Conversing with the User Based on Eye-Gaze Patterns

Pernilla Qvarfordt
Dept. of Computer and Information Science
Linköpings universitet
SE-581 83 Linköping, Sweden
perqv@ida.liu.se

Shumin Zhai
IBM Almaden Research Center
San Jose, California
USA 95120
zhai@us.ibm.com

ABSTRACT
Motivated by and grounded in observations of eye-gaze patterns in human-human dialogue, this study explores using eye-gaze patterns in managing human-computer dialogue. We developed an interactive system, iTourist, for city trip planning, which encapsulated knowledge of eye-gaze patterns gained from studies of human-human collaboration systems. User study results show that it was possible to sense users' interest based on eye-gaze patterns and manage computer information output accordingly. Study participants could successfully plan their trip with iTourist and positively rated their experience of using it. We demonstrate that eye-gaze could play an important role in managing future multimodal human-computer dialogues.

ACM Classification Keywords
H.5.2 [**Information interfaces and Presentation**]: User Interfaces – *Input devices and strategies*

Author Keywords
Multimodal interaction, dialogue systems, eye tracking, interest detection

INTRODUCTION
Being "a window to the mind," the eye and its movements are tightly coupled with human cognitive processes. The possibility of taking advantage of information conveyed in eye-gaze has attracted many researchers in human-computer interaction [3, 11, 27, 29]. Most of the work in this area to date are in three interrelated categories: eye-gaze as a pointing mechanism in direct manipulation interfaces (e.g. Ware [27], Jacob [11], and Zhai *et al* [29]), eye-gaze as a disambiguation channel in speech or multimodal input (e.g. Tanaka [21], Kaur *et al* [12] and Zhang [30], see also Oviatt [16] for mutual disambiguation in general) and eye-gaze as a facilitator in computer supported human-human communication and collaboration (e.g. Velichkovsky [24] and Vertegaal *et al* [25]).

As a new step in utilizing eye-gaze in HCI, the current work is motivated by the following combination of propositions. 1. The eye-gaze contains richer and more complex information regarding a person's interest and intentions than what is used in pointing [29]. 2. Eye-gaze can be one of the critical channels in future multimodal systems. 3. Eye-gaze can be particularly useful in human-computer dialogue interfaces rather than direct manipulation interfaces. A particular challenge in developing computer dialogue systems lies in taking human-computer conversation beyond pre-defined scripts and adapting the conversation to the user's interest. Sensing where and how the user's eye-gaze moves when the dialogue subject is related to spatial content (such as maps) may provide contextual clues for the computer system to start and manage a dialogue with the user.

GAZE PATTERNS IN HUMAN-TO-HUMAN DIALOGUE
To investigate whether and how eye gaze can be used in managing multimodal human-computer dialogue systems, we first studied human-human conversation as a foundation of the current research. Previous work on the relationship between eye gaze and conversation mainly focused on patterns in face-to-face conversation concerning issues such as when and how long people look at each other [13, 26]. In conversations involving visual spatial information, such as a map, people spend most of the time looking at the spatial information rather than their partner [2].

In order to form a more direct foundation to our current work on using eye gaze to mediate human-computer dialogue, we developed and experimented with a system, RealTourist, which allowed a tourist to talk to a remote tourist consultant to plan a conference trip. The tourist and the consultant saw the same map displayed on their monitors respectively. On the consultant's side the system also superimposed the tourist's eye gaze onto the map, so the consultant could use it to determine the tourist's interests. In an experiment that involved 12 tourists and two tourist consultants, the tourist's eye gaze in relation to the trip planning conversation were collected, inspected, annotated, visualized, and analyzed. For more details of the RealTourist study and its related literature, see [17].

The RealTourist experiment can be viewed as a simulated ("Wizard of Oz") human-computer dialogue study. The results clearly showed that when conversing with another agent about spatial information the user's eye movement on the map was tightly coupled with the dynamics of the conver-

sation. Various functions of the eye-gaze were identified in the study including implicit deictic referencing, interest detection by the consultant, bringing common focus between the tourist and the consultant, increasing redundancy in communication, and facilitating assurance and understanding.

In particular, the study revealed two types of eye-gaze patterns related to a person's interest in an object on the screen (e.g. a hotel, a restaurant etc). Two patters indicated interest in the current topic of the conversation. One of these was that the person looked at the object with high intensity and long accumulated duration. Sometimes a person exclusively looked at a place (e.g. a museum) on a map and its photo when intensely interested in it. In the second pattern that indicates the tourist's interest in the current topic of conversation, the tourist not only looked at the object of the topic (e.g. a hotel), but also at objects that were related to the current topic in some way. For example, when considering a particular attraction, e.g. the Museum of Modern Art, the person might look back at his or her hotel location to figure out the distance relationship. In this pattern, the person often returned the eye gaze to the object that was the focus of the conversation (focus object). Similarly, disinterest could be characterized by two different eye-gaze patterns. The first was to completely leave the focus object, and when the person found a new object, he or she asked about the new object almost immediately. The second pattern was similar, however here the person did not immediately ask for the new object. Instead he or she kept an eye on the new object by returning the eye gaze to it while continuing to explore new possibilities until there was an opportunity to change the topic of the conversation to the new object.

We have also found a gaze pattern indicating interest in the relationship between two objects. Before and while people asked for the distance between the focus object and another object, they switched the eye gaze between the focus object and the other object frequently.

