Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 65 (2014) 50–66

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Turn-taking, feedback and joint attention in situated human–robot interaction

Gabriel Skantze *, Anna Hjalmarsson, Catharine Oertel

Department of Speech Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

Received 7 January 2014; received in revised form 20 May 2014; accepted 27 May 2014
Available online 6 June 2014

Abstract

In this paper, we present a study where a robot instructs a human on how to draw a route on a map. The human and robot are seated face-to-face with the map placed on the table between them. The user's and the robot's gaze can thus serve several simultaneous functions: as cues to joint attention, turn-taking, level of understanding and task progression. We have compared this face-to-face setting with a setting where the robot employs a random gaze behaviour, as well as a voice-only setting where the robot is hidden behind a paper board. In addition to this, we have also manipulated turn-taking cues such as completeness and filled pauses in the robot's speech. By analysing the participants' subjective rating, task completion, verbal responses, gaze behaviour, and drawing activity, we show that the users indeed benefit from the robot's gaze when talking about landmarks, and that the robot's verbal and gaze behaviour has a strong effect on the users' turn-taking behaviour. We also present an analysis of the users' gaze and lexical and prosodic realisation of feedback after the robot instructions, and show that these cues reveal whether the user has yet executed the previous instruction, as well as the user's level of uncertainty.

Keywords: Turn-taking; Feedback; Joint attention; Prosody; Gaze; Uncertainty

1. Introduction

Conversation can be described as a joint activity between two or more participants, and the ease of conversation relies on a close coordination of actions between them (c.f. Clark, 1996). Much research has been devoted to identify the behaviours that speakers attend to in order to achieve this fine-grained synchronisation. Firstly, any kind of interaction has to somehow manage the coordination of turn-taking. Since it is difficult to speak and listen at the same time, interlocutors take turns speaking and this turn-taking has to be coordinated (Sacks et al., 1974). Many studies have shown that turn-taking is a complex process where a number of different verbal and non-verbal behaviours including gaze, gestures, prosody, syntax and semantics influence the probability of a speaker change (e.g., Duncan, 1972; Kendon, 1967; Koiso et al., 1998). Secondly, in addition to the coordination of verbal actions, many types of dialogues also include the coordination of task-oriented non-verbal actions. For example, if the interaction involves instructions that need to be carried out, the instruction-giver needs to attend to the instruction-follower's task progression and level of understanding in order to decide on a future course of action. Thus, when speaking, humans continually evaluate how the listener perceives and reacts to what they say and adjust their future behaviour to accommodate this feedback. Thirdly, speakers also have to coordinate their joint focus of attention. Joint attention is fundamental to efficient communication: it allows people to interpret and predict each other's

* Corresponding author. Tel.: +46 87907874.
E-mail address: gabriel@speech.kth.se (G. Skantze).

http://dx.doi.org/10.1016/j.specom.2014.05.005
0167-6393/

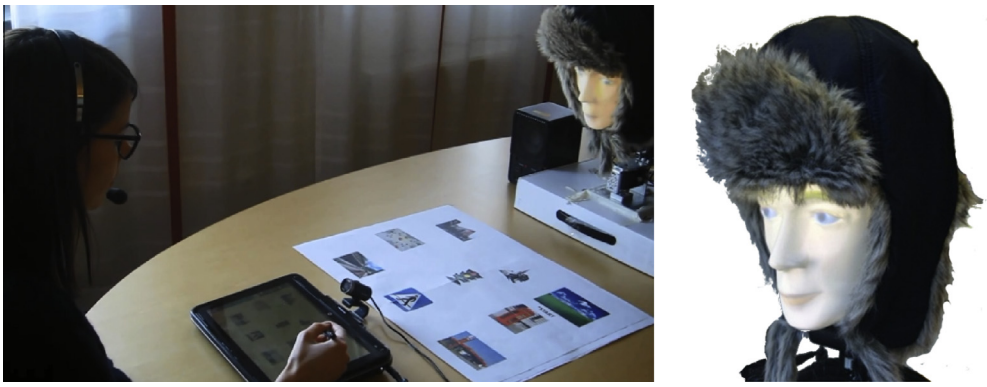

Fig. 1. The human–robot Map Task setup used in the study (left) and a close-up of the robot head Furhat (right).

actions and prepare reactions to them. For example, joint attention facilitates simpler referring expressions (such as pronouns) by circumscribing a subdomain of possible referents. Thus, speakers need to keep track of the current focus of attention in the discourse (Grosz and Sidner, 1986). In the case of situated face-to-face interaction, this entails keeping track of possible referents in the verbal interaction as well as in the shared visual scene (Velichkovsky, 1995).

Until recently, most computational models of spoken dialogue have neglected the physical space in which the interaction takes place, and employed a very simplistic model of turn-taking and feedback, where each participant takes the turn with noticeable pauses in between. While these assumptions simplify processing, they fail to account for the complex coordination of actions in human–human interaction outlined above. However, researchers have now started to develop more fine-grained models of dialogue processing (Schlangen and Skantze, 2011), which for example makes it possible for the system to give more timely feedback (e.g. Meena et al., 2013). There are also recent studies on how to model the situation in which the interaction takes place, in order to manage several users talking to the system at the same time (Bohus and Horvitz, 2010; Al Moubayed et al., 2013), and references to objects in the shared visual scene (Kennington et al., 2013).

These advances in incremental processing and situated interaction will allow future conversational systems to be endowed with more human-like models for turn-taking, feedback and joint attention. However, as conversational systems become more human-like, it is not clear to what extent users will pick up on behavioural cues and respond to the system in the same way as they would with a human interlocutor. In the present study we address this question. We present an experiment where a robot instructs a human on how to draw a route on a map, similar to a Map Task (Anderson et al., 1991), as shown in Fig. 1. The human and robot are placed face-to-face with a large printed map placed on the table between them. 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 of joint attention.

A schematic illustration of how speech and gaze could be used in this setting for coordinating turn-taking, task execution and attention (according to studies on human–human interaction) is shown in Fig. 2. In the first part of the robot's instruction, the robot makes an ambiguous reference to a landmark ("the tower"), but since the referring expression is accompanied with a glance towards the landmark on the map, the user can disambiguate this. At the end of the first part, the robot (for some reason) needs to make a pause. Since the execution of the instruction is dependent on the second part of the instruction, the robot produces turn-holding cues (e.g., gazes down and/or produces a filled pause) that inhibit the user to start drawing and/or taking the turn. After the second part, the robot instead produces turn-yielding cues (e.g., gazes up and/or produces a syntactically complete phrase) which encourage the user to react. After executing the instruction, the user gives an acknowledgement ("yeah") that informs the robot that the instruction has been understood and executed. The user's and the robot's gaze can thus serve several simultaneous functions: as cues to disambiguate which landmarks are currently under discussion, but also as cues to turn-taking, level of understanding and task progression.

In this study,[1] we pose the questions: Will humans pick up and produce these coordination cues, even though they are talking to a robot? If so, will this improve the interaction, and if so, how? To answer these questions, we have systematically manipulated the way the robot produces turn-taking cues. We have also compared the face-to-face setting described above with a setting where the robot employs a random gaze behaviour, as well as a voice-only setting where the robot is hidden behind a paper board. This way, we can explore what the contributions of a face-to-face setting really are, and whether they can be explained by the robot's gaze behaviour or the presence

[1] This article is an extension of Skantze et al. (2013a) and (2013b).

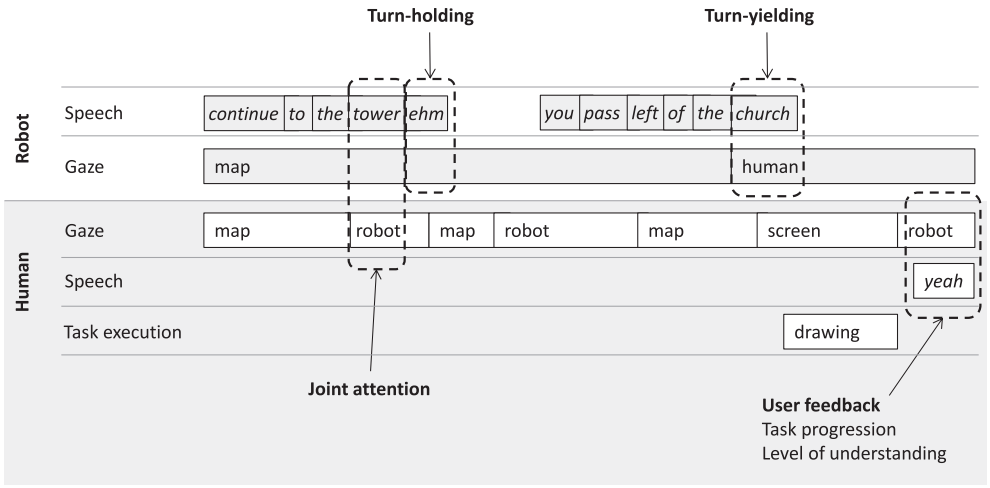

Fig. 2. The use of speech and gaze for coordinating turn-taking, task execution and attention.

of a face per se. A data-collection with 24 subjects interacting with the system has resulted in a rich multi-modal corpus. We have then analysed a wide range of measures: the users' verbal feedback responses (including prosody and lexical choice), gaze behaviour, drawing activity, subjective ratings and objective measures of task success.

