Exploring the effects of gaze and pauses in situated human-robot interaction

Gabriel Skantze, Anna Hjalmarsson, Catharine Oertel
KTH Speech, Music and Hearing
Stockholm, Sweden
gabriel@speech.kth.se, annah@speech.kth.se, catha@kth.se

Abstract

In this paper, we present a user study where a robot instructs a human on how to draw a route on a map, similar to a Map Task. This setup has allowed us to study user reactions to the robot's conversational behaviour in order to get a better understanding of how to generate utterances in incremental dialogue systems. We have analysed the participants' subjective rating, task completion, verbal responses, gaze behaviour, drawing activity, and cognitive load. The results show that users utilise the robot's gaze in order to disambiguate referring expressions and manage the flow of the interaction. Furthermore, we show that the user's behaviour is affected by how pauses are realised in the robot's speech.

1 Introduction

Dialogue systems have traditionally relied on several simplifying assumptions. When it comes to temporal resolution, the interaction has been assumed to take place with a strict turn-taking protocol, where each speaker takes discrete turns with noticeable gaps in between. While this assumption simplifies processing, it fails to model many aspects of human-human interaction such as turn-taking with very short gaps or brief overlaps and backchannels in the middle of utterances (Heldner & Edlund, 2010). Recently, researchers have turned to more incremental models, where the dialogue is processed in smaller units (Schlangen & Skantze, 2011). On the output side, this allows dialogue systems to start speaking before processing is complete, generating and synthesizing the response segment by segment, until the complete response is realised. If a segment is delayed, there will be a pause in the middle of the system's speech. While previous studies have clearly shown the potential benefits of incremental speech generation (Skantze & Hjalmarsson, 2012; Dethlefs et al., 2012; Buschmeier et al., 2012), there are few studies on how users react to pauses in the middle of the system's speech.

Apart from the real-time nature of spoken interaction, spoken dialog technology has for a long time also neglected the physical space in which the interaction takes place. In application scenarios which involve *situated interaction*, such as human-robot interaction, there might be several users talking to the system at the same time (Bohus & Horvitz, 2010), and there might be physical objects in the surroundings that the user and the system refer to during the interaction (Boucher et al., 2012). In such settings, gaze plays a very important role in the coordination of joint attention and turn-taking. However, it is not clear to what extent humans are able to utilize the gaze of a robot and respond to these cues.

Here, we present a user study where a robot instructs a human on how to draw a route on a map, similar to a Map Task. The nature of this setting allows us to study the two phenomena outlined above. First, we want to understand how a face-to-face setting facilitates coordination of actions between a robot and a user, and how well humans can utilize the robot's gaze to disambiguate referring expressions in situated interaction. The second purpose of this study is to investigate how the system can either inhibit or encourage different types of user reactions while pausing by using filled pauses, gaze and syntactic completeness.

2 Background

2.1 Gaze in situated interaction

Gaze is one of the most studied visual cues in face-to-face interaction, and it has been associated with a variety of functions, such as managing attention (Vertegaal et al., 2001), expressing intimacy and exercising social control (Kleinke,

1986), highlighting the information structure of the propositional content of speech (Cassell, 1999) as well as coordinating turn-taking (Duncan, 1972). One of the most influential publications on this subject (Kendon, 1967) shows that speakers gaze away when initiating a new turn. At the end of a turn, in contrast, speakers shift their gaze towards their interlocutors as to indicate that the conversational floor is about to become available. Furthermore, it has been shown that gaze plays an important role in collaborative tasks. In a map task study by Boyle et al. (1994), it was shown that speakers in a face-to-face setting interrupt each other less and use fewer turns, words, and backchannels per dialogue than speakers who can not see each other.

A lot of research has also been done on how gaze can be used to facilitate turn-taking with robots (Mutlu et al., 2006; Al Moubayed et al., 2013) and embodied conversational agents (Torres et al., 1997). Several studies have also explored situated human-robot interaction, where the interlocutors sit around a table with objects that can be referred to, thus constituting a shared space of attention (Yoshikawa et al., 2006; Johnson-Roberson et al., 2011). However, there are very few studies on how the robot's gaze at objects in the shared visual scene may improve task completion in an interactive setting. One exception is a controlled experiment presented by Boucher et al. (2012), where the iCub robot interacted with human subjects. While the study showed that humans could utilize the robot's gaze, the interaction was not that of a free continuous dialogue.

Similarly to the study presented here, Nakano et al. (2003) presented a system that describes a route to a user in a face-to-face setting. Based on studies of human-human interaction, they implemented a model of face-to-face grounding. However, they did not provide a detailed analysis of the users' behaviour when interacting with this system.

Even if we successfully manage to model human-like behaviour in a system, it is not certain to what extent humans react to these signals when interacting with a robot. In the current work, we investigate to what extent the robot's gaze can be used to: (1) help the user disambiguate referring expressions to objects in the shared visual scene, and (2) to either inhibit or encourage different types of user reactions while the system pauses or at turn endings.