In addition to the RealTourist study, the current research is also based on conclusions from other gaze and speech studies in different but relevant settings. One of them is that when exposed to a visual stimulus and asked to talk about or listen to a story about it, people tend to look at the part of the stimulus that is relevant to the discourse. For example, Cooper found that people look at objects that are relevant to what they listen to [5]. In his study when participants heard the word "lion," they looked at the lion in the picture. When they heard the word "Africa," they looked at the lion, the zebra and the snake. Later work has shown that the latency from the onset of the word referring to a specific object and when the eye looks at the picture of it is around 400-800 ms [1, 22, 23]. When people refer to an object, they look at an object around 900 ms before they refer to it [8].

ITOURIST – AN EXPERIMENTAL SYSTEM

To explore if eye-gaze information can provide useful information to multimodal dialogue systems, we developed an experimental tourist information system, iTourist, based on some of the findings summarized in the last section. iTourist provides the user with tourist information about an imaginary city, Malexander, in the form of a map and photos of different places. Information about the different places is presented by pre-recorded, synthesized speech. iTourist attempts to adapt its information output based on the user's interests and needs analyzed from the user's eye-gaze pattern. Although our ultimate goal is to make eye gaze an integrated channel of multimodal systems including speech, gesture, and eye-gaze, iTourist was developed as a *stress* test to investigate how much information can be gained from a user's eye-gaze alone. iTourist does allow the user to use a mouse, but only as a backup channel when necessary (Figure 1).

Figure 1. User in front of iTourist.

iTourist contains information about 36 places shown on a map, including hotels (10), restaurants (10), attractions (8), nightclubs (5), a conference centre, a bus terminal and a tourist information office (Figure 1). Each object has a number of spoken utterances (6 for each hotel or restaurant, 9 for each attraction) and images (5 for each hotel or restaurant, 8 for each attraction). When displayed, the images are connected with the place's location on the map with an arrow line. Synchronized with the changing photos displayed, the information utterances (sentences) for each place are played in a pre-defined order. After the last utterance a sound (a chime) is played to indicate the end of the information about the particular place. We found that it was difficult for the user to look at the same still image for an extended period of time, even if the user was interested in the place. We hence designed multiple images for each place that switched with new sentences. The content of the images reflects the content of the speech, e.g., when iTourist talks about the exhibition at the Royal Palace, it also shows the photos from the exhibition.

In addition to the information about each place, iTourist also contains distance and transit information between all places. It describes distance by walking time (minutes). If the walking distance is greater than 30 minutes, it gives bus transportation information instead. Simultaneously with a distance utterance, a line connecting the two places is shown on the map.

In total, 1618 utterances are included in iTourist, including 234 information utterances, 122 transit utterances, 1261 distance utterances, and one generic error message. In addition, a spoken introduction to the city is also included. This text is spoken by iTourist when it starts up to give an overview of the city. All the utterances were pre-synthesized by the Festival Speech Synthesis system and made into wave-files.

The users' eye-gaze was tracked in iTourist by a Tobii Eye Tracker 1750 running on a server. iTourist communicates with the eye tracker server via TCP/IP. Eye-gaze data is received from the eye tracker at a rate of 50 Hz. All functions in iTourist are implemented in C++.

iTourist allows the user to interact with it solely based on eye-gaze patterns, except when the user wants to commit to a place for booking or for a visit. This is done by clicking the left mouse button. iTourist marks the place and writes its name at the bottom of the screen as a reminder.

Based on the insights and patterns identified from the RealTourist study, we designed and implemented iTourist's architecture and algorithms (Figure 2). In iTourist, all objects (places) maintain two basic variables: the Interest Score (IScore) and the Focus Interest Score (FIScore). Based on these two variables of all objects, the Event and Interaction Manager (EIM) in iTourist determines what, if anything, to talk about with the user. The object that iTourist currently talks about is called the active object. .
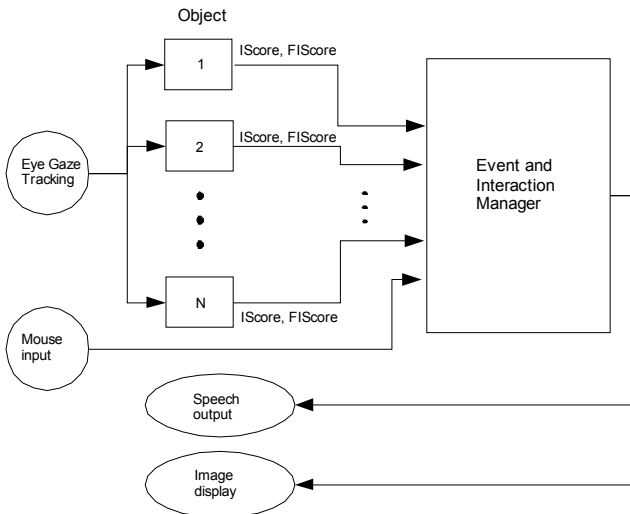

Figure 2. iTourist System structure

Modeling User Interest

At the core of iTourist is user interest detection. Although iTourist aims to be general enough for tourist information systems of its kind, it makes some assumptions about the nature of the information the user looks at and how a dialogue progresses in this type of situation. It assumes that the conversation is object-based and these objects are related to one another by category (e.g. hotels), location, or a task-dependent relationship, e.g. hotels and conference centre.

The two variables, IScore and FIScore, of each object represent the user's interest level in the object as an active object and inactive object, respectively.

IScore
IScore is used to determine an object's "arousal" level, or the likelihood that the user is interested in hearing about it. Its basic component is eye-gaze intensity (p), defined by accumulated gaze duration on the object (T_{ISon}) in a moving time window (T_{IS}). The moving time window ensures that measured p reflects the user's current interest.

$$p = \frac{T_{ISon}}{T_{IS}} \qquad (1)$$

where T_{ISon} is the accumulated time on the object within time window T_{IS}, and T_{IS} is the size of the moving time window.