The article is structured as follows. We start by reviewing previous research related to joint attention, turn-taking and feedback in human–human and human–machine interaction in Section 2. The review of these areas ends with five specific research questions that we have addressed in this study. We then describe the experimental setup, data collection and analysis in Section 3. The results from the analysis are presented Section 4, together with a discussion of how the results answer our five research questions. We then end with conclusions and discussion about future work in Section 5.

2. Background

2.1. Joint attention

In situated interaction, speakers naturally look at the objects which are under discussion. The speaker's gaze can therefore be used by the listener as a cue to the speaker's current focus of attention. Speakers seem to be aware of this fact, since they naturally use deictic expressions accompanied by a glance towards the object that is being referred to (Clark and Krych, 2004). In the same way, listeners naturally look at the referent during speech comprehension (Allopenna et al., 1998), and their gaze can therefore be used as a cue by the speaker to verify common ground. Thus, eye gaze acts as an important coordination device in situated interaction. This phenomenon is often referred to as "joint attention", which means that the speakers are attending to the same object and are mutually aware of it (Clark and Marshall, 1981; Baron-Cohen, 1995).

In situated human–machine interaction, the robot's or agent's gaze could be used as a cue to facilitate the user's comprehension of the robot's instructions. This could be regarded as a very basic form of joint attention (although one can argue about the "mutual awareness", cf. Staudte and Crocker, 2011). An example of this is the system presented in Nakano et al. (2003), where an animated agent describes a route to a user in a face-to-face setting, with a map in between, similar to the setup used in this study. However, a problem with animated agents on 2D displays is that it is impossible for the user to see exactly where the agent is looking, a problem typically referred to as the Mona Lisa effect (Al Moubayed et al., 2013). This is because the agent and user do not share the same physical space. Thus, in a multi-party setting, this means that the agent cannot establish exclusive mutual gaze with one of the users, and in situated interaction the object that is the target of the gaze cannot be inferred. However, this is not the case for physical (robot) heads, where the agent and user are physically co-present. In previous studies we have shown that if an animated face is back-projected on a mask – a technique which we also employ in this study – exclusive mutual gaze in multiparty dialogue is indeed possible (Al Moubayed et al., 2013). Also, when seated face-to-face at a table, humans can determine where on the table the robot is looking with a fairly high precision, close to reading the gaze of a human (Al Moubayed et al., 2013). However, it is one thing to determine the target of the gaze as an isolated task, and another to use it for facilitating comprehension of spatially ambiguous language. In a controlled experiment, Staudte and Crocker (2011) let subjects watch videos of a robot describing objects while gazing at them, showing that subjects were indeed able to utilize the gaze for resolving referring expressions. In another experiment, Boucher et al. (2012) let the iCub robot interact with human subjects and give them instructions while gazing at objects. The results showed that subjects reached objects faster when they could follow the gaze of the robot. However, in a study on infants, Okumura et al. (2013) found that while 12-month-olds understand the referential nature of human

gaze, this is not the case for robot gaze. This suggests that humans do not naturally interpret robot gaze in the same way as the gaze of other humans, but that this anthropomorphism might be learned. The extent to which gaze following comes naturally is also likely to depend on the design of the robot. In this study, we use the robot head Furhat (Al Moubayed et al., 2013), as seen in Fig. 1, which uses a facial animation model that is back-projected on a static mask. Compared to the robot heads used in the studies discussed above, Furhat is arguably more human-like. Another important difference from previous studies is that we here investigate how gaze-following can facilitate language disambiguation in a more complex ongoing dialogue (i.e., not just isolated instructions), where the gaze is also used for managing turn-taking. Thus, the first research question can be summarised as:

Q1: *Can the user utilize the gaze of a back-projected robot face in order to disambiguate ambiguous referring expressions in an ongoing dialogue?*

2.2. Generating turn-taking cues

Numerous studies have investigated how humans synchronise turn-taking in dialogue. In a seminal study, Duncan (1972) showed how speakers use prosody, syntax and gestures to signal whether the speaker wants to hold the turn or yield it to the interlocutor. For example, flat final pitch, syntactic incompleteness and filled pauses are strong cues to turn hold. In his analysis, Duncan found that as more turn yielding cues are presented together, the likelihood that the listener will try to take the turn increases. Later studies have presented more thorough statistical analyses of turn-yielding and turn-holding cues (Koiso et al., 1998; Gravano and Hirschberg, 2011; Hjalmarsson, 2011). Similar studies have been done to analyse the timing of *backchannels* (brief verbal responses, such as "uhu", or head nods), which are used to signal continued attention, but not to claim the floor (Ward and Tsukahara, 2003; Cathcart et al., 2003; Morency et al., 2010).

Contrary to this sophisticated combination of cues for managing turn-taking, dialogue systems have traditionally only used a fixed silence threshold, after which the system starts to process the utterance and generate a response. While this model simplifies processing, it fails to account for many aspects of human–human interaction such as hesitations, turn-taking with very short gaps or brief overlaps and backchannels in the middle of utterances (Heldner and Edlund, 2010). In order to analyse a wider set of cues for detecting where the user yields the turn, the system must run different levels of processing *asynchronously* and *incrementally* while the user is speaking (Schlangen and Skantze, 2011). This allows dialogue systems to interpret syntactic and prosodic cues to make continuous decisions on when to take the turn or give feedback, resulting in both faster response time and less interruptions (Skantze and Schlangen, 2009; Meena et al., 2013).

While there are several studies on how dialogue systems can *interpret* turn-taking cues, and find suitable places for backchannels in the *user's* speech, there are few studies that explore how such systems should *produce* such cues. If the system just produces simple questions or short answers, this might not be a big problem. However, in the case of longer instructions, the system should be able to pace the instructions to the suit the listener (Iwase and Ward, 1998). This can be done by chunking the instructions into smaller units, or *instalments* (Clark, 1996), after which a pause is produced and the user is invited to react and give feedback. In order to do this in a sophisticated manner, the system should be able to produce the spoken output *incrementally* (unit by unit), listening for user feedback and continuously adjusting its behaviour to the user's level of understanding (Buschmeier et al., 2012). Another benefit of incremental speech production is that the system can start to produce the spoken output before knowing exactly what to say. In a recent study, we implemented a model of incremental speech generation in a dialogue system (Skantze & Hjalmarsson, 2013). By allowing the system to generate and synthesise 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. If such a model of speech production is adopted, the system might have to produce pauses in the middle of instructions, where user reactions should be inhibited (e.g. if the initial system response was wrong and needs to be revised).

Thus, we can discriminate two kinds of pauses in the system's speech: pauses which are produced by the system in order to allow the user to execute some action and provide feedback, and pauses which the system has to make due to processing constraints. By utilising syntactic completeness and filled pauses – turn-taking cues typically found to have a strong effect in human–human interaction – it is possible that these two types of pauses can be employed. In the current study, we therefore ask the question:

Q2: *Can filled pauses and syntactic completeness be used as cues for inhibiting or encouraging user reactions in pauses?*

When it comes to face-to-face interaction, gaze has also been found to be an important cue for turn-taking. In one of the most influential publications on this subject, Kendon (1967) found that speakers in a face-to-face setting use gaze to signal whether they want to yield the turn. When initiating a turn, speakers typically gaze away. 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. Later studies have further supported these findings and also found that speakers' gaze patterns play an important role for the timing of backchannels, that is, backchannels are more likely to occur when the speaker is looking at the listener (Bavelas et al., 2002; Oertel et al., 2012). In multi-party interactions, gaze seems to be a strong cue for selecting the addressee of an utterance (Vertegaal et al., 2001). The importance of gaze in interaction is further supported by Boyle et al. (1994), who found that speakers in a Map Task setting interrupt each other less and use fewer turns, words, and backchannels per dialogue, if they interact face-to-face compared to when they cannot see each other.