2.2 Pauses in the system's speech

Speakers in dialogue produce speech piece by piece as the dialogue progresses. When starting to speak, dialogue participants typically do not have a complete plan of how to say something or even what to say. Yet, they manage to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions (Levelt, 1989). Still, pauses occur frequently within utterances and it has been shown that these play a significant role in human-human dialogue (for an overview, see Rochester, 1973). For example, the timing and duration of pauses have important structural functions (Goldman-Eisler, 1972), pauses (filled and silent) are associated with high cognitive load and planning difficulties (Brennan & Williams, 1995), and whether a pause is detected or not does not only depend on duration but also on its linguistic context (Boomer & Dittmann, 1962).

Recently, several studies have looked into the possibilities of replicating the incremental behaviour of humans in human-machine interaction. Work on incremental speech generation has focused on the underlying system architecture (Schlangen & Skantze, 2011), how to incrementally react to events that occur while realizing an utterance (Dohsaka & Shimazu, 1997, Buschmeier et al., 2012), and how to make the incremental processes more efficient in order to reduce the system's response time (e.g. Dethlefs et al., 2012). In a recent study, we implemented a model of incremental speech generation in a dialogue system (Skantze & Hjalmarsson, 2012). By allowing the system to generate and synthesize the response segment by segment, the system could start to speak before the processing of the input was complete. However, if a system segment was delayed for some reason, the system generated a response based on the information obtained so far or by generating a pause (filled or unfilled). The system also employed self-repairs when the system needed to revise an already realised speech segment. Despite these disfluencies (filled pauses and self-repairs), an evaluation of the system showed that in comparison to a non-incremental version, the incremental version had a shorter response time and was perceived as more efficient by the users.

However, pauses do not only have to be a side-effect of processing delays. Pauses could also be used wisely to chunk longer instructions into shorter segments, giving the user enough

time to process the information. In this case, the system should instead invite user reactions during the course of its utterance. In the current work, we investigate to what extent the system can use filled pauses, syntactic completeness and gaze as cues to either inhibit or encourage the user to react when the system pauses.

3 Human-robot Map Task data

Map Task is a well establish experimental paradigm for collecting data on human-human dialogue [30]. Typically, an *instruction-giver* has a map with landmarks and a route, and is given the task of describing this route to an *instruction-follower*, who has a similar map but without the route drawn on it. In a previous study, (Skantze, 2012) we used this paradigm for collecting data on how humans elicit feedback in human-computer dialogue. In that study, the human was the instruction-giver. In the current study, we use the same paradigm for a human-robot dialogue, but here the robot is the instruction-giver and the human is the instruction-follower. This has resulted in a rich multi-modal corpus of various types of user reactions to the robot's instructions, which vary across conditions.

Figure 1: The experimental setup.

3.1 A Map Task dialogue system

The experimental setup is shown in Figure 1. The user is seated opposite to the robot head Furhat (Al Moubayed et al., 2013), developed at KTH. Furhat uses a facial animation model that is back-projected on a static mask. The head is mounted on a neck (with 3 degrees of freedom), which allows the robot to direct its gaze using both eye and head movements. The dialogue system was implemented using the IrisTK framework developed at KTH (Skantze & Al Moubayed, 2012), which provides a set of modules for input and output, including control of Furhat (facial gestures, eye and head movements), as well as a statechart-based authoring language for controlling the flow of the interaction. For speech synthesis, we used the CereVoice unit selection synthesizer developed by CereProc (www.cereproc.com).

Between the user and the robot lies a large map printed on paper. In addition, the user has a digital version of the map presented on a screen and is given the task to draw the route that the robot describes with a digital pen. However, the landmarks on the user's screen are blurred and therefore the user also needs to look at the large map in order to identify the landmarks. This map thereby constitutes a target for joint attention. While the robot is describing the route, its gaze is directed at the landmarks under discussion (on the large map), which should help the user to disambiguate between landmarks. In a previous study, we have shown that human subjects can identify the target of Furhat's gaze with an accuracy that is very close to that of observing a human (Al Moubayed et al., 2013). At certain places in the route descriptions, the robot also looks up at the user. A typical interaction between the robot and a user is shown in Table 1. As the example illustrates, each instruction is divided into two parts with a pause in between, which results in four phases per instruction: *Part I*, *Pause*, *Part II* and *Release*. Whereas user responses are not mandatory in the *Pause* phase (the system will continue anyway after a short silence threshold, as in U.2), the *Release* requires a verbal response, after which the system will continue. We have explored three different realisations of pauses, which were systematically varied in the experiment:

COMPLETE: Pauses preceded by a syntactically complete phrase (R.5).

INCOMPLETE: Pauses preceded by a syntactically incomplete phrase (R.9).

FILLED: Pauses preceded by a filled pause (R.1). The phrase before the filled pause was sometimes incomplete and sometimes complete.