Critical to our design is that the intensity variable p is modified by other factors related to user interest. Some of these factors inhibit the object's excitability and others increase it. These factors are collected into one variable α, which is used to adjust the gaze intensity residual ($1-p$):

$$p_{is} = p(1 + \alpha(1 - p)) \qquad (2)$$

or $$p_{is} = p + \alpha p - \alpha p^2 \qquad (3)$$

where p_{is} is the arousal level of the object (i.e. IScore) and α is the excitability modification. The value of p_{is} is between 1 and 0, and the value of α is between -1 and 1. A negative α inhibits the IScore, while a positive α excites the IScore. If no other factors than the accumulated time on the object exist, then α is equal to zero and the IScore would solely be dependent on the eye-gaze intensity.

Based on our observations from the RealTourist study summarized earlier, the iTourist system uses four factors in an object's excitability modification in the following manner (Figure 3):

$$\alpha = c_0 \frac{c_f \alpha_f + c_c \alpha_c + c_s \alpha_s + c_a \alpha_a}{c_f + c_c + c_s + c_a} \qquad (4)$$

where α_f is the frequency of the user's eye gaze entering and leaving the object; α_c is the categorical relationship with the previous active object; α_s is the relative size to a baseline object; and α_a records previous activation of the object. c_f, c_c, c_s and c_a are constants empirically adjusted.

The frequency of the user's eye gaze entering and leaving (α_f), identified as one indication of a user's interest in an object in the RealTourist study, was calculated over a time window, N_f. N_f can be viewed as the "memory span" of an object:
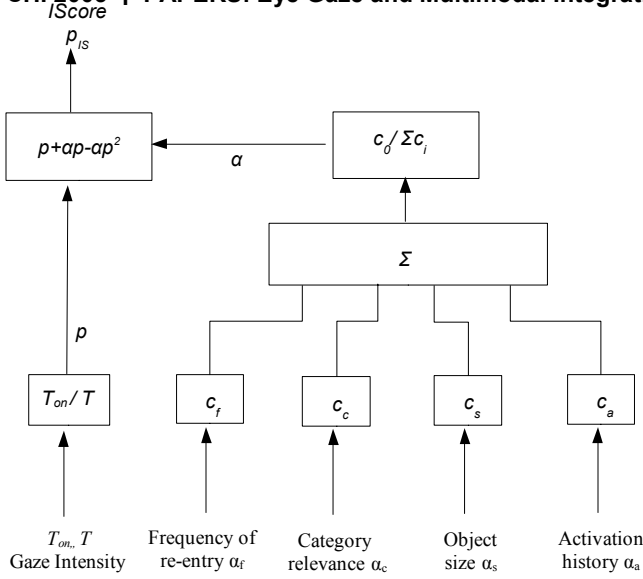
Figure 3. How the measured gaze intensity is filtered by other factors to finally form the IScore

Figure 4. iTourist map, shown here with superimposed object bounding box and eye-gaze fixation trace

$$\alpha_f = \frac{N_{sw}}{N_f} \qquad (5)$$

where N_{sw} is the number of times the eye gaze enters and leaves the object, and N_f is the maximum possible N_{sw} in the set time window.

The categorical relationship with the previous objects (α_c) is currently treated as binary; either the object is of the same category as the previous one, or not (although fuzzy membership is also possible in principle). The user is more likely to be interested in the same category before he/she finds a solution (a hotel, a restaurant) that matches his/her criteria in that category. When the user has committed to a place, he or she is not as likely to be interested in objects of the same category. For this reason, the value of a_c can be either +1 (when the category of the current object matches the previous object), 0 (no match), or -1 (when the category of the current object matches any of the committed objects). This modification factor is particularly task-domain dependent. It models the regularity in the trip-planning task.

The interest detection algorithm was targeted at a map-based application. On the map, most places (objects) have the same sizes (implicit bounding box, shown in Figure 4). However, there are a number of larger objects, such as attractions. Since the data from the eye tracker is not noise-free, the larger objects have a higher probability of being hit by random noise. The excitability of objects is therefore adjusted according to their size:

$$\alpha_s = \frac{S_b - S}{S} \qquad (6)$$

where S_b is the area size of the common objects which are also the smallest, and S is the size of the current object. α_s is 0 when S is the same as S_b. Otherwise it is negative, and therefore inhibits large objects' excitability.

Finally, if a place previously has been active, it is not very likely that the user intends to activate it again. This is the last component of α, defined by α_a. Previous activation of a place should inhibit its excitability. α_a is -1 when the place has been activated and 0 when it has not been activated.

Our approach in modeling IScore is prescriptive, encapsulating many of the behavioral patterns observed in our RealTourist study. It is easily extendable to other factors that can influence an object's excitability. For example, user preferences of the responsiveness of the system could be added as a fifth factor to α. This approach can be made more sophisticated in the future. For example, the relative weights of the factors constituting α could be updated with machine-learning algorithms to better adapt to the user's behavior.

As soon as the IScore of an object moves above a set threshold, the object is qualified to become the active focus of dialogue, pending the Event and Interaction Manager's control.