In multi-party human–machine interaction (where several users interact with an animated agent or robot at the same time), studies have found that the users' gaze and head pose patterns can be used to infer whether the user is talking to the system or another person (Katzenmaier et al., 2004; Skantze and Gustafson, 2009). However, this effect is much weaker when the conversation involves objects in the shared visual scene (Johansson et al., 2013). It has also been found that the robot's gaze has a big influence on which user will speak next in multi-party interactions (Mutlu et al., 2006; Al Moubayed et al., 2013). When it comes to turn-taking in dyadic interactions, it has been shown that users interpret the gaze of an animated agent as a turn-yielding cue (Edlund and Beskow, 2009). In a controlled experiment where subjects were asked to listen to an animated agent and press a button at backchannel-relevant places, Hjalmarsson and Oertel (2012) found that when the agent gazed at the user, the user was more likely to press the button. However, we are not aware of any previous studies which explore the user's reactions to the system's gaze in combination with other turn-taking cues in an ongoing dialogue. In this study, we therefore ask the question:

Q3:*How does the robot's gaze affect the user's reactions to turn-taking cues?*

2.3. Interpreting user feedback

Feedback is an essential part of all communication. In human–human interaction, speakers continuously provide feedback to each other in the form of verbal expressions (e.g. backchannels, clarification requests), gaze and gestures (e.g. smile, head nods). On a shallow level, feedback can be used to just signal continued attention, but it can also be used to signal perception, understanding and acceptance (or lack thereof) (Allwood et al., 1992). A frequent type of verbal feedback, which we will focus on in this study, is *acknowledgements* – short utterances such as "yes", "okay" and "mhm" (Allen and Core, 1997). When these are used in an unobtrusive manner just to signal continued attention, they are typically referred to as *continuers* (Schegloff, 1982) or *backchannels* (Yngve, 1970), but they can also be used in task-oriented dialogues to signal that an action has been performed. While carrying little propositional content, it has been shown that acknowledgements play a significant role in the collaborative processes of dialogue (Schober and Clark, 1989). Furthermore, it has been found that the lexical and prosodic realisations of these tokens provide the speaker with information about the listener's attitude (Ward, 2004) and level of uncertainty (Lai, 2010; Neiberg and Gustafson, 2012).

As mentioned above, there are several studies on how to automatically find suitable places in the user's speech for the system to give acknowledgements. There are also studies on how to prosodically realise system acknowledgements in speech synthesis (Wallers et al., 2006; Stocksmeier et al., 2007). However, in this is study, we are interested in the problem of *interpreting user feedback*, based on prosody, lexical choice and gaze, in order to be able to adapt the system's output to the task progression and the user's level of understanding. While there are several examples of systems that pace or adapt their instructions to feedback from the user (Iwase and Ward, 1998; Nakano et al., 2003; Skantze and Schlangen, 2009; Reidsma et al., 2011; Buschmeier and Kopp, 2011), none of these perform any prosodic or lexical analysis of the user's acknowledgements in order to infer a more precise meaning.

In many task-oriented dialogue settings, one of the speakers has the role of an expert that guides the other speaker through some process in a step-wise manner. In this type of setting, acknowledgements do not only give the speaker information about the level of understanding, but also about whether an action has been completed. This setting is typical for many dialogue system domains, such as troubleshooting (Boye, 2007) and turn-by-turn navigation (Boye et al., 2012). For such systems, it is essential to provide the instructions in appropriately sized chunks and in a timely manner, to avoid overloading the user with information and to give the user enough time to complete the requested action. When the user gives an acknowledgement after such an instruction, it might be hard to determine whether the user has actually executed the action. The following example from Boye (2007), where a system for automated broadband support is presented, illustrates the problem:

1. **System:** Can you locate the telephone plug in the wall
2. **User:** Uh, yes
3. **System:** One of the cables going from the telephone plug should lead to a little box that probably has some lights on it
4. **User:** Ok

The acknowledgements in utterance 2 could either mean that the user has understood the instruction and will execute it, or it could mean that the user has executed it. This should have an effect on how long the system should wait

before proceeding with utterance 3. To our knowledge, there are no previous studies that investigate if it is possible to infer this from how the acknowledgement is realised. In the current study, we therefore ask the question:

Q4: *Can the users' gaze as well as their lexical and prosodic realisation of acknowledgements be used to discriminate between acknowledgements uttered prior to and after action completion in task-oriented human–robot interaction?*

Feedback can also reveal the speaker's level of uncertainty. In the example above, even if utterance 4 signals that the user thinks he has found the box, he could be more or less confident that he has actually found the right box. Although there are several studies on how uncertainty is realised in speech in general (Liscombe et al., 2006; Pon-Barry, 2008), there are very few studies which specifically investigate acknowledgements. One exception is Lai (2010), who found that different intonation contours of cue words (e.g. "yeah", "right", "really") influence listeners' perception of uncertainty. There are also very few examples of studies of how uncertainty is expressed in human–machine dialogue. One example is Forbes-Riley and Litman (2011) who adjusted the content of the output in a tutoring system based on the students' level of uncertainty. However, they did not specifically analyse acknowledgements. In this study, we therefore ask the question:

Q5: *Can the users' gaze as well as their lexical and prosodic realisation of acknowledgements be used to determine the user's level of uncertainty?*

3. Human–robot Map Task data

In order to address the questions outlined above, we collected a multimodal corpus of human–robot Map Task interactions. Map Task is a well establish experimental paradigm for collecting data on human–human dialogue (Anderson et al., 1991). 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) 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 setting was chosen because it naturally gives rise to the different phenomena we have set out to explore. First, since the human and robot are seated face-to-face talking about landmarks on a map, joint attention is likely to help in identifying the landmarks under discussion. Second, since the robot has the initiative when giving instructions, it allows us to systematically manipulate how the robot generates turn-taking cues, and measure how subjects react to these. Third, these instructions naturally evoke a large set of naturally occurring feedback utterances (mostly acknowledgements) in different contexts, which allow us to do statistical analysis of their form and function.

3.1. A Map Task dialogue system

The experimental setup is shown in Fig. 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 and Al Moubayed, 2012; www.iristk.net), which provides a set of modules for input and output, including control of Furhat (facial gestures, eye and head movements), as well as an XML-based language for authoring the flow of the interaction. For speech synthesis, we used the CereVoice unit selection synthesiser developed by CereProc (www.cereproc.com).

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*. This is also illustrated in Fig. 2. 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

Given the current limitations of conversational speech recognition, and the 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 Fig. 3. 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 Fig. 3).

The system's behaviour and the design of the maps were tuned during a series of pilot studies, which resulted in a convincing system behaviour and smooth interaction. In general, since the hidden spots were placed at the landmark locations pointed out in the route descriptions, they worked very well for locating the users' drawing activity. These pilot studies also 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.

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 s.
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. Most of the subjects were students from the School of Computer Science and Communication at KTH, but none of them had any longer experience with interacting with robots.

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

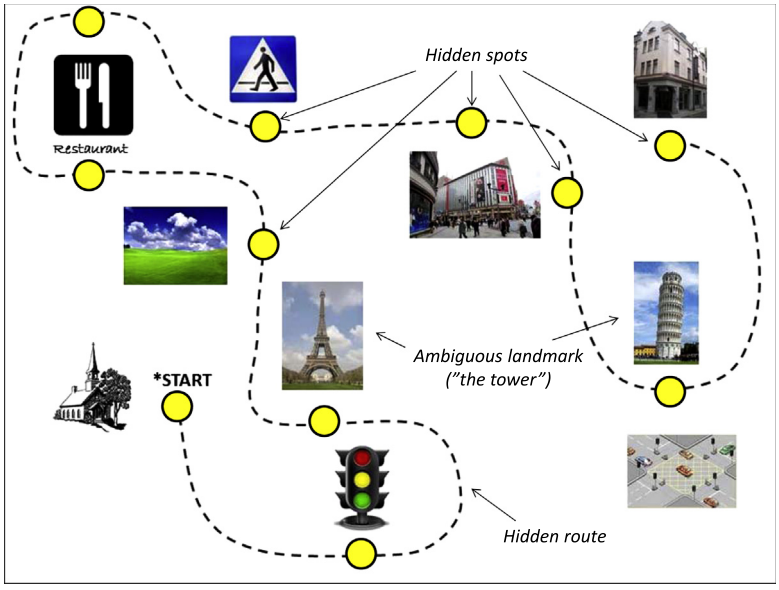

Fig. 3. An example map.

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 (<1 s.)	At the wrong spot	CONTINUE (with misunderstanding)
Long (>1 s.)	At the wrong spot	RESTART or CHECK
No resp.	Any	CHECK

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. The reason for splitting the CONSISTENT/RANDOM conditions between groups is because there is a risk that subjects otherwise would adapt to the version they use first. For example, if they use the RANDOM version first, they might stop paying attention to the robot's gaze, which would also affect the interaction with the CONSISTENT version. 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. These three categories were very clear from watching the video of the subject, so multiple annotations were not deemed necessary.