To make the conditions comparable, the amount of information given before the pauses was balanced between conditions. Thus, the incomplete phrases still contained an important piece of information and the pause was inserted in the beginning of the following phrase (as in R.9).

Table 1: An example interaction.

Turn	Activity	Phase
R.1	[gazing at map] continue towards the lights, ehm...	Part I
U.2	*[drawing]*	Pause
R.3	until you stand south of the stop lights [gazing at user]	Part II
U.4	*[drawing] alright [gazing at robot]*	Release
R.5	[gaze at map] continue and pass east of the lights...	Part I
U.6	*okay [drawing]*	Pause
R.7	...on your way towards the tower [gaze at user]	Part II
U.8	*Could you take that again?*	Release
R.9	[gaze at map] Continue to the large tower, you pass...	Part I
U.10	*[drawing]*	Pause
R.11	...east of the stop lights [gaze at user]	Part II
U.12	*[drawing] okay, I am at the tower*	Release

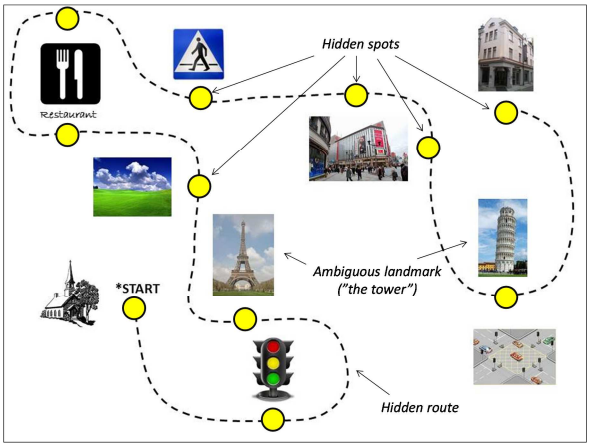

Figure 2: An example map.

Given the current limitations of conversational speech recognition, and lack of data relevant for this task, we needed to employ some trick to be able to build a system that could engage in this task in a convincing way in order to evoke natural reactions from the user. One possibility would be to use a Wizard-of-Oz setup, but that was deemed to be infeasible for the time-critical behaviour that is under investigation here. Instead, we employed a trick similar to the one used in (Skantze, 2012). Although the users are told that the robot cannot see their drawing behaviour, the drawing on the digital map, together with a voice activity detector that detects the user's verbal responses, is actually used by the system to select the next action. An example of a map can be seen in Figure 2. On the intended route (which obviously is not shown on the user's screen), a number of hidden "spots" were defined – positions relative to some landmark (e.g. "east of the field"). Each instruction from the system was intended to guide the user to the next hidden spot. Each map also contained an ambiguous landmark reference (as "the tower" in the example).

Pilot studies showed that there were three basic kinds of verbal reactions from the user: (1) an acknowledgement of some sort, encouraging the system to continue, (2) a request for repetition, or (3) a statement that some misunderstanding had occurred. By combining the length of the utterance with the information about the progression of the drawing, these could be distinguished in a fairly robust manner. How this was done is shown in Table 2. Notice that this scheme allows for both short and long acknowledgements (U.4, U.6 and U.12 in the example above), as well as clarification requests (U.8). It also allows us to explore misunderstandings, i.e. cases where the user thinks that she is at the right location and makes a short acknowledgement, while she is in fact moving in the wrong direction. Such problems are usually detected and repaired in the following turns, when the system continues with the instruction from the intended spot and the user objects with a longer response. This triggers the system to either RESTART the instruction from a previous spot where the user is known to have been *("I think that we lost each other, could we start again from where you were at the bus stop?"*), or to explicitly CHECK whether the user is at the intended location *("Are you at the bus stop?"*), which helps the user to correct the path.

Table 2: The system's action selection based on the user's voice activity and drawing.

User response	Drawing	Action
Short/Long	Continues to the next spot	CONTINUE
Short/Long	Still at the same spot	REPHRASE
Short (<1s.)	At the wrong spot	CONTINUE (with misunderstanding)
Long (>1s.)	At the wrong spot	RESTART or CHECK
No resp.	Any	CHECK

3.2 Experimental conditions

In addition to the utterance-level conditions (concerning completeness) described above, three dialogue-level conditions were implemented:

CONSISTENT gaze (**FACE**): The robot gazes at the landmark that is currently being described during the phases Part I, Pause and Part II. In

accordance with the findings in for example Kendon (1967), the robot looks up at the end of phase Part II, seeking mutual gaze with the user during the Release phase.

RANDOM gaze (**FACE**): A random gaze behaviour, where the robot randomly shifts between looking at the map (at no particular landmark) and looking at the user, with an interval of 5-10 seconds.

NOFACE: The robot head was hidden behind a paper board so that the user could not see it, only hear the voice.