FIScore
FIScore measures how the user keeps up his or her interest in an active object. The FIScore is similar to IScore in that the base component of FIScore is the intensity of the user's eye-gaze on the object. However it only involves one other factor, which is the eye-gaze intensity on all related objects during the same time window. A weight constant c_r makes sure that the intensity of related objects contributes less than the intensity of the active object to the FIScore:

$$p_{FIS} = \frac{T_{on} + c_r T_{ron}}{T_F} \qquad (7)$$

where T_{on} is the accumulated time on the active object within time window T_F and T_{ron} is the accumulated time of all related objects within time window T_F. Objects were

considered related if their location was close to the active object, or if they were on the list of objects the user had committed to. In addition, all hotels had the Conference Centre as a related object. Objects with the same category as the active object were not considered related in calculating the FIScore, although they could be located close to one another. This allowed the user to easily switch between, for example, different hotel alternatives.

When an active object's FIScore dropped below a set threshold, the object could be deactivated by the Event and Interaction Manager.

Interest in distance relation

Eye-gaze patterns can be used not only to detect interest in an object, but also to detect interest in the relationship between two objects. We observed in the RealTourist study that when judging distance, the tourist switched back and forth between two places on the map. We use this eye-gaze pattern as the basis for iTourist to give information about the distance between two objects. However, looking back and forth between two objects is a relatively common eye-gaze pattern. It does not always indicate an interest in distance. One restriction imposed on iTourist in this regard is that distance information can only be triggered between the active object and another object. This assumes that the user is only interested in how other objects are located in relation to the currently active object.

The specific distance interest detection algorithm involves an object memory that sequentially stores objects being recently fixated on by the eye-gaze. The algorithm checks transitions between these objects in the memory store. If a pair of objects (one of them has to be the current active object) with two or more transitions between them is identified, iTourist will utter the distance between them to the user. If more than one pair exits with two or more transitions, the distance between the pair with the highest number of transitions is uttered.

Once the distance information is presented to the user, we found that the user tends to gaze back and forth between the two objects, which may result in a loop of distance information. For this reason, our system keeps track of the history and suppresses the immediate second occurrence.

System Tuning

The coefficients in the iTourist algorithms were tuned by testing with the authors and five pilot participants. They were first set at the most plausible values according to the relative importance of each factor and then modified during pilot testing.

The IScore window (T_{IS}) was set to 1000 ms. The activation threshold for IScore was set to 0.45. This means that if the IScore was only influenced by the time the user looked at an object, the user needed to look at it for 450 ms to trigger it.

The FIScore window (T_F) was set to 2000 ms. This makes the FIScore slightly more conservative than the IScore. The

FIScore threshold was set to 0.22. The weight (c_r) on related objects' accumulated time in T_F was set to 0.3. When the user only looked at related objects and not the active object, he or she needed to look at related objects longer than 1466 ms (within the 2000 ms window) to keep it active.

Dialogue Model

The algorithms for calculating IScore, FIScore, and interest in distance between two objects are based on a certain implicit dialogue model the users follow in the tourist information domain. In addition iTourist also models the order of information presented.

Each place has a number of sentences iTourist can speak. When an object becomes deactivated and reactivated again within one minute, the Event and Interaction Manager will continue from where it stopped. If the object is deactivated for more than one minute and re-activated, the Event and Interaction Manager will start from the first sentence.

When an object becomes active, the first utterance depends on the last active object. When the last active object is related to a newly activated object, iTourist plays a transit utterance. For example, if the user listens to information about the Royal Palace and then changes interest to the Museum of Modern Art, iTourist starts out saying: "The older art collection at the Royal Palace can be complemented with more modern art at the Museum of Modern Art" before giving the regular information about the museum.

When ITourist finishes talking about an object that is still active (FIScore above deactivation threshold), it will repeat the same information from the start.

Event and Interaction Manager

The Event and Interaction Manager (EIM) manages the state of the whole system based on the states of all individual objects. Figure 5 illustrates the basic logic flow of the EIM.

The states of the individual objects are evaluated each time new eye-gaze data arrives from the eye tracker. EIM starts out to check if there is a current active object. If there is no active object, EIM finds the object whose IScore has passed the activation threshold. It activates that object and starts to play its information (visual images and speech).

If there is an active object, EIM first checks if the object is currently playing information (i.e. in the middle of a sentence). As long as the object plays information, it stays active. When the object is active and is not playing information, EMI determines if the object should play new information or if it should be deactivated by checking if the object's FIScore has dropped below the deactivation threshold. This procedure enables iTourist to make graceful endings with the spoken output.
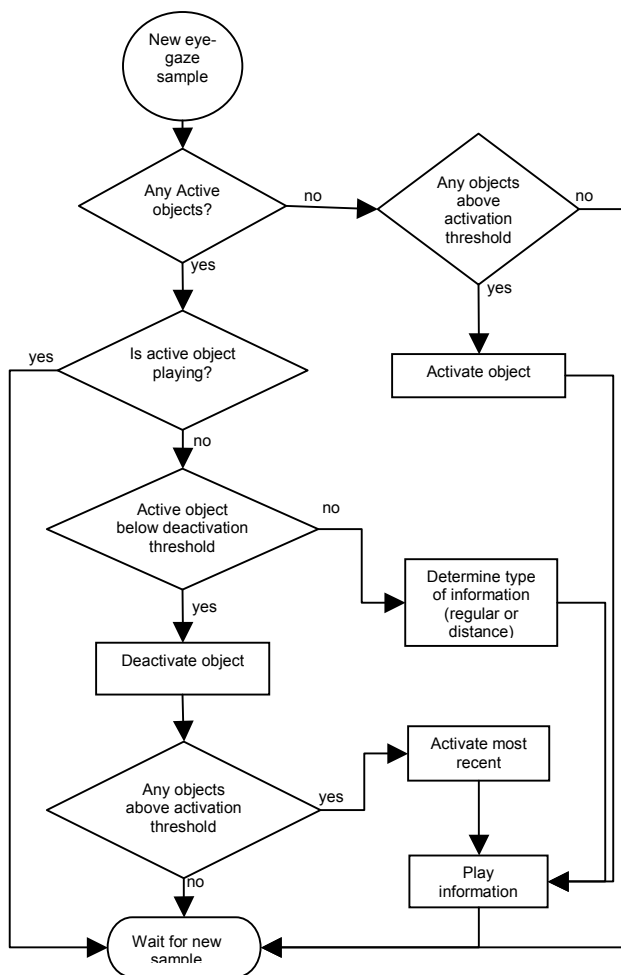
Figure 5. Event and Interaction Manager logic flow

When the active object's FIScore drops below the deactivation threshold, EIM deactivates it. EIM then checks if there is any object whose IScore has gone above the activation threshold. If there is more than one object, EIM chooses the object to which the user most recently looked.