In order to analyse user feedback, all dialogues were manually transcribed and one-word acknowledgements were identified. As acknowledgements we included all words transcribed as "okay", "yes", "yeah", "mm", "mhm", "ah", "alright" and "oh". Although there were other possible candidates for this category, their frequencies were very low. For each acknowledgement, we extracted the pitch using ESPS in Wavesurfer/Snack (Sjölander and Beskow, 2000) and converted the values into semitones. In order to get a measure of **pitch slope**, we calculated the difference between the average of the second half of these values and the average of the first half for each token (i.e. negative = falling, positive = rising). Each value was then z-normalised based on the overall data for the speaker. Using these values, the **average (normalised) pitch** was calculated for each token. A measure of

duration was also calculated by counting the number of voiced frames (each 10 ms) for the token. To measure the **average (normalised) intensity**, we used Praat (Boersma, 2001) and then calculated the average dB (z-normalised for the speaker). The intensity measures used were extracted from voiced intervals only.

4. Results

4.1. Joint attention

We first address **Question 1**: *Can the user utilize the gaze of a back-projected robot face in order to disambiguate ambiguous referring expressions in an ongoing dialogue?*

By comparing the main conditions in Group A (FACE/CONSISTENT vs. NOFACE) with Group B (FACE/RANDOM vs. NOFACE), and measuring what effect they have on the users' subjective ratings, task completion and drawing activity, we can investigate whether the users utilised the gaze of the robot in order to disambiguate ambiguous referring expressions. If so, we should see a positive effect on these measures in the FACE condition in Group A (where the robot gazed at the landmarks while referring to them), but not in Group B (where the robot did not look at any place in particular).

First, 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 normalised to a scale between 0 and 1. A Wilcoxon Signed Ranks Test was carried out, using these differences for ranking.[2] 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 ($n = 8$) found it easier to complete the task in the FACE condition than in the NOFACE 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 Fig. 4.

Apart from the subjective ratings, we also wanted to see whether reading the robot's gaze affects 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 s faster in the FACE condition than in the NOFACE condition. For Group B

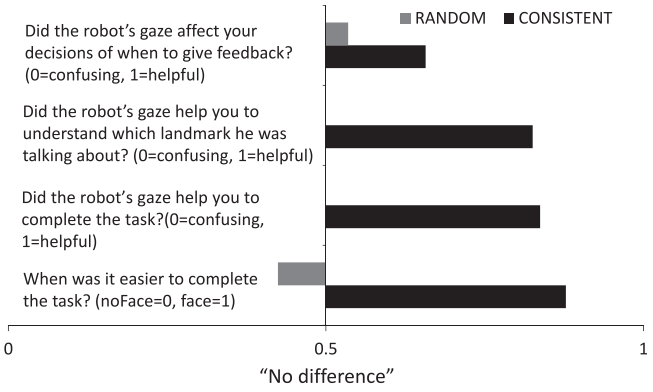

Fig. 4. The results from the questionnaire. The bars show the median rating for Group A (consistent) and Group B (random).

(RANDOM), the dialogues were instead 2.3 system utterances and 17.3 s longer in the FACE condition (Mann–Whitney U-test, $p < .05$). Thus, it seems like the presence of 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 task 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 ($\chi^2(1, N = 25) = 4.8$, $p = .028$; $\chi^2(1, N = 48) = 6.75$, $p = .0094$). This indicates that the users with the RANDOM/FACE condition had more difficulties to follow the system's instructions than the users with the CONSISTENT/FACE condition, most likely because they did not get guidance from the robot's gaze in disambiguating referring expressions.

The analyses above only take into account the full dialogues, and therefore do not tell us whether it is the gaze reading during ambiguous references per se that actually improves task completion. Therefore, we also investigated the users' drawing activity during the ambiguous references. The mean drawing activity over the four phases of the descriptions of ambiguous landmarks is plotted in Fig. 5. 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 normalised in order to make it possible to analyse the average user behaviour. Thus, the time on the x-axis is not linear. 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 the robot's

[2] Analyses of the different measures used throughout Sections 4.1 and 4.2 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.

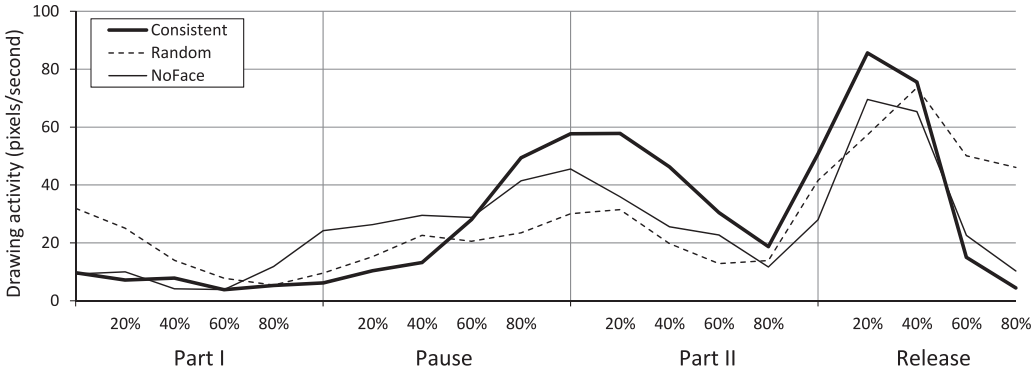

Fig. 5. Average drawing activity during ambiguous references depending on condition (Group A and B). *X*-axis shows the relative time as percentage of each phase.

gaze at the target landmark during ambiguous references makes it possible for the subjects to start to draw quicker.

To sum up, the results show that the robot's gaze behaviour (looking at the landmark when making references to them) was rated as helpful rather than confusing for task progression and landmark identification, compared to interacting with voice only. These effects were not present when the robot did not look at any landmark in particular (the RANDOM condition). 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 when system's reference was ambiguous, suggest that the users indeed utilised the system's gaze to discriminate between landmarks.

4.2. Generating turn-taking cues

Next, we investigate the effect of the robot's turn-taking cues on the users' responses in pauses. Thus, we address research questions Q2 (*Can filled pauses and syntactic completeness be used as cues for inhibiting or encouraging user reactions in pauses?*) and Q3 (*How does the robot's gaze affect the user's reactions to turn-taking cues?*). To answer these questions, we use data from Group C, where the pauses are longer (2–4 s) and the robot used a consistent gaze behaviour (always gazing down at landmarks during pauses, looking up at the user after completing the instruction).

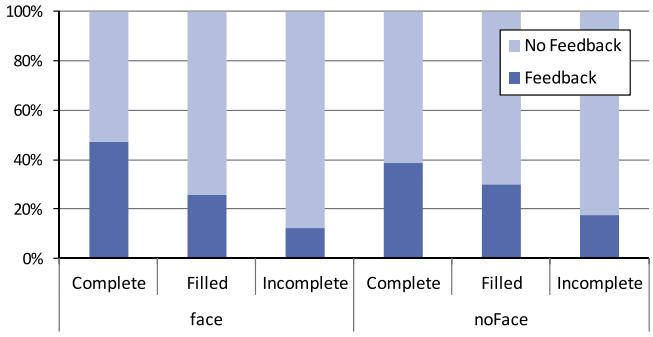

Fig. 6. Presence of feedback depending on pause type (Group C).

First, we investigated whether syntactic completeness before pauses had an effect on whether the users gave verbal responses in the pause. Fig. 6 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 ($\chi^2(2, N = 157) = 10.32$, $p = .0057$). In fact, the difference is not significant in the NOFACE condition ($p > .05$).

Since we are investigating a multi-modal face-to-face setting, we also wanted to analyse the effect of turn-taking cues on the participants' gaze. In this analysis, it turned out that there was no difference in the users' gaze between FILLED and INCOMPLETE (which both have a turn-holding function). The percentage of gaze at the robot over the four different utterance phases for complete and incomplete utterances is plotted in Fig. 7. For each phase, a Mann–Whitney *U*-test was conducted. The results show that the percentage of gaze at the robot 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.