3.3 Data collection and analysis

We collected a corpus of 24 subjects interacting with the system, 20 males and 4 females between the ages of 21-47. Although none of them were native speakers, all of them had a high proficiency in English. First, each subject completed a training dialogue and then six dialogues that were used for the analysis. For each dialogue, different maps were used. The subjects were divided into three groups with 8 subjects in each:

Group A: Three maps with the CONSISTENT (FACE) version and three maps with the NOFACE version. All pauses were 1.5 s. long.

Group B: Three maps with the RANDOM (FACE) version and three maps with the NOFACE version. All pauses were 1.5 s. long.

Group C: Three maps with the CONSISTENT version and three maps with the NOFACE version. All pauses were 2-4 s. long (varied randomly with a uniform distribution).

For all groups, the order between the FACE and the NOFACE condition was varied and balanced. Group A and Group B allow us to explore differences between the CONSISTENT and RANDOM versions. This is important, since it is not evident to what extent the mere presence of a face affects the interaction and to what extent differences are due to a consistent gazing behaviour. Group C was added to the data collection since we wanted to be able to study users' behaviour during pauses in more detail. Thus, Group C will only be used to study within-group effects of different pause types and will not be compared against the other groups.

After the subjects had interacted with the system, they filled out a questionnaire. First, they were requested to rate with which version (FACE or NOFACE) it was easier to complete the task. Second, the participants were requested to rate whether the robot's gaze was helpful or confusing when it came to task completion, landmark identification and the timing of feedback. All ratings were done on a continuous horizontal line with either FACE or "the gaze was helpful" on the left end and NOFACE or "the gaze was confusing" on the right end. The centre of the line was labelled with "no difference".

During the experiments, the users' speech and face were recorded and all events in the system and the drawing activity were automatically logged. Afterwards, the users' voice activity that had been automatically detected online was manually corrected and transcribed. Using the video recordings, the users' gaze was also manually annotated, depending on whether the user was looking at the map, the screen or at the robot.

In this study, we also wanted to explore the possibility of measuring cognitive load in human-robot interaction using EDA (electrodermal activity). Hence, in an explorative manner, we investigated how the realisation of the system's pauses and the presence of the face affected the cognitive costs of processing the system's instructions. For measuring this, we used a wearable EDA device, which exerts a direct current on the skin of the subject in order to measure skin conductance responses. For these measurements as well as the logging of the data the Q-Sensor developed by Affectiva[1] was used. The measurements were taken from the fingertips of the subjects. The sampling rate was 8 Hz. All post processing was carried out in Ledalab[2]. We first applied the Butterworth filter and then carried out a Continuous Decomposition Analysis. All skin conductance responses (SCR) with a minimum amplitude of 0.01 muS and a minimal distance of 700ms were used for further analysis. Due to problems with the EDA device, we only have data for six subjects in Group A, six in Group B and none in Group C.

4 Results

Analyses of the different measures used here revealed that they were not normally distributed. We have therefore consistently used non-parametric tests. All tests of significance are done using two-tailed tests at the .05 level.

[1] http://www.affectiva.com/
[2] http://www.ledalab.de/

4.1 Subjective ratings

The questionnaire was used to analyse differences in subjective ratings between Group A and B. The marks on the horizontal continuous lines in the questionnaire were measured with a ruler based on their distance from the midpoint (labelled with "no difference") and normalized to a scale between 0 and 1. A Wilcoxon Signed Ranks Test was carried out, using these rankings as differences. The results show that the Consistent version differed significantly from the midpoint ("no difference") in four dimensions whereas there were no significant differences from the midpoint for RANDOM version. More specifically, Group A (CONSISTENT) (n=8) found it easier to complete the task in the face condition than in the no face condition (Mdn=0.88, Z=-2.54, p=.012). The same group thought that the robot's gaze was helpful rather than confusing when it came to task completion (Mdn=0.84, Z=-2.38, p=.017), landmark identification (Mdn=0.83, Z=-2.52, p=.012) and to decide when to give feedback (Mdn=0.66, Z=-1.99, p=.046). The results of the questionnaire are presented in Figure 3.

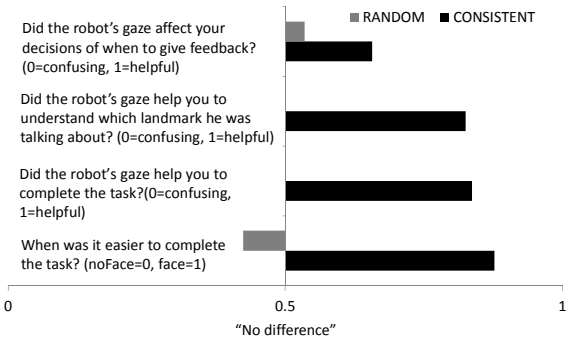

Figure 3: The results from the questionnaire. The bars show the median rating for Group A (consistent) and Group B (random).

