, speakers have to avoid gaze at other conversational partners to avoid signaling they are being addressed.

Gaze as a Predictor of Conversational Attention

Overall, our results mean that the user's eye gaze can form a reliable source of input for conversational systems that need to establish whether the user is speaking or listening to them. However, the predictive power of gaze signals may depend on the individual user and the visual design of the conversational system. Firstly, we found considerable individual differences in looking behavior, with standard deviations in gaze time of 19% while listening and 23% while speaking to individuals. Secondly, our findings pertain to a pure communication situation only. Argyle and Graham [5] found that if a pair of subjects was asked to plan a European holiday and there was a map of Europe in between them, the amount of gaze dropped from 77 percent to 6.4 percent. However, they also suggested that the predictive function of gaze might be transformed into a more generic indicator of joint interest: subjects were keeping in touch by looking at the same object. Even so, faces in general, and the eyes in particular, are powerful attractors of visual attention during conversations. Qualitative observations of fixations superimposed onto video recordings of sessions showed that even during periods of heavy gesticulation by an attended conversational partner, subjects would typically fixate on the face, rather than the hands. Within the face, subjects tended to produce iterative fixations on the left eye, right eye and mouth. This suggests that when using gaze as a means of managing turn taking in conversational systems, an anthropomorphic design of the system may be beneficial.

Next, we will summarize our work towards the application of our findings in a multi-agent conversational system.

APPLICATIONS: FRED, AN ATTENTIVE AGENT

We are currently working to implement our findings in FRED [18], a multi-agent conversational system that establishes where the user looks by means of an eye tracking system mounted below the computer screen (see Figure 4). In FRED, multiple conversational agents can be embodied by means of realistic 3D texture-mapped models of human faces. Based on work by Waters [20], muscle models are used for generating accurate 3D facial expressions. The system uses our SCHISMA speech recognition and production engine to converse with the user [13]. Each agent is capable of detecting whether or not the user is looking at it, and combines this information with speech data to determine when to speak or listen to the user. Agents use a real-time version of our utterance analysis algorithm to predict when a user has a turn. To help the user regulate conversations, agents generate gaze behavior as well. This is exemplified by Figure 5. In this example, the agent speaking on the left is the focal point of the user's eye fixations. The right agent observes that the user is looking at the speaker, and signals it does not wish to interrupt by looking at the left agent, rather than the user.

CONCLUSIONS

In this paper, we focused on one particular problem of speech recognition and production systems: knowing when the system is being addressed or expected to speak. This problem becomes apparent particularly in multi-agent, multi-user environments, where users as well as agents should have a means of establishing who is talking to whom. We presented an experiment aimed at evaluating whether gaze directional cues of users could be used to indicate their conversational attention, i.e., whom they listen or speak to. Using an eye tracker, we measured the subjects' gaze at the faces of conversational partners during four-person conversations. On average, subjects looked about 7 times more at the individual they listened to (62%), than at others (9%). They looked about 3 times more at an individual they spoke to (40%), than at others (12%). We conclude that gaze, or looking at faces, is an excellent predictor of conversational attention in multiparty conversations. When someone is listening to an individual, there is an 88% chance that the person gazed at is the person listened to. When someone speaks to an individual, there is a 77% chance that the person gazed at is the addressed individual. In this essentially dyadic situation, speakers gazed about 1.6 times less than listeners, presumably to avoid distraction by eye contact with the listener. However, when addressing a group of three, we saw speaker gaze rise dramatically to about the level of gaze while listening (59%). Although levels of gaze per individual dropped significantly in such cases, each listener still received 1.7 times more gaze than could be expected had he not been addressed. Speakers need not look away when addressing a group because they can easily avoid prolonged eye contact by looking at other addressees. This, in turn, is clear evidence that gaze is in fact used to signal conversational attention in conversations: when addressing a single individual, speakers have to avoid gaze at other conversational partners to avoid signaling they are being addressed. Overall, our results mean that the user's eye gaze can form a reliable source of input for conversational systems that need to establish whom the user is speaking or listening to. However, the predictive power of gaze signals may depend on the individual user and the visual design of the conversational system. For example, when using gaze as a means of managing turn taking in conversational systems, an anthropomorphic design of the system may be beneficial. We demonstrated how this might be implemented in a multi-agent conversational system that can observe and use gaze directional cues.