As pointed out above, in a situated task-oriented setting, turn-taking is not just about taking turns in speaking, but also about managing the execution of the task. Thus, we would expect the robot's turn-taking cues to not just affect the verbal and gaze behaviour of the user, but also the drawing activity. If the system produces a cue indicating incompleteness, the user should to a larger extent stop the drawing activity and await the complete instruction. For this analysis, FILLED and INCOMPLETE have again been merged (since they have a similar function and there was no clear difference in their effect). The mean drawing activity over the four phases of the descriptions is plotted in

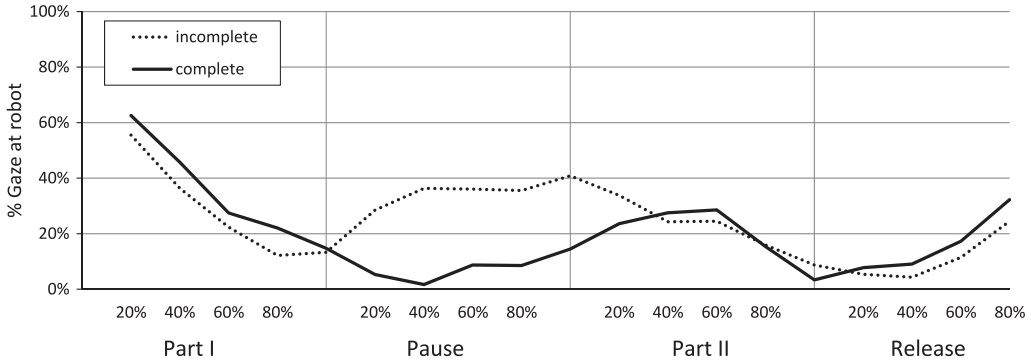

Fig. 7. Average user gaze depending on pause type (Group C). *X*-axis shows the relative time as percentage of each phase.

Fig. 8. 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. Note also that the users were aware of the fact that they were affected by the robot's gaze, as can be seen in the result analysis of the questionnaire shown in Fig. 4.

These results now allow us to answer **Questions 2 and 3**:

Q2: *Can filled pauses and syntactic completeness be used as cues for inhibiting or encouraging user reactions in pauses?*

The results show that pauses preceded by incomplete syntactic segments or filled pauses appear to inhibit user activity. In these pauses, users give less feedback and draw less. During these segments, they also look at the robot to a larger extent, which further indicates that they await further instructions before acting. 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. Comparing filled pauses with syntactic incompleteness, the results show that incomplete phrases have an even more inhibiting effect on the likelihood of the user giving feedback.

Q3: *How does the robot's gaze affect the user's reactions to turn-taking cues?*

The results show that the effects of syntactic completeness versus incompleteness and filled pauses on drawing activity and feedback outlined above were more pronounced in the face-to-face setting than in the voice only setting. Indeed, in the voice only setting, incomplete phrases do not have an inhibiting effect on the user's drawing activity, and thus does not seem to be a strong enough inhibiting cue in itself. In the face-to-face setting, on the other hand, when the face consistently looked down at the map during pauses, the difference between complete versus incomplete phrases is big. This suggests that the combined cues of looking down and producing an incomplete phrase have an inhibiting effect.

4.3. User feedback and task progression

In the previous section, we investigated how the system can elicit feedback from the user. We now turn to the inter-

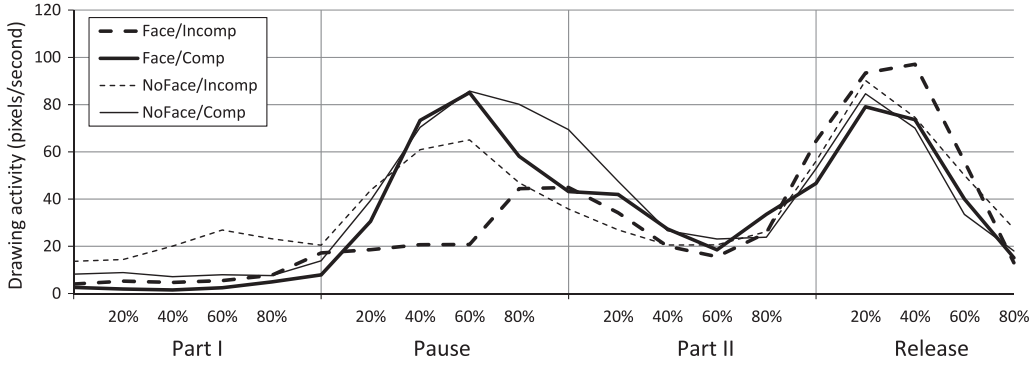

Fig. 8. Average drawing activity depending on pause type and the presence of the face (Group C). *X*-axis shows the relative time as percentage of each phase.

pretation of user feedback. As discussed in Section 2.3 above, in order for the system to know how long it should wait before proceeding with the instructions, it should be able to detect whether an acknowledgement indicates that the previous instruction has been executed (in which case it should proceed immediately) or is about to be executed (in which case it should wait). Thus, we address **Question 4:** *Can the users' gaze as well as their lexical and prosodic realisation of acknowledgements be used to discriminate between acknowledgements uttered prior to and after action completion in task-oriented human–robot interaction?*

To investigate this, we defined four different classes of acknowledgements: (1) those that signal that the activity is about to be initiated (**before drawing**), (2) those that are given **while drawing**, (3) those that signal that that the activity has been completed (**after drawing**), or (4) those that signal that the activity has already been completed in a previous step (**no drawing**). Each acknowledgement was automatically assigned to one of these categories, based on the accompanying drawing activity, as illustrated in Fig. 9. For a pen movement to be considered as a drawing activity, it had to encompass at least 50 pixels, either in the time window between the end of the system's instruction and the middle of the acknowledgement, or in the window between the middle of the acknowledgement and three seconds after. In total, there were 1568 feedback tokens in the whole dataset. The prosodic extraction failed for some tokens, so for the prosodic analysis we were able to use 1464 tokens.

Intuitively, the time it takes from the end of the system's instruction until the user gives the acknowledgement (i.e., response time) should correlate strongly with task progression. A bit surprisingly, this was not the case. Fig. 10 illustrates the distribution of the four classes of acknowledgements in relation to the response time. As can be seen, the four classes are fairly evenly distributed across different time segments, especially for early acknowledgements (which are also the most common.

Next, we investigated the relationship between the lexical realisation of acknowledgements and drawing activity, which is illustrated in Fig. 11. A chi-square test showed that there are significant differences in the distribution of lexical tokens between the different drawing activity classes ($\chi^2(21, N = 1568) = 248.19$, $p < .001$). For example, a "yes" is more likely to signal that the activity has already been completed (*no drawing*), whereas "okay" shows an opposite distribution. The use of "mhm" is more common before or while executing an action.

To explore the relationship between the prosodic realisation of the acknowledgement and the different drawing activity classes, we used a MANOVA. The prosodic features described in Section 3.2 were used as dependent variables and the drawing activity class was used as the independent variable. The results show a general significant effect ($F(28, 5240) = 39.96$; Wilk's $\Lambda = 0.50$; $p < .05$), as well as several individual significant effects in post hoc tests (Tukey HSD). All prosodic features except average pitch showed effects. Significant results are summarised in Tables 3 and 4. For example, acknowledgements with "no drawing" have a higher intensity, shorter duration and a relatively flat pitch. Acknowledgements "before drawing" and "while drawing" both have a lower intensity and longer duration, but differ in that "while drawing" has a higher rising pitch. Since the uneven distribution of lexical tokens may influence these differences, we also conducted separate MANOVA analyses for the two most frequent tokens – "okay" and "yes" – which also showed significant effects. As can be seen in the post hoc tests reported in Table 4, the general patterns are the same, even though some individual effects are different.

For the face-to-face conditions, we also counted the number of times that the user gazed at the robot in vicinity of the feedback (within a window between 1 s before and 1 s after the token) and a chi-square test was used to investigate the relationship between gaze and drawing activity.

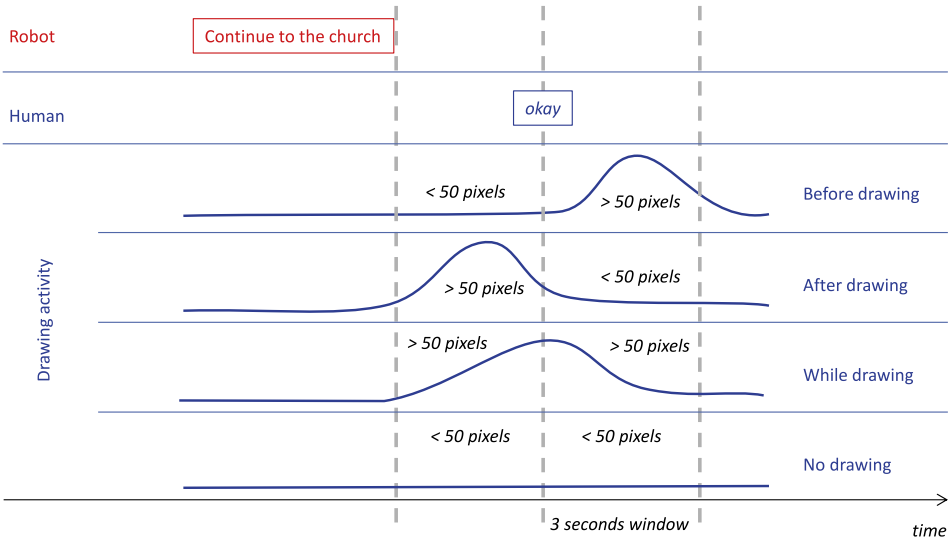

Fig. 9. Illustration of the definition of the four classes of acknowledgements, as related to task progression (in terms of drawing activity).