4.2 Task completion

Apart from the subjective ratings, we also wanted to see whether the face-to-face setting affected task completion. In order to explore this, we analysed the time and number of utterances it took for the users to complete the maps. On average, the dialogues in Group A (CONSISTENT) were 2.5 system utterances shorter and 8.9 seconds faster in the FACE condition than in the NOFACE condition. For Group B (RANDOM), the dialogues were instead 2.3 system utterances and 17.3 seconds longer in the FACE condition (Mann-Whitney U-test, p<.05). Thus, it seems like the face facilitates the solving of the task, and that this is not just due to the mere presence of a face, but that the intelligent gaze behaviour actually contributes. In fact, the RANDOM gaze worsens the performance, possibly because subjects spent time on trying to make sense of signals that did not provide any useful information.

Looking at more local phenomena, it seems like there was also a noticeable difference when it comes to miscommunication. The dialogues in the RANDOM/FACE condition had a total of 18 system utterances of the type RESTART (vs. 7 in CONSISTENT), and a total of 33 CHECK utterances (vs. 15 in CONSISTENT). A chi-square test shows that the differences are statistically significant ($\chi^2(1, N=25) = 4.8$, p =.028; $\chi^2(1, N=48) = 6.75$, p=.009). This indicates that the users that did not get the CONSISTENT gaze to a larger extent did not manage to follow the system's instructions, most likely because they did not get guidance from the robot's gaze in disambiguating referring expressions.

4.3 Gaze behaviour

In order to analyse the users' direction of attention during the dialogues, the manual annotation of the participants' gaze was analysed. First, we explored how the completion type of the robot's utterance affected the users' gaze. In this analysis, FILLED and INCOMPLETE have been merged (since there was no difference in the users' gaze between these conditions). The percentage of gaze at the robot over the four different utterance phases for complete and incomplete utterances is plotted in Figure A in the Appendix. Note that the different phases actually are of different lengths depending on the actual content of the utterance and the length of the pause. However, these lengths have been normalized in order to make it possible to analyse the average user behaviour. For each phase, a Mann-Whitney U-test was conducted. The results show that the percentage of gaze at Furhat during the mid-utterance pause is higher when the first part of the utterance is incomplete than when it is complete (U=7573.0, p<.001). There were, however, no significant differences in gaze direction between complete and incomplete utterance during the other three phases (p>.05). This indicates that users gaze at the robot to elicit a continuation of the instruction when it is incomplete.

Second, we wanted to explore if gaze direction can be used as a cue of whether the user will provide a verbal response in the pause or not. The percentage of gaze at the robot over the four utterance phases for system utterances with and

without user response in the pause is plotted in Figure B in the Appendix. For each phase, a Mann-Whitney U-test was conducted. The results show that the percentage of gaze at Furhat during the mid-utterance pause (U=1945.5, p=.008) and Part II (U=2090.0, p=.008) of the utterance is lower when the user gives a verbal response compared to when there is no response. There were however no significant differences in gaze direction between complete and incomplete utterance during the other two phases (p>.05).

4.4 Verbal feedback behaviour

Apart from the user's gaze behaviour, we also wanted to see whether syntactic completeness before pauses had an effect on whether the users gave verbal responses in the pause. Figure 4 shows the extent to which users gave feedback within pauses, depending on pause type and FACE/NOFACE condition. As can be seen, COMPLETE triggers more feedback, FILLED less feedback and INCOMPLETE even less. Interestingly, this difference is more distinct in the FACE condition ($\chi^2(2, N=157) = 10.32$, p<.01). In fact, the difference is not significant in the NOFACE condition (p >.05).

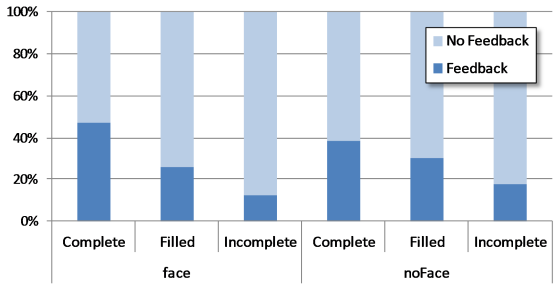

Figure 4: Presence of feedback depending on pause type (Group C).

In Skantze et al. (2013), we have also done a more thorough analysis of the verbal acknowledgements from the users. The analysis shows that the prosody and lexical choice in these acknowledgements ("okay", "yes", "yeah", "mm", "mhm", "ah", "alright" and "oh") to some extent signal whether the drawing activity is about to be initiated or has been completed. The analysis also shows how these parameters are correlated to the perception of uncertainty.