Mouse events can also trigger transitions from the inactive state to the active state, or change the currently active object. In this case the IScore or the FIScore does not need to be reached the thresholds for activation/deactivation. EIM processes the mouse events as soon as they are received. This means that iTourist can be interrupted in the middle of the sentence with a mouse click.

RELATED WORK

Although not all in the context of managing human-computer dialogue, there have been several previous attempts at detecting users' interest based on patterns in eye gaze. Most of them have been focused on text reading or target selection with eye gaze.

Detecting gaze patterns in text is often used to identify words that a user has difficulty with. One example of this is the Reading Assistant by Sibert et al. [19]. It pronounces words

the user looks at for longer than a predefined threshold. Another popular topic is dictionaries, such as iDict by Hyrskykari *et al* [9, 10]. iDict decides which word's translation is displayed based on the eye-gaze dwell time. A different application using word detection from gaze patterns is SUITOR by Maglio and colleagues [4, 15]. SUITOR identifies content words and displays information, such as news, relevant to these words.

Gaze patterns have also been used to make selection with eye gaze more accurate. Methods for selection have been developed to fit a context; for example, Salvucci [18] developed a method for eye typing based on gaze patterns and language models. Gaze patterns have also been used for identification of an object [14], and for performing actions such as zooming [7]. Edwards [6] presented an algorithm that attempts to distinguish if the user is merely looking around, or wanting to select a target. The goal of this algorithm was to avoid the Midas touch problem in eye-gaze selection.

The methods for detecting gaze patterns used in previous work range from relatively simple accumulated time on target [19] to more sophisticated statistical models [18]. All of these methods include additional processing of the gaze data in addition to locating where the user is fixating. For example, all the methods detecting gaze patterns of text use algorithms to detect which line the user is currently reading. The additional processing of the eye-gaze data is specific to the purpose of the system and the type of visual information the users look at. This also means that only parts of the methods can be transferred to other domains.

The work that is most similar to the present study is by Starker and Bolt [20]. They used users' eye-gaze to modify a story told by a narrator. Starker and Bolt proposed three methods for measuring users' interest in particular objects. Model 1 essentially used the accumulated time on an object. In the two other models, the interest level raised when the user looked at an object and decayed when the user looked away from the object. The two latter models emphasized the user's most current interest while the first model emphasized the whole history of interest in an object. Starker and Bolt did not use a variable (as our FIScore) to determine if the user had lost interest in the current object. Instead they used "a low-end cut-off point of the sample standard deviation of item interest levels" (p. 8). The reason to use this procedure was that the next topic was decided based on which object had the highest interest level. Another weakness with the three different methods for calculating the interest levels was that they were never formally tested with interactive users. Starker and Bolt reported that the three algorithms were tested with pre-recorded eye-gaze patterns and simulated eye gaze by an "eye cursor." They noted that "differences in the narrative behavior showed up between the three models" (p. 8), and that model 3 talked longer than model 1. Several factors limit Starker and Bolt's model to more complex systems such as iTourist, which can also keep up the activation of an object based on the user's eye-gaze on related objects. Another difference is that while Starker and

Bolt's method might be suitable for a narrative purpose, it was not designed to handle a more interactive human-computer system like ours.

USER STUDY OF ITOURIST

Although iTourist is still at a preliminary stage of technical development based on relatively simple algorithms, it is important to test if our goal of making eye gaze a channel of human-computer dialogue and if the iTourist approach based on observations from RealTourist study are *feasible* at all. We hence conducted a somewhat informal but realistic user study of iTourist.

Design

Twelve people, two women and ten men from twenty-seven to forty-one (median of 30.5) years of age, participated in a user study of iTourist. They had diverse cultural backgrounds and eight of them spoke English as a second language.

Their task was to plan a conference trip to the city Malexander. The task was made rather realistic. Various constraints such as hotel and meal limits were imposed on the trip. The participants were asked to find a hotel for lodging, a restaurant for a group dinner, one night club and enough attractions for a spare day on the weekend. They were encouraged to consider their own convenience and preferences in addition to price restrictions. They were also encouraged to explore the city, get a good idea of it, and be prepared to be quizzed later.

A total number of fifteen persons volunteered for the study. Three of them were excluded for failing to comply with the instructions, having a high rate of lost eye-gaze data, or having difficulty in understanding the synthesized speech. Given that the focus of our study was not on the quality of eye trackers or speech synthesis, the exclusion of these three participants should not impact our conclusions on the feasibility of iTourist and the soundness of the principles it embodies.

Procedure

The experimental session started with an introduction to the study and eye-tracker calibration. The participant was encouraged to try to find the best position for the eye tracker (Figure 1). After that, the experiment proceeded to planning the trip to Malexander. After completion, the participants were asked a few questions about the city, and asked to fill in a post-test questionnaire. The session ended with a short interview in which the participants could comment on their experience of using iTourist.