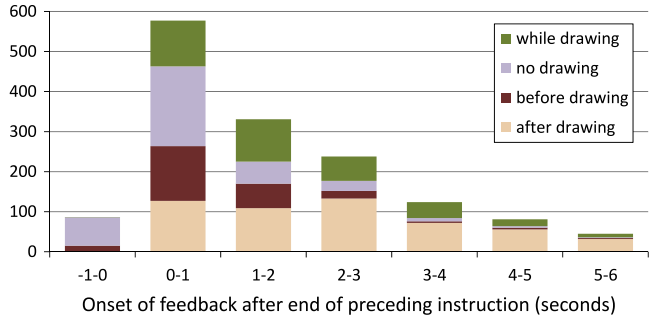

Fig. 10. Acknowledgement frequency and activity completion over response time.

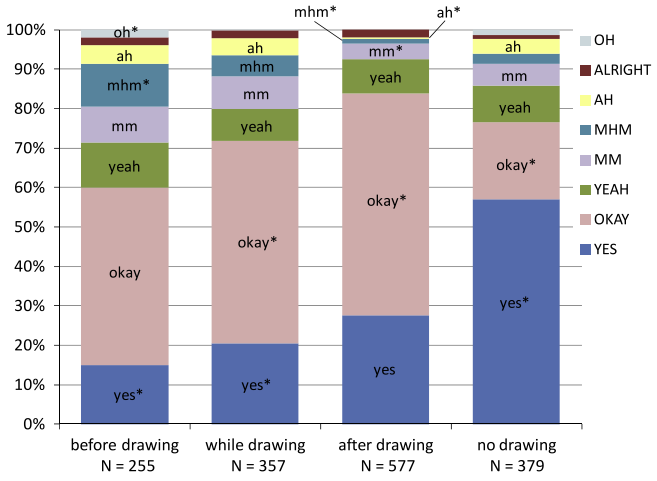

Fig. 11. Distribution of lexical tokens depending on activity completion. * Marks significant differences from the overall distribution.

For this analysis, we could only use the dialogues with a face-to-face condition, which resulted in 807 tokens. The results are presented in Table 5, showing clear significant effects ($\chi^2(3, N = 807) = 55.99$, $p < .001$) – users tend to look more at the robot when they do not have to draw, and even more when they have completed the drawing activity. This is in line with general gaze patterns found in turn-taking behaviour (Kendon, 1967; Oertel et al., 2012).

While the results above show an interesting correlation between the form and function of acknowledgements, they do not tell us whether the different parameters are redundant. To analyse this, we used a logistic regression model with prosody, lexical token, and response time as features. We did not use gaze, as this would only allow us to use half of the data. To implement the model, the Weka machine learning toolkit was used (Hall et al., 2009). In order to evaluate how well the model generalises over subjects, a 24-fold cross validation was used, where the model was trained on 23 subjects and evaluated on the 24th for each fold. Since the classes BeforeDrawing and WhileDrawing show similar characteristics and would arguably result in similar system reactions, they were collapsed into one class. Thus, the model had to distinguish between three classes: AfterDrawing ($N = 549$), BeforeOrWhile ($N = 572$) and NoDrawing ($N = 343$). The performances of the different combinations of features are shown in Table 6. We also include the majority class baseline for comparison.

As the results show, the combination of prosody, response delay and dialogue context gives the best performance, which is a clear improvement over the majority class baseline (62.1% vs. 28.2%). Interestingly, the lexical token is redundant and does not contribute in this combination, which means that if such a classifier would be used online in the system, ASR would not be needed for this.

To sum up, the delay between an instruction and an acknowledgement is not in itself a very good indicator of whether the instruction has been or is about to be executed. However, if the prosodic realisation of the acknowledgement is also taken into account, task progression can to some extent be inferred. This shows an important aspect of the form and function of acknowledgements that we have not seen being discussed before in the literature.

Table 4
Post hoc analyses for differences in effect of drawing activity on prosody.

Token	Parameter	Post hoc effects
All	Intensity	No > (After, While, Before) After > Before
	Duration	(While, Before) > After > No
	Pitch Slope	(While, After) > (Before, No)
"yes"	Intensity	No > (While, After)
	Duration	While > (Before, After, No)
	Pitch Slope	While > No
"okay"	Intensity	*No significant differences*
	Duration	While > (After, No)
	Pitch Slope	After > (Before, No)

Table 3
The relationship between drawing activity and prosody.

	Before Drawing	While Drawing	After Drawing	No Drawing
N	238	334	549	343
Intensity	Low	Medium	Medium	High
(z-score)	$M = -0.19$	$M = -0.08$	$M = 0.00$	$M = 0.25$
	SD = 0.64	SD = 0.53	SD = 0.52	SD = 0.69
Duration	Long	Long	Medium	Short
(voiced frames)	$M = 27.30$	$M = 28.11$	$M = 22.19$	$M = 18.97$
	SD = 12.97	SD = 12.51	SD = 9.68	SD = 10.91
Pitch Slope	Flat	Rising	Rising	Flat
(semitones)	$M = 0.97$	$M = 1.98$	$M = 1.76$	$M = 0.35$
	SD = 3.05	SD = 3.14	SD = 3.21	SD = 3.07

Table 5
The relationship between drawing activity and gaze.

	Before drawing	While drawing	After drawing	No drawing
N	131	151	321	204
Gaze at robot	34.4%	37.1%	66.0%	55.4%
	SR = −2.9	SR = −2.7	SR = 3.3	SR = 0.5
	Post hoc: After > No > Before, While			

Table 6
The performance of the logistic regression classifier for determining task progression, using different feature sets. Performance is shown in terms of accuracy (percent correctly classified instances), as well as weighted average *F*-score and area under the ROC curve (AUC) for the different classes.

Feature set	Accuracy	*F*-score	AUC
Prosody + respDelay	62.1	0.629	0.765
RespDelay	49.9	0.495	0.679
Token	47.5	0.443	0.621
Prosody	46.8	0.471	0.634
Majority class baseline	28.2	0.153	0.500

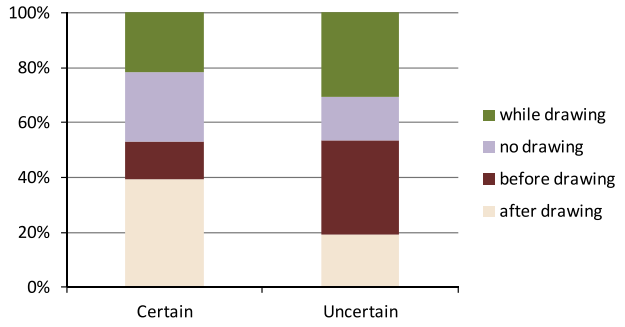

Fig. 12. Distribution of task progression vs. uncertainty.

4.4. User feedback and uncertainty

As described in Section 2.3 above, acknowledgements can also reveal the speaker's level of uncertainty. If uncertainty is detected, the system could for example give the user some extra time to execute the task, or check that the user has actually understood the instruction. Thus, we now address **Question 5:** *Can the users' gaze as well as their lexical and prosodic realisation of acknowledgements be used to determine the user's level of uncertainty?*

In order to analyse uncertainty in acknowledgements, a binary distinction between **certain** and **uncertain** was made, and all acknowledgement tokens were manually annotated for these two categories, resulting in 1376 certain tokens and 192 uncertain tokens. In order to validate the annotation scheme, we randomly selected 110 pairs of certain/uncertain tokens in the same lexical category. The annotators did not get any dialogue context and were not given information about the other annotators' labelling. The cross-annotator agreement was measured in a multirater kappa analysis (Randolph, 2005), which resulted in a kappa score of 0.63 – a substantial agreement according to Landis and Koch (1977).