4.5 Drawing behaviour

Whereas gaze and verbal responses can be regarded as communicative signals, the users were told that the robot could not observe their drawing activity. However, the drawing of the route can be regarded as the purpose of the interaction and it is therefore important to understand how this is affected by the system's behaviour under different conditions. First, we wanted to see how the completeness of the robot's utterance in combination with the presence of the face affected the drawing activity. In this analysis, FILLED and INCOMPLETE have been merged (since there was no clear difference). The mean drawing activity over the four phases of the descriptions is plotted in Figure C in the Appendix. For each phase, a Kruskal-Wallis test was conducted showing that there is a significant difference between the conditions in the Pause phase (H(3) = 28.8, p<.001). Post-hoc tests showed that FACE/INCOMPLETE has a lower drawing activity than the other conditions, and that NOFACE/INCOMPLETE has a lower drawing activity than the COMPLETE condition. Thus, INCOMPLETE phrases before pauses seem to have an inhibiting effect on the user's drawing activity in general, but this effect appears to be much larger in the FACE condition.

Second, we aimed to investigate to what extent the robot's gaze at landmarks during ambiguous references helps users to discriminate between landmarks. The mean drawing activity over the four phases of the descriptions of ambiguous landmarks is plotted in Figure D in the Appendix. For each phase, a Kruskal-Wallis test was conducted showing that there is a significant difference between the conditions in the Part II phase (H(2)=10.2, p=.006). Post-hoc tests showed that CONSISTENT has a higher drawing activity than the RANDOM and NOFACE conditions. However, there is no such difference when looking at non-ambiguous descriptions. This shows that robot's gaze at the target landmark during ambiguous references makes it possible for the subjects to start to draw quicker.

4.6 Cognitive load

As mentioned above, we also wanted to study the cognitive costs of processing the system's instructions, as measured with a wearable EDA device. For each system utterance part (Part I and Part II), we calculated the sum of the amplitudes of the skin conductance responses (SoSCR) during the following three seconds. The SoSCR during the pause, depending on pause type are shown in Figure 5. A Kruskal-Wallis test revealed that there is an overall effect (H(2)=8.7, p=.13), and post-hoc tests showed that there is a significant difference between utterances which are incomplete and those with filled pauses, indi-

cating that the syntactic incompleteness without a filled pause leads to a higher cognitive load. We have no good explanation for this, and we do not know whether this is due to how the syntactically incomplete segments were realised by the synthesizer, or whether the same effect would appear in human-human interaction.

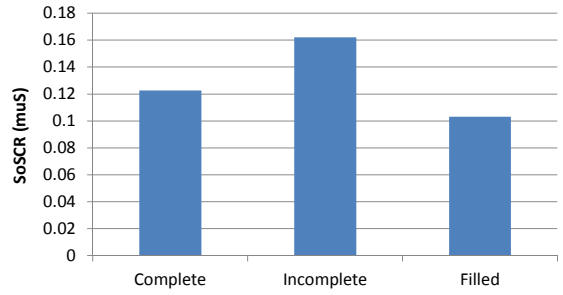

Figure 5: EDA at different pause types (Group A and B).

A similar analysis was done after both Part I and Part II to see if there is any difference in SoSCR between ambiguous and non-ambiguous references in the different conditions, as shown in in Figure 6. No such differences were found for Group B, but for Group A, ambiguous references were followed by a higher SoSCR in the NOFACE condition, indicating that the robot's gaze helps in disambiguating the referring expressions and reduces cognitive load (Mann-Whitney U-test; U = 6585, p = .001).

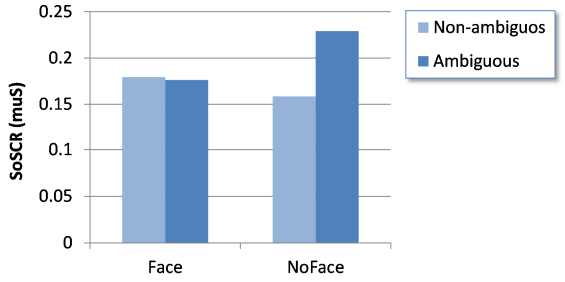

Figure 6: EDA for Group A (CONSISTENT).

5 Conclusions and Discussion

In this study, we have investigated to what extent the robot's gaze can be used to: (1) help the user disambiguate referring expressions to objects in the shared visual scene, and (2) to either inhibit or encourage different types of user reactions while the system pauses. The results show that the robot's gaze behaviour was rated as helpful rather than confusing for task completion, landmark identification and feedback timing. These effects were not present when the robot used a random gaze behaviour. The efficiency of the gaze was further supported by the time it took to complete the task and the number of misunderstandings. These results in combination with a faster drawing activity and lower cognitive load when system's reference was ambiguous, suggest that the users indeed utilized the system's gaze to discriminate between landmarks.

The second purpose of this study was to investigate to what extent filled pauses, syntactic completeness and gaze can be used as cues to either inhibit or encourage the user to react in pauses. First, the results show that pauses preceded by incomplete syntactic segments or filled pauses appear to inhibit user activity. Thus, our analyses of gaze and drawing activity show that users give less feedback, draw less and look at the robot to a larger extent when the preceding system utterance segment is incomplete than when it is complete. An interesting observation is that the inhibiting effect on drawing activity appears to be more pronounced in the face-to-face condition, which indicates that gaze also plays an important role here (since the robot looked down at the map during the pauses). Additionally, there is less cognitive load when the silence is preceded by a filled pause. These results suggest that incomplete system utterances prevent further user processing; instead the user waits for more input from the system before starting to carry out the system's instruction. After complete utterance segments, however, there is more drawing activity and the user looks less at the robot, suggesting that the user has already started to carry out the system's instruction.

