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/

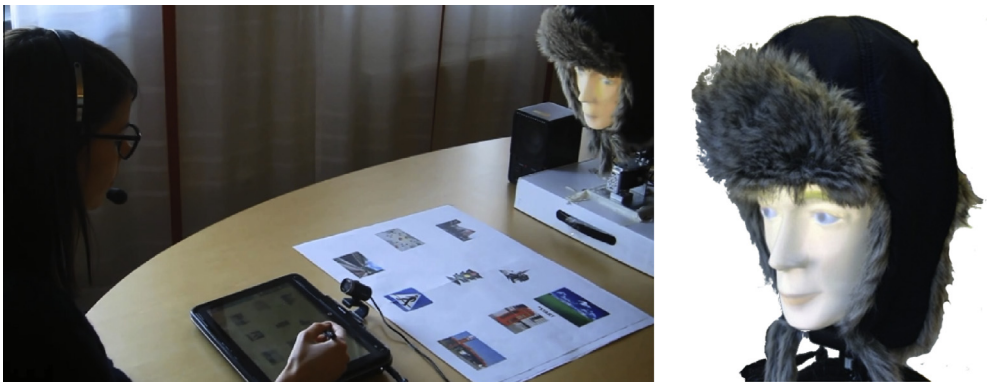

Fig. 1. The human–robot Map Task setup used in the study (left) and a close-up of the robot head Furhat (right).

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).