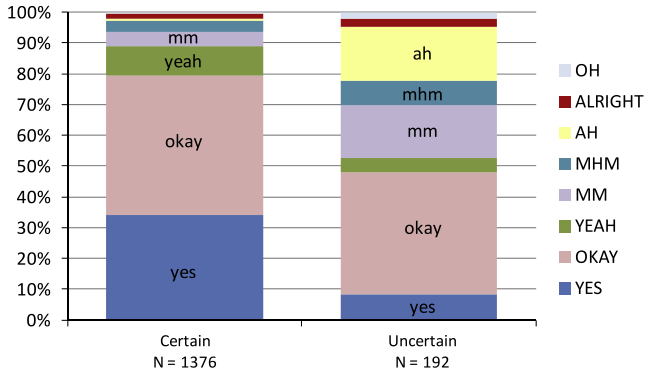

Fig. 13. Distribution of lexical tokens depending on uncertainty.

Table 7
Significant differences in prosody between certain/uncertain for all tokens in general, and for "okay" in particular.

	All tokens		"okay"	
	Certain	Uncertain	Certain	Uncertain
N	1280	184	595	70
Intensity	Medium	Low	Low	Low
(*z*-score)	*M* = 0.08	*M* = −0.47	*M* = −0.83	*M* = −0.53
	SD = 0.57	SD = 0.64	SD = 0.45	SD = 0.55
Duration	Medium	Long	Medium	Long
(voiced frames)	*M* = 22.01	*M* = 34.71	*M* = 23.44	*M* = 27.96
	SD = 9.80	SD = 17.16	SD = 6.78	SD = 12.68
Pitch slope	Rising	Flat	Not significant	
(semitones)	*M* = 1.43	*M* = 0.82		
	SD = 3.24	SD = 2.90		

Table 8
Certain/uncertain vs. user's gaze.

	Certain	Uncertain
N	710	97
Gaze at robot	55.9%	29.9%
	SR = 1.1	SR = −3.1

First, we wanted to investigate the relationship between the notion of task progression explored above and that of uncertainty. A chi-square test revealed that the distributions of task progression were different between certain and uncertain acknowledgements ($\chi^2(3, N = 1568) = 75.1$, $p < .001$), as illustrated in Fig. 12. Not surprisingly, uncertain acknowledgements occur more frequently before drawing, whereas certain acknowledgements are more common after drawing.

Similar to the analysis of task progression above, we wanted to investigate the relationship between the lexical realisations and uncertainty. A chi-square test showed a difference in the distribution of lexical tokens between the certain and uncertain categories ($\chi^2(7, N = 1568) = 272.87$, $p < .001$). As illustrated in Fig. 13, the distribution varies a lot – all lexical tokens except "alright" and "okay" are significantly different between the *certain* and *uncertain* categories.

Next, we wanted to see how uncertainty is related to the prosodic features described in Section 3.2, as well as the drawing activity. All these features were used as dependent variables in a MANOVA, which showed a general significant effect ($F(6, 1457) = 71.85$; Wilk's $\Lambda = 0.77$; $p < .05$). The results are summarised in Table 7. All prosodic features except average pitch show effects. "Uncertain" acknowledgements have lower intensity, longer duration, and a more flat pitch. Table 7 also shows a separate analysis of "okay" (which was equally common in both categories and frequent enough). As can be seen, all significant differences except pitch slope remain. However, for intensity the trend now goes in the opposite direction – "okay" has a lower intensity when expressed as "certain". We do not have a good explanation for this, and this difference needs to be investigated further.

For the face-to-face conditions, we also counted how many times the user gazed at the robot in vicinity of the acknowledgement, depending on uncertainty. The results are shown in Table 8, and show significant effects ($\chi^2(1, N = 807) = 23.18$, $p < .001$) – users tend to look more at the robot when they are certain than when they are uncertain.

Similarly to the case of task progression, we explored the use of logistic regression to explore how well the features generalise over subjects, and whether the different parameters are redundant, using the same features sets. However, in this case, there are only two classes: Certain and Uncertain. The results are reported in Table 9.

Since most utterances were perceived as Certain, the dataset is somewhat unbalanced, which gives a majority class baseline of 87.9%. Still, the combination of prosody and token gives a significantly better accuracy of 90.9% (Wilcoxon signed-rank test; $Z = -3.24$, $p < 0.05$). In this case, dialogue context and response delay do not contribute to the performance. As can be seen, the area under the ROC curve is much higher for the classifier than the baseline, indicating that the balance between precision and recall can be tuned, depending on the dialogue costs imposed by false positives vs. false negatives.

Table 9
The performance of the logistic regression classifier for determining uncertainty, using different feature sets. The performance is shown in terms of accuracy (percent correctly classified instances), as well as weighted average *F*-score and area under the ROC curve (AUC) for the different classes.

Feature set	Accuracy	*F*-score	AUC
Prosody + token	90.9	0.892	0.804
Prosody	89.6	0.871	0.782
RespDelay	87.9	0.825	0.467
Token	87.1	0.836	0.670
Majority class baseline	87.9	0.825	0.479

To sum up, the results show that uncertainty can to some extent be inferred from the realisation of the acknowledgement. For example, uncertainty is more often expressed with "mm" and "ah", whereas certainty is more often expressed with "yes". In the cases where the distributions are similar, as for "okay", the prosodic patterns could be used to identify uncertain user feedback. For example, uncertainty seems to be associated with longer duration, a finding which is line the analysis reported by Ward (2004). The users' gaze also seems to be informative – users gaze more at the robot when they are certain.

5. Conclusions and future work

In this study, we have investigated to which extent humans produce and respond to human-like coordination cues in face-to-face interaction with a robot. We did this by conducting an experiment where a robot gives instructions to a human in a Map Task-like scenario. By manipulating the robot's gaze behaviour and whether the user could see the robot or not, we were able to investigate how the face-to-face setting affects the interaction. By manipulating the robot's turn-taking cues during pauses, we were able to investigate how these affected the users' drawing activity, gaze and verbal feedback behaviour. Overall, the results show that users indeed are affected by these cues and that the face-to-face setting enhances these effects. Moreover, the face-to-face setting allows the user to utilize the robot's gaze to disambiguate references to landmarks on the map. In addition to this, we have explored the users' realisation of verbal feedback and showed that the realisation (lexical and prosodic) to some extent reveal whether the systems instructions have been completed and the user's level of uncertainty.

This study is novel in several ways. First of all, in most previous studies related to turn-taking and feedback, the user has the initiative and the system has the role of responding to the user's requests or providing feedback to the user (Huang et al., 2011; Meena et al., 2013). Hence, these studies typically investigate how the system should be able to detect turn-taking or feedback-inviting cues in the users' speech, or how to produce adequate and timely feedback. There are relatively few previous studies where the roles are the opposite, and which therefore study how to interpret user feedback or how the system should produce turn-management cues. We think that the method presented here is very useful for studying these phenomena, since it does not depend on conversational speech recognition. Moreover, while there are previous studies that investigate whether humans can disambiguate referring expressions using gaze, we have not seen any studies which show that humans can utilize the robot's gaze for disambiguation in an ongoing dialogue. When it comes

to interpreting user feedback, we are not aware of any other studies that have explored how users' gaze patterns and prosodic realisations of acknowledgements are related to task progression and uncertainty in a human–machine dialogue setting.

The results presented in this study have implications for generating multimodal behaviours incrementally in dialogue systems. 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, and then let the user's feedback determine the system's future course of actions. For example, in a tutoring application the system could react to an uncertain user response in rephrasing the instruction rather than moving on. Also, the problem of knowing whether an action has been completed is of course not limited to drawing a route on a map, but should be applicable to many types of task-oriented dialogue settings, such as turn-by-turn navigation or troubleshooting.

In future studies, we plan to extend our previous model of incremental speech generation (Skantze and Hjalmarsson, 2013) with these capabilities. In that model, we separated the speech planning from the real-time realisation of the speech plan. Depending on the level of commitment in the speech plan, the component responsible for realisation could automatically produce turn-holding or turn-yielding cues, and automatically adapt the realisation of the plan to feedback from the user. The results from this study show that users are likely to react to these cues and that it is indeed possible to infer task progression and uncertainty form the user's feedback. We think that the experimental setup we have presented in this article will be an excellent test-bed for testing such a model.

Acknowledgements

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. Int. J. Humanoid Rob. 10 (1).

Allen, J.F., Core, M., 1997. Draft of DAMSL: Dialog act Markup in Several Layers. Unpublished manuscript.

Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K., 1998. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. J. Mem. Lang. 38 (4), 419–439.

Allwood, J., Nivre, J., Ahlsen, E., 1992. On the semantics and pragmatics of linguistic feedback. J. Semantics 9 (1), 1–26.

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., . The HCRC map task corpus. Lang. Speech 34 (4), 351–366.