The results presented in this study have implications for generating multimodal behaviours incrementally in dialogue systems for human-robot interaction. Such a system should be able to generate speech and gaze intelligently in order to inhibit or encourage the user to act, depending on the state of the system's processing. In future studies, we plan to extend our previous model of incremental speech generation (Skantze & Hjalmarsson, 2012) with such capabilities.

Acknowledgments

Gabriel Skantze is supported by the Swedish research council (VR) project *Incremental processing in multimodal conversational systems* (2011-6237). Anna Hjalmarsson is supported by the Swedish Research Council (VR) project *Classifying and deploying pauses for flow control in conversational systems* (2011-6152). Catharine Oertel is supported by *GetHomeSafe* (EU 7th Framework STREP 288667).

References

Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction. International Journal of Humanoid Robotics, 10(1).

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, *S., Isard,* S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

Bohus, D., & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *Proc ICMI'*10. Beijing, China.

Boomer, D. S., & Dittmann, A. T. (1962). Hesitation pauses and juncture pauses in speech. *Language and Speech, 5*, 215-222.

Boucher, J. D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., Dominey, P. F., & Ventre-Dominey, J. (2012). I reach faster when I see you look: gaze effects in human-human and human-robot face-to-face cooperation. *Frontiers in neurorobotics, 6*.

Boyle, E., Anderson, A., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and speech, 37*(1), 1-20.

Brennan, S., & Williams, M. (1995). The Feeling of Another's knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language, 34*, 383-398.

Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., & Schlangen, D. (2012). Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of SigDial* (pp. 295–303). Seoul, South Korea.

Cassell, J. (1999). Nudge, nudge, wink, wink: Elements of face-toface conversation for embodied conversational agents. In Cassell, J., Suillivan, J., Prevost, S., & Churchill, E. (Eds.), *Embodied Conversational Agents*. Cambridge, MA: MIT Press.

Dethlefs, N., Hastie, H., Rieser, V., & Lemon, O. (2012). Optimising Incremental Dialogue Decisions Using Information Density for Interactive Systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 82-93). Jeju, South Korea.

Dohsaka, K., & Shimazu, A. (1997). System architecture for spoken utterance production in collaborative dialogue. In *Working Notes of IJCAI 1997 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*, 103-113.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*, 555-568.

Johnson-Roberson, M., Bohg, J., Skantze, G., Gustafson, J., Carlson, R., Rasolzadeh, B., & Kragic, D. (2011). Enhanced Visual Scene Understanding through Human-Robot Dialog. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

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

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological Bulletin, 100*, 78-100.

Mutlu, B., Forlizzi, J., & Hodgins, J. (2006). A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Proceedings of 6th IEEE-RAS International Conference on Humanoid Robots* (pp. 518-523).

Nakano, Y., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)* (pp. 553-561).

Rochester, S. R. (1973). The significance of Pauses in Spontaneous Speech. *Journal of Psycholinguistic Research, 2*(1).

Schlangen, D., & Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse, 2*(1), 83-111.

Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA.

Skantze, G., & Hjalmarsson, A. (2012). Towards Incremental Speech Generation in Conversational Systems. *Computer Speech & Language, 27*(1), 243-262.

Skantze, G., Oertel, C., & Hjalmarsson, A. (2013). User feedback in human-robot interaction: Prosody, gaze and timing. In *Proceedings of Interspeech*.

Skantze, G. (2012). A Testbed for Examining the Timing of Feedback using a Map Task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Portland, OR.

Torres, O., Cassell, J., & prevost, S. (1997). Modeling gaze behavior as a function of discourse structure. *Proc. of the First International Workshop on Human-Computer Conversation*.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of ACM Conf. on Human Factors in Computing Systems*.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., & Miyamoto, T. (2006). Responsive robot gaze to interaction partner. In *Proceedings of robotics: Science and systems*.