ACKNOWLEDGEMENTS

This work was supported by a grant from NSERC of Canada. We thank Martijn Polak and Eelco Herder for their work on FRED, Nancy and Dixon Cleveland of LC Technologies for their support, and Harro Vons, Bert Lenting, Herman Adèr, Jolijn Hendriks, Luuk Lagerwerf as well as the members of the former Ergonomics Department at Twente University for their contributions. Thanks also to Nancy Barker, Kevin Brewer and Boris Velichkovsky.

REFERENCES

1. Argyle, M. *The Psychology of Interpersonal Behaviour*. London: Penguin Books, 1967.
2. Argyle, M. *Social Interaction*. London: Tavistock Publications, 1969.
3. Argyle, M. and Cook, M. *Gaze and Mutual Gaze*. London: Cambridge University Press, 1976.
4. Argyle, M. and Dean, J. Eye-contact, Distance and Affiliation. *Sociometry* 28, 1965, pp. 289-304.
5. Argyle, M. and Graham, J. The Central Europe Experiment - Looking at Persons and Looking at Things. *Journal of Environmental Psychology and Nonverbal Behaviour* 1, 1977, pp. 6-16.
6. Argyle, M., Lalljee, M., and Cook, M. The Effects of Visibility on Interaction in a Dyad. *Human Relations* 21, 1968, pp. 3-17.
7. Brady, P.T. A Statistical Analysis of On-off Patterns in 16 Conversations. *The Bell System Technical Journal* (Jan.) 1968, pp. 73-91.
8. Cassell, J., Bickmore, T., *et al*. Embodiment in Conversational Interfaces: Rea. In *Proceedings of CHI'99*. Pittsburgh, PA: ACM, 1999.
9. Jaffe, J. and Feldstein, S. *Rhythms of Dialogue*. New York: Academic Press, 1970.
10. Kendon, A. Some Functions of Gaze Direction in Social Interaction. *Acta Psychologica* 32, 1967, pp. 1-25.
11. LC Technologies, Inc. *The Eyegaze Communication System*. Fairfax, VA: *http://www.eyegaze.com*, 1997.
12. Mobbs, N.A. Eye-contact in Relation to Social Introversion-Extraversion. *British Journal of Social Clinical Psychology* 7(305-306), 1968.
13. Nijholt, A. and Hulstijn, J. Multimodal Interactions with Agents in Virtual Worlds. In *Future Directions for Intelligent Information Systems and Information*

Science, N. Kasabov (ed.), Physica-Verlag: Studies in Fuzziness and Soft Computing 45, 2000, pp. 148-173.

14. Sellen, A.J. Speech Patterns in Video-Mediated Conversations. In *Proceedings of CHI'92.* Monterey, CA: ACM, 1992, pp. 49-59.
15. Velichkovsky, B. and Hansen, J.P. New Technological Windows to Mind: There is More in Eyes and Brains for Human Computer Interaction. In *Proceedings of CHI'96.* Vancouver, Canada: ACM, 1996, pp. 496-503.
16. Vertegaal, R. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In *Proceedings of CHI'99.* Pittsburg, PA: ACM, 1999, pp. 294-301.
17. Vertegaal, R. *Look Who's Talking to Whom*. PhD Thesis. Enschede, The Netherlands: Cognitive Ergonomics Department, Twente University, 1998.
18. Vertegaal, R., Slagter, R., Van der Veer, G.C., and Nijholt, A. Why Conversational Agents Should Catch the Eye. In *Extended Abstracts of CHI'2000.* The Hague, The Netherlands: ACM, 2000, pp. 257-258.
19. Vertegaal, R., Van der Veer, G.C., and Vons, H. Effects of Gaze on Multiparty Mediated Communication. In *Proceedings of Graphics Interface 2000.* Montreal, Canada: Morgan Kaufmann Publishers, 2000, pp. 95-102.
20. Waters, K. and Frisbee, J. A Coordinated Muscle Model for Speech Animation. In *Proceedings of Graphics Interface'95.* Canada, 1995, pp. 163-170.
21. Weisbrod, R.M. Looking Behavior in a Discussion Group. Unpublished paper, Department of Psychology, Cornell University, 1965.