Baron-Cohen, S., 1995. The eye direction detector (EDD) and the shared attention mechanism (SAM): two cases for evolutionary psychology. In: Moore, C., Dunham, P.J. (Eds.), Joint Attention: Its Origins and Role in Development. Erlbaum, Hillsdale, NJ, pp. 41–60.

Bavelas, J., Coates, L., Johnson, T., 2002. Listener responses as a collaborative process: the role of gaze. J. Commun. 52 (3), 566–580.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5 (9/10), 341–345.

Bohus, D., Horvitz, E., 2010. Facilitating Multiparty Dialog with Gaze, Gesture, and Speech. In: Proc ICMI'10. Beijing, China.

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. Front. Neurorobotics 6.

Boye, J., 2007. Dialogue management for automatic troubleshooting and other problem-solving applications. In: Proceedings of the 8th SIGDial Workshop on Discourse and Dialogue, Antwerp, Belgium.

Boye, J., Fredriksson, M., Götze, J., Gustafson, J., Königsmann, J., 2012. Walk this Way: Spatial Grounding for City Exploration. In: IWSDS2012 (International Workshop on Spoken Dialog Systems).

Boyle, E., Anderson, A., Newlands, A., 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. Lang. Speech 37 (1), 1–20.

Buschmeier, H., Kopp, S., 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In: Proceedings of IVA, Reykjavik, Iceland, pp. 169–182.

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, Seoul, South Korea, pp. 295–303.

Cathcart, N., Carletta, J., Klein, E., 2003. A shallow model of backchannel continuers in spoken dialogue. In: 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest.

Clark, H.H., 1996. Using Language. Cambridge University Press, Cambridge, UK.

Clark, H.H., Krych, M.A., 2004. Speaking while monitoring addressees for understanding. J. Mem. Lang. 50, 62–81.

Clark, H.H., Marshall, C.R., 1981. Definite reference and mutual knowledge. In: Joshi, A.K., Webber, B.L., Sag, I.A. (Eds.), Elements of Discourse Understanding. Cambridge University Press, Cambridge, England, pp. 10–63.

Duncan, S., 1972. Some signals and rules for taking speaking turns in conversations. J. Pers. Soc. Psychol. 23 (2), 283–292.

Edlund, J., Beskow, J., 2009. MushyPeek – a framework for online investigation of audiovisual dialogue phenomena. Lang. Speech 52 (2–3), 351–367.

Forbes-Riley, K., Litman, D., 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Commun. 53 (9–10), 1115–1136.

Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Comput. Speech Lang. 25 (3), 601–634.

Grosz, B.J., Sidner, C.L., 1986. Attention, intentions, and the structure of discourse. Comput. Linguist. 12 (3), 175–204.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1).

Heldner, M., Edlund, J., 2010. Pauses, gaps and overlaps in conversations. J. Phonetics 38, 555–568.

Hjalmarsson, A., 2011. The additive effect of turn-taking cues in human and synthetic voice. Speech Commun. 53 (1), 23–35.

Hjalmarsson, A., Oertel, C., 2012. Gaze direction as a back-channel inviting cue in dialogue. In: Proc. of the IVA 2012 Workshop on Realtime Conversational Virtual Agents (RCVA 2012). Santa Cruz, CA, USA.

Huang, L., Morency, L.-P., Gratch, J., 2011. Virtual Rapport 2.0. In: Intelligent Virtual Agents, Reykjavik, Iceland, pp. 68–79.

Iwase, T., Ward, N., 1998. Pacing spoken directions to suit the listener. In: Proceedings of ICSLP, Sydney, Australia, pp. 1203–1207.

Johansson, M., Skantze, G., Gustafson, J., 2013. Head pose patterns in multiparty human–robot team-building interactions. In: International Conference on Social Robotics – ICSR 2013. Bristol, UK.

Katzenmaier, M., Stiefelhagen, R., Schultz, T., Rogina, I., Waibel, A., 2004. Identifying the addressee in human–human–robot interactions based on head pose and speech. Proceedings of International Conference on Multimodal Interfaces ICMI 2004. State College, PA, USA.

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

Kennington, C., Kousidis, S., Schlangen, D., 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. In: Proceedings of the SIGDIAL 2013 Conference, Metz, France, pp. 173–182.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y., 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. Lang. Speech 41, 295–321.

Lai, C., 2010. What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. In: Proceedings of Interspeech, Makuhari, Japan.

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

Liscombe, J., Venditti, J., Hirschberg, J., 2006. Detecting question-bearing turns in spoken tutorial dialogues. In: Proceedings of Interspeech 2006, Pittsburgh, PA, USA.

Meena, R., Skantze, G., Gustafson, J., 2013. A data-driven model for timing feedback in a map task dialogue system. In: 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue – SIGdial, Metz, France, pp. 375–383.

Morency, L.P., de Kok, I., Gratch, J., 2010. A probabilistic multimodal approach for predicting listener backchannels. Auton. Agent. Multi-Agent Syst. 20 (1), 70–84.

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.

Neiberg, D., Gustafson, J., 2012. Cues to perceived functions of acted and spontaneous feedback expressions. In: The Interdisciplinary Workshop on Feedback Behaviors in Dialog.

Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., Gustafson, J., 2012. Gaze Patterns in Turn-Taking. In: Proc. of Interspeech 2012, Portland, Oregon, US.

Okumura, Y., Kanakogi, Y., Kanda, T., Ishiguro, H., Itakura, S., 2013. Infants understand the referential nature of human gaze but not robot gaze. J. Exp. Child Psychol. 116, 86–95.

Pon-Barry, H., 2008. Prosodic manifestations of confidence and uncertainty in spoken language. In: Proceedings of Interspeech, Brisbane, Australia, pp. 74–77.

Randolph, J.J., 2005. Free-marginal multirater kappa: an alternative to Fleiss' fixed-marginal multirater kappa. In: Joensuu University Learning and Instruction Symposium. Joensuu, Finland.

Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., van Welbergen, H., 2011. Continuous interaction with a virtual human. J. Multimodal User Interfaces 4 (2), 97–118.

Sacks, H., Schegloff, E., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Schegloff, E., 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In: Tannen, D. (Ed.), Analyzing Discourse: Text and Talk. Georgetown University Press, Washington, DC, USA, pp. 71–93.

Schlangen, D., Skantze, G., 2011. A general, abstract model of incremental dialogue processing. Dialogue Discourse 2 (1), 83–111.

Schober, M., Clark, H., 1989. Understanding by addressees and overhearers. Cogn. Psychol. 21 (2), 211–232.

Sjölander, K., Beskow, J., 2000. WaveSurfer – an open source speech tool. In: Yuan, B., Huang, T., Tang, X. (Eds.), Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing, Beijing, pp. 464–467.

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.

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., & Gustafson, J., 2009. Attention and interaction control in a human–human–computer dialogue setting. In: Proceedings of SigDial 2009, London, UK.

Skantze, G., Hjalmarsson, A., 2013. Towards incremental speech generation in conversational systems. Comput. Speech Lang. 27 (1), 243–262.

Skantze, G., Schlangen, D., 2009. Incremental dialogue processing in a micro-domain. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Athens, Greece.

Skantze, G., Hjalmarsson, A., Oertel, C., 2013a. Exploring the effects of gaze and pauses in situated human–robot interaction. In: 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue – SIGDial, Metz, France.

Skantze, G., Oertel, C., Hjalmarsson, A., 2013b. User feedback in human–robot interaction: prosody, gaze and timing. In: Proceedings of Interspeech.

Staudte, M., Crocker, M.W., 2011. Investigating joint attention mechanisms through spoken human–robot interaction. Cognition 120, 268–291.

Stocksmeier, T., Kopp, S., Gibbon, D., 2007. Synthesis of prosodic attitudinal variants in German backchannel ja. In: Proceedings of Interspeech 2007.

Velichkovsky, B.M., 1995. Communicating attention: gaze position transfer in cooperative problem solving. Pragmatics Cognition 3, 199–224.

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.

Wallers, Å., Edlund, J., Skantze, G., 2006. The effects of prosodic features on the interpretation of synthesised backchannels. In: André, E., Dybkjaer, L., Minker, W., Neumann, H., Weber, M. (Eds.), Proceedings of Perception and Interactive Technologies. Springer, pp. 183–187.

Ward, N., 2004. Pragmatic functions of prosodic features in non-lexical utterances. In: Proceedings of Speech Prosody, pp. 325–328.

Ward, N., Tsukahara, W., 2003. A study in responsiveness in spoken dialog. Int. J. Hum Comput Stud. 59, 603–630.

Yngve, V.H., 1970. On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, Chicago, pp. 567–578.