A Gaze-contingent Dictating Robot to Study Turn-taking

Alessandra Sciutti
RBCS Dept, Istituto Italiano di
Tecnologia, Via Morego 30, 16163
Genoa, Italy
alessandra.sciutti@iit.it

Lars Schillingmann
Graduate School of Engineering,
Osaka University 2-1 Yamadaoka,
Suita, Osaka, Japan
lars@ams.eng.osaka-u.ac.jp

Oskar Palinko
RBCS Dept, Istituto Italiano di
Tecnologia, Via Morego 30, 16163
Genoa, Italy
oskar.palinko@iit.it

Yukie Nagai
Graduate School of Engineering,
Osaka University
2-1 Yamadaoka, Suita, Osaka, Japan
yukie@ams.eng.osaka-u.ac.jp

Giulio Sandini
RBCS Dept,
Istituto Italiano di Tecnologia,
Via Morego 30, 16163 Genoa, Italy
giulio.sandini@iit.it

ABSTRACT

In this paper we describe a human-robot interaction scenario designed to evaluate the role of gaze as implicit signal for turn-taking in a robotic teaching context. In particular we propose a protocol to assess the impact of different timing strategies in a common teaching task (English dictation). The task is designed to compare the effects of a teaching behavior whose timing is dependent on the student's gaze with the more standard fixed timing approach. An initial validation indicates that this scenario could represent a functional tool for investigating the positive and negative impacts that personalized timing might have on different subjects.

Categories and Subject Descriptors

H.1.2 [User/Machine Systems]: Human factors

Keywords

Mutual gaze; Turn taking; Timing; Personalized behavior

1. INTRODUCTION

Establishing eye contact with another person is a fundamental step in initiating communication and in regulating inter-individual exchanges, particularly during conversations [1]. Also in the HRI domain, the appropriate use of robot head and eye movements and the monitoring of human head and gaze behavior has been proven important in mediating conversational turn taking in two-party and multi-party settings [2]–[4]. Although it is generally accepted that a contingent gaze behavior leads to more natural interaction in a conversation (e.g. [3]), what if it is applied in more structured contexts, where the robot should try to play a leading role, motivating the human partner to keep a certain pace? There are several environments in which the rhythm of the task is fixed, in order to maximize work frequency. Consider the quality control phases of a company producing food: often those who visually perform the final quality check need to adapt to the time the items are presented to them on a conveyor belt at a fast pace. In a school-related scenario, usually the task of taking a dictation during a foreign language class is performed by listening to recorded speech, which guarantees again a fixed timing. Hence structured interaction is often guided by a predefined rhythm, which facilitates the coordination of the partners involved and is thought to maximize their efficiency. On the other hand, a pre-established timing forces all participants to adjust their natural speed to the external requirement. Where does the optimal trade-off between these two paradigms lie? This question acquires particular relevance in those contexts of human-robot interaction where the robot has the role of a trainer or teacher, with the need to find the appropriate balance between adaptation to the needs of the human partner and the avoidance of "slacking". To address this question we have developed a simple English-as-a-second language dictation scenario, where a humanoid robot plays the role of the teacher and adopts either a fixed timing in dictating a set of sentences (Rhythmic condition) or a gaze contingent behavior (Contingent condition), pronouncing a new sentence only when it detects that a student is looking at the robot's eyes. This protocol could allow to test if making the robot behavior contingent to human gaze facilitates the interaction and leads to a more efficient turn taking or whether it slows down task completion for participants who tend to be naturally slower. In the following sections we will describe the structure of the system and the results of its validation on trained human partners.

2. THE SCENARIO

The robot used in the current implementation is the humanoid robot iCub [5]. It has been programmed to pronounce a set of predefined sentences while concurrently emulating lip movement, represented as LED lights on the robot's face. Vocalizations and lip motions were obtained through MARY Text-to-Speech System [6] and the iSpeak iCub module. In the "Rhythmic" condition, the robot waits for a fixed time after each sentence (between 11 and 13s as a function of sentence length). This time was chosen considering that the average speed for transcription in case of slow writers is slightly over 20 words per minute [7]. In the "Contingent" condition, the robot does not initiate a new sentence

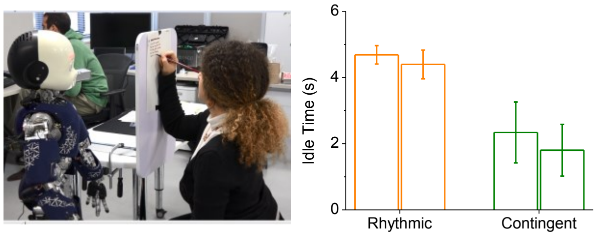

Figure 1: Left - Snapshot of the scenario. Right - Average idle time for the two sessions of each condtion.