Performance of the eye tracker

The performance of iTourist depended on the quality of eye tracking. On average, the twelve participants' eye-gaze was missed by the eye-tracker 7.6% of the time (SD=3.14). The time period of missing eye-gaze was on average 86 ms long (SD=188.9). The longer periods of loss happened when the users read instructions on paper. In a total of six cases the loss of eye tracking caused iTourist to deactivate the current focus object.

Results of eye-gaze pattern detection

IScore

As the core concept of eye-gaze interest pattern detection in iTourist, IScore worked well in the study. One measurement of its success was how often the users used the mouse to activate a place. Of the twelve users, only four ever moved the mouse to activate any object. One of these four users, User 4, used the mouse seven times. The other three used the mouse only once. On average the mouse was involved in only 3% of all object activation (SD=5.7).

The time the users spent looking at a place before it was activated depended on which state iTourist was in. When the system state was inactive, iTourist rapidly detected the user's interest in a new place. In this case the average activation time from the first moment the user looked at a new object to the moment iTourist started talking about it was 421 ms. Since IScore's threshold was 0.45 and its time window was set at 1000 ms, factors other than gaze intensity such as category relationship, frequent revisit and activation history must have contributed to the excitability of some of the objects as planned. When the iTourist system was in an active state (talking about something), it took on average 1507 ms to switch the activation to another object. This delay was due to the fact that iTourist did not stop in the middle of a sentence.

FIScore

One user thought that iTourist stopped talking somewhat too early, suggesting that the FIScore's (deactivating) threshold was too high for him. One measurement of the success of the FIScore algorithm was how many times a user reactivated a place immediately after it was deactivated. Analysis showed that there were 1.6 (SD=1.56) or 6.3% such cases per person or 19 in total of all participants. Four of these were caused by low accuracy and missed eye gaze data by the eye-tracker. Four others were preceded by a distance utterance. This kind of utterance often made the users look at the two places iTourist was talking about and hence caused more eye movements than regular utterances. The small number of prematurely deactivated places shows that the FIScore algorithm worked well in general.

The fact that iTourist always had to finish an already started sentence was noticed by the users. On average, iTourist detected the user's disinterest (or losing interest) in an active object 1.4 sec after the start of a sentence. Since the mean length of the sentences played by iTourist was 4.7 sec (SD=1.6), it means that often iTourist had to continue for another 3.3 sec (mean) to finish its sentence, despite "knowing" the user was no longer interested in the object it was talking about. Some participants commented on this, but also stated that it would be odd if iTourist cut itself off before finishing a sentence.

In the post-test questionnaire, the participants were asked two questions regarding interest and disinterest detection: how often iTourist talked about a place they were not interested in (false alarm), and how often iTourist did not talk about places they were interested in (misses). The rating distributions on these questions are shown in Figure 6. They indicate that the algorithms were not always successful, but in general were well received. Most of the false alarms were related to distance relation detection.
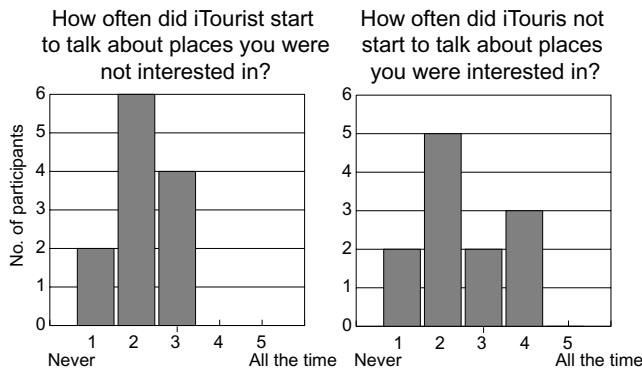

Figure 6. Users' rating on false alarms and misses

Distance relation detection

The distance interest detection in iTourist was probably most naïve. It was also the most difficult to handle. Users' reaction to it varied from person to person. Some were apparently not aware of it (even though the function was mentioned in the written introduction). One, User 13, made the following suggestion: "If you looked at a couple of different things and then it could draw some connection between them… Like if you looked at a hotel and you looked at a nightclub and then…you might be wondering whether it was easy to get from the hotel to the nightclub and then it could talk about how you can get to one from another." What he suggested was, in fact, how iTourist worked in distance relation detection. Sometimes the distance information clearly failed to match the user's need or natural expectation: "It apparently tries to make a connection for you when you go to something close by and it says, oh, this is five minutes away. No, I don't want to know that; I want to know what's there" (User 6). "I do like it [the distance information] although I didn't seem to be able to control it. It wouldn't do it on my will. It would only do it randomly" (User 12). "It is clear that it had something to do with when you looked at one thing and then the other, but I couldn't figure out that I had to look back and forth" (User 6).

Other users took advantage of the distance interest detection. "Actually, when I looked at a restaurant, the distance between the hotel and the restaurant appeared. That is very nice. It was the kind of information that I wanted to know" (User 11)

One of the users listened through 18 distance information utterances. On average, each user listened to 7.2 distance information utterances (SD=4.5), of which 1.2 utterances

were repeated (SD=1.4). The distance information contributed to, on average, 30.6% (SD=18.01) of all the sentences played by iTourist.

Overall Results: Acceptance and Impression of iTourist

The most remarkable evidence of the system's success was that all of the participants completed their tasks with iTourist. Only three users had minor deviations from the stated rule of checking price information of some of the restricted choices (two restaurants and one hotel), although all these three cases were in fact within the set price limit. Considering it was the same trip planning task as in our previous RealTourist study that involved a remote human "Tourist Consultant" on the line busy looking up information and talking about places in natural language, this is quite encouraging for the role that eye-gaze may play in human-machine communication.