Appendix

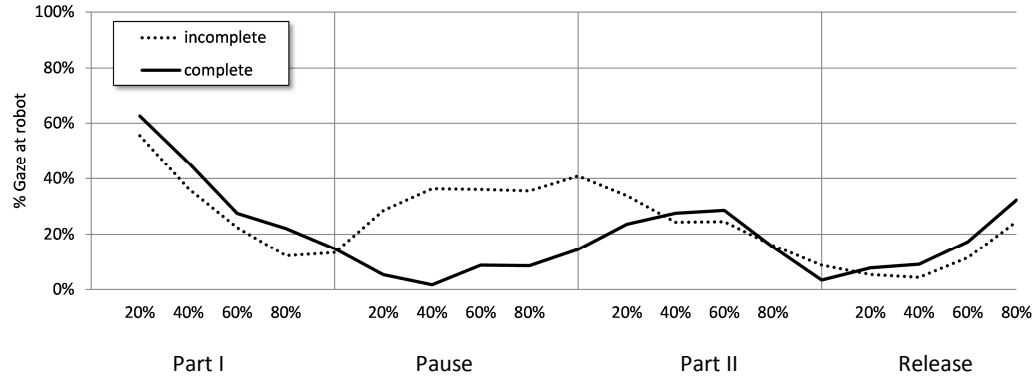

Figure A: Average user gaze depending on pause type (Group C).

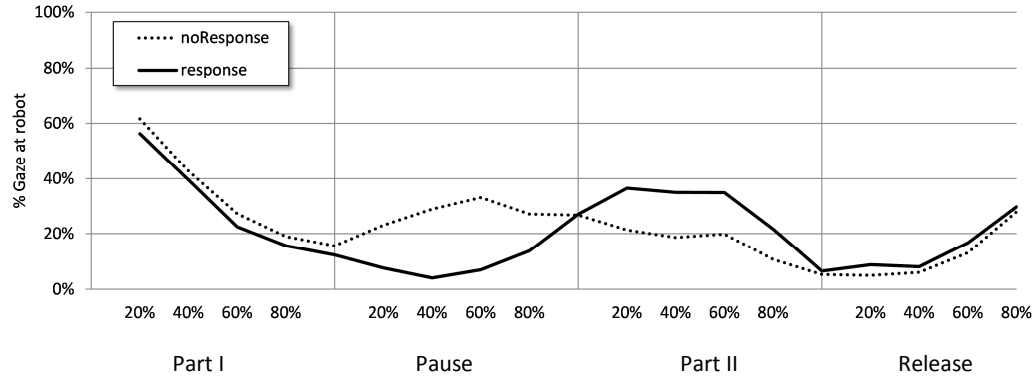

Figure B: Average user gaze depending whether the user responds in the pause (Group A and B).

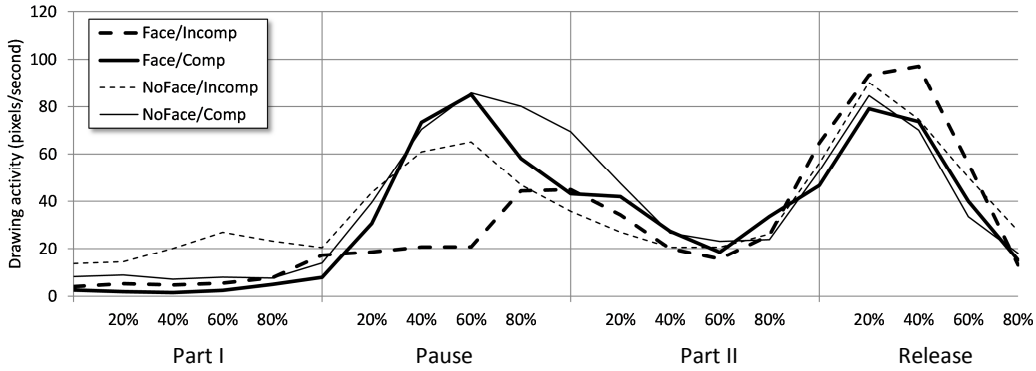

Figure C: Average drawing activity depending on pause type and the presence of the face (Group C).

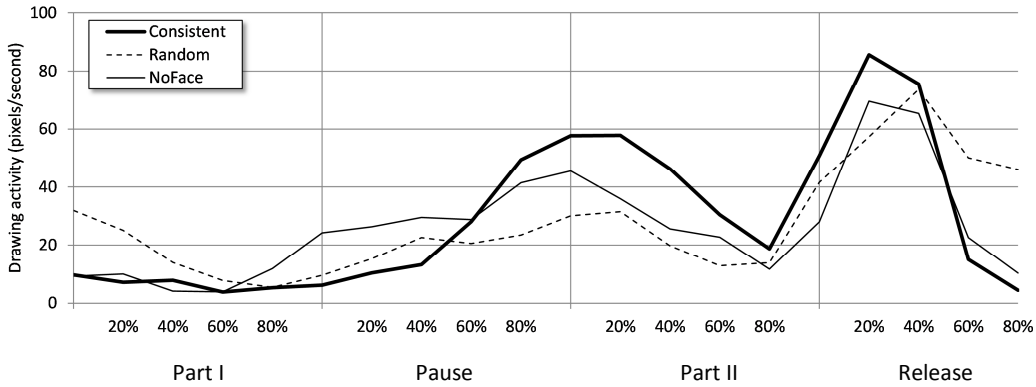

Figure D: Average drawing activity during ambiguous references depending on condition (Group A and B).