The participants spent between 5:15 minutes and 13:50 minutes (average 8:27 minutes, SD=2:34) using iTourist to plan their trip. They looked at, on average, 16.6 (SD=3.6) places and listened, on average, to 82.4 utterances (SD=2.3). For each place they looked at, on average they listened to 3.8 utterances (SD=1.2). The post-test interview showed that they had a correct impression of and clear motivation to visit the places they had learned about. Many of the users enjoyed using iTourist; one of the users expressed a wish to visit Malexander. In the post-test questionnaire, the iTourist average was a 3.9 (SD=.79) rating on a 5-grade scale (5 the best rating and 1 the worst). Figure 7 shows the distribution of the twelve users on the question as to how well they thought iTourist worked. In the post-test interview the users gave overall quite positive comments about iTourist.

"I thought it was pretty convenient…the idea is really cute... like you look at something and it pops up this information bit" (*User 7*).

One of the design goals of iTourist was to utilize natural eye-gaze movement determined by the task as implicitly as possible. In the actual tests users were clearly aware of the role of their eye gaze, as indicated by the comment above. As to whether they have to consciously *control* the system with their gaze (see our previous discussions on avoiding eye-control [28, 29]), four of the twelve participants felt they had to "stare at" an object at least once. Usually this happened when iTourist was in the middle of a sentence, but the participant wanted to move on. Overall, most participants felt that it was quite natural to use iTourist without being too conscious of their eye gaze. This is evident in the following comment.

"Overall, I would say that it is a nice program, very cool, and you get to know quite a lot when you listen and you look everywhere and you get quite a good feeling of what you want to do quite fast --- I think this was a really cool program. I mean, it tracks down what you *think* and what you *want* to do." (*User 10)*

"Actually, I think it is pretty good. It does get my attention. It doesn't bore me or annoy me. I think I like it. I'm quite

impressed how it correctly interprets my attention to talk about places that I'm interested in." (*User 11*)

"I liked it. It worked pretty well." (*User 12*)

"This is the best experience that I ever had with eye tracking." *User 15*

Some of the users wanted to have more interaction capabilities in iTourist, although they realized that the system was pushing the limit with only eye gaze as the input method.

"I think for now when it doesn't have back channel [from the user to iTourist] it works all right; you cannot do any better."(*User 8*)

Apparently, there were cases where iTourist talked about places the participant did not consciously want to know about. Some of the users enjoyed a certain degree of "randomness" from iTourist and did not describe the experience as being "out of control".

"Actually, the sushi [restaurant] just came up. I didn't look at it consciously. It just happened to pop up. I actually like that. It's nice that it can give you information when you browse over the map." (*User 7*)

Preference for responsiveness differed from one individual to another. Some of the users felt iTourist, with its current parameter settings, was too sensitive, while others enjoyed a high degree of responsiveness. Adapting to individual users' preferences appears to be a critical factor for the success of future multimodal information systems that use eye gaze.

DISCUSSION AND CONCLUSION

Motivated by and grounded in observations of eye-gaze patterns in human-human dialogue over spatial content, we explored the use of eye-gaze patterns in managing human-computer interaction. In contrast to previous HCI works on eye-gaze that have primarily focused on direct manipulation user interfaces, this research targets the human-machine dialogue type of interaction. We envision future multimodal HCI systems with eye gaze as a critical contributing channel. Towards that goal, we developed a system, iTourist, that encapsulates knowledge gained in a previous study on a computer-mediated, human-human collaboration system (RealTourist) for city trip planning. iTourist provides information about a city to a user based solely or primarily on the user's eye-gaze patterns. The results show that eye-gaze can indeed play an important role in managing human-computer dialogue. In this study, by eye-gaze information alone the system could manage its visual and audio (speech) output and help users to plan an entire trip to a city, including finding hotels and restaurants within price limits and to the user's liking in terms of distance, location and preferences. It is remarkable that the "tourist consulting" task, quite demanding on the part of a real human operator in attending the tourist's eye-gaze, hearing what he or she asks about, looking up information, and telling the user relevant information [17], could be accomplished by iTourist successfully. The overall reactions of the users to iTourist
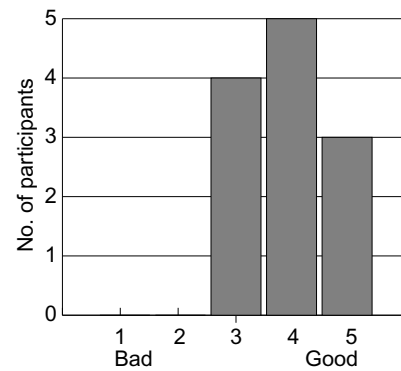

Figure 7. Distribution of the users' ratings on the question of how well iTourist worked (1-lowest, 5-highest)

were quite positive. On the other hand, although it is not as flexible as interactive storytelling previously researched with eye-gaze tracking [20], trip planning is a more tolerant task than many common HCI applications. A user is not likely to be very disturbed if iTourist talked about something he or she has no intention to learn about, which happens with human agents too.

Our investigation also showed that the prescriptive approach to developing interaction algorithms based on observations in human-human communication, although relatively naïve, could work at least in a specific domain. Note that iTourist could be made to work with any city, as long as relevant information is given to it.

Undoubtedly our approach in general and the iTourist implementation in particular are far from being mature or perfect. Eye-gaze pattern-based interaction systems, as any other recognition based systems, can produce both false alarms and misses. Some of these limitations can be over-come by developing more advanced techniques such as statistical learning, but more importantly ambiguity can be dramatically reduced when multiple modalities are combined due to the mutual disambiguation effects [16]. If eye-gaze pattern alone in our stress test of iTourist could be successful most of the time, its role can be expected to be even more powerful when combined with other modalities such as speech recognition.

ACKNOWLEDGEMENT

REFERENCES
1. Allopenna, P.D., Magnuson, J.S. and Tanenhaus, M.K. Tacking the time course of spoken word recognition using eye movements: Evidence for continuous mapping

models. *Journal of Memory and Language*, *38*, (1998). 419-439.

2. 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). 6-16.

3. Bolt, R.A., Eyes at the interface. *Proc. Human Factors in Computer Systems* (*1982*), ACM, 360-362.

4. Campbell, C.S. and Maglio, P.P., A robust algorithms for reading detection. *Proc. ACM Workshop on Perceptive User Interfaces* (*2001*), 1-7.

5. Cooper, R.M. The control of eye fixation by the meaning of spoken language - a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, (1974). 84-107.

6. Edwards, G., A tool for creating eye-aware applications that adapt to changes in user behaviour. *Proc. 3rd International ACM Conference on Assistive Technologies* (*1998*), 67–74.

7. Goldberg, J.H. and Schryver, J.C. Eye-gaze determination of user intent at the computer interface. In Findlay, J.M., Walker, R. and Kentridge, R.W. (ed). *Eye Movement Research -- Mechanisms, Processes and Applications*, Elsevier Science, New York (1995).

8. Griffin, Z.M. and Bock, K. What the eye says about speaking. *Psychological Science*, *11*, 4 (2000). 274-279.

9. Hyrskykari, A., Majaranta, P., Aaltonen, A. and Räihä, K.-J., Design issues of iDICT: A gaze-added translation aid. *Proc. ACM Symposium on Eye Tracking Research & Applications* (*2000*), 9-14.

10. Hyrskykari, A., Majaranta, P. and Räihä, K.-J., Proactive response to eye movements. *Proc. INTERACT - IFIP Conference on Human-Computer Interaction* (*2003*), 129-136.

11. Jacob, R.J.K., What you look at is what you get: Eye movement-based interaction techniques. *Proc. CHI* (*1990*), ACM, 11-18.

12. Kaur, M., Tremaine, M., Huang, N., Wilder, J., Gacovski, Z., Flippo, F. and Mantravadi, C.S., Where is "it"? event syncronisation in gaze-speech input systems. *Proc. Fifth International Conference on Multimodal Interfaces* (*2003*), 151-157.

13. Kendon Some function of gaze direction in social interaction. *Acta Psychologica*, *32*, (1967). 1–25.

14. Li, M. and Selker, T., Eye pattern analysis in intelligent virtual agents. *Proc. IVA 2001, Lecture Notes in Artificial Intelligence, 2190* (*2001*), Springer-Verlag.

15. Maglio, P.P., Barrett, R., Campbell, C.S. and Selker, T., SUITOR: An attentive information system. *Proc. International Conference on Intelligent User Interfaces* (*2000*), 169-176.

16. Oviatt, S., Mutual disambiguation of recognition errors in a multimodal architecture. *Proc. CHI* (*1999*), ACM, 576-583.

17. Qvarfordt, P. Eyes on multimodal interaction (Ph.D. Thesis) *Department of Computer and Information Science*, Linkoping University, Linkoping Studies in Science and Technology Dissertation No. 893 (2004).

18. Salvucci, D., Inferring intent in eye-based interfaces: Tracing eye movements with process models. *Proc. CHI* (*1999*), ACM, 254-261.

19. Sibert, J.L., Gokturk, M. and Lavine, R.A., The reading assistant: Eye gaze triggered auditory prompting for reading remediation. *Proc. ACM Symposium on User Interface Software and Technology* (*2000*), 101-107.

20. Starker, I. and Bolt, R.A., A gaze-responsive self-disclosing display. *Proc. CHI* (*1990*), ACM, 3-9.

21. Tanaka, K., A robust selection system using real-time multi-modal user-agent interactions. *Proc. 4th International Conference on Intelligent User Interfaces* (*1999*), 105-108.

22. Tanenhaus, M.K., Magnuson, J.S., Dahan, D. and Chambers, C. Eye movements and lexical access in spoken-language comprehension: Linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistics Research*, *29*, 6 (2000). 557-580.

23. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. and Sedivy, J.C. Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 5217 (1995). 1635-1634.

24. Velichkovsky, B.M. Communicating attention-gaze position transfer in cooperative problem solving. *Pragmatics and Cognition*, *3*, 2 (1995). 99-224.

25. Vertegaal, R., The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. *Proc. CHI* (*1999*), ACM, 294-301.

26. Vertegaal, R., Slagter, R., van der Veer, G.C. and Nijholt, A., Eye gaze patterns in conversations: there is more the conversational agents than meets the eyes. *Proc. CHI* (*2001*), ACM, 301-308.

27. Ware, C. and Mikaelian, H.H., An evaluation of an eye tracker as a device for computer input. *Proc. CHI+GI* (*1987*), ACM, 183-188.

28. Zhai, S. What's in the Eyes for Attentive Input *Communications of the ACM* (2003), 34-39.

29. Zhai, S., Morimoto, C. and Ihde, S., Manual and gaze input cascaded (MAGIC) pointing. *Proc* (*1999*), ACM, 246-253.

30. Zhang, Q., Imamiya, A., Go, K. and Gao, X., Overriding errors in speech and gaze multimodal architecture. *Proc. 9th International Conference on Intelligent User Interfaces* (*2004*), 346-348.