HRI'15 Extended Abstracts, Mar 02-05 2015, Portland, OR, USA
ACM 978-1-4503-3318-4/15/03.
http://dx.doi.org/10.1145/2701973.2702055

until the subject gazes at him during the waiting period (which starts 5s after the robot completed the pronunciation of the sentence). To monitor subjects' gaze we implemented a mutual gaze detection module that classifies the eye area in conjunction with an existing face detector [8]. Mutual gaze is detected only if it is maintained for at least 150ms. The experiment is recorded both through the eyes of the robot and through an external camera. The whole task consists of the dictation of four paragraphs, each composed of 8 short sentences (e.g., "The flowers are red.") The order of condition presentations (Rhythmic or Contingent) is randomized between participants to control for order effects. Participants are instructed to listen to each sentence and then write it down, while leaving blank space for any word that they did not understand. The main variables for the current analysis are ***idle time***, i.e., the time between the moment in which the participant stops writing and the beginning of the next utterance by the robot; and ***pause duration*** (fixed in the Rhythmic condition and user-dependent in the Contingent sessions).

3. VALIDATION

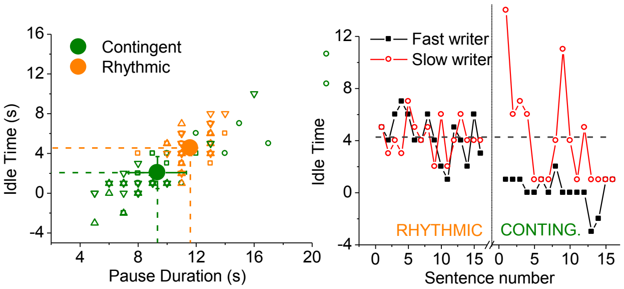

Figure 2: Left - Idle time plotted against pause duration. Single subjects (small symbols) and averages (big circles). Different symbols represent different subjects. Right - Trial by trial variation in idle time for fast and slow writers.

We validated the scenario by testing four of the authors in the task. Testing on non-naïve subjects was aimed at verifying the suitability of the task to detect different subjects' reactions to the two teaching strategies. The results show that on average the idle time in the contingent case is shorter than in the rhythmic condition (see Fig. 1, right) and this difference is statistically significant in 3 of 4 subjects (pair sample t-tests, $p<0.001$). This was foreseeable given the slow timing selected for the Rhythmic condition. However, the Contingent condition was characterized by a higher variability in timing, both among different subjects (compare different green symbols in Fig. 2 left) and within the same subject (see for instance how the green circles in Fig. 2 span the whole graph). In particular, the comparison of two participants exhibiting different average writing speeds (see Fig. 2 right) clearly shows that only the rhythmic approach leads to the adoption of a common pace. Conversely, the contingent approach can have different effects on different subjects. If on the one hand it leads to an average reduction in idle time, especially for those who naturally tend to be fast, it may also lead to erratic and long waiting times for slower subjects. The presence of negative idle times (see for instance the black dots in Fig. 2 right) implies that the robot began to pronounce a sentence before the subject finished his writing. These results indicate the occurrence of false positive errors (mutual gaze detection in absence of real mutual gaze). The analysis of the whole data indicated that this type of error occurred quite rarely (2% of the sentences). In sum, the analysis conducted provides evidence that this scenario could be actually used to characterize different types of turn-taking approaches and to evaluate their impact on human-robot interaction.

4. DISCUSSION & ONGOING WORK

"Taking dictation requires choreography between speaker and listener" [9] therefore poor synchronization has a strong impact on task performance yielding to delays. Here we have shown that the use of a similar task in an HRI scenario could be functional to the investigation of human response to different robotic approaches to turn taking. As the proposed system is now validated, it will be used for data collection on naïve subjects. The analysis of performance metrics (as idle time and pause duration), will be complemented with the analysis of ***subjects gaze patterns*** during the dictation and by a short questionnaire to assess the qualitative ***subjective opinion*** of the different timing conditions. The aim of the complete study will be two-fold. On the one hand we will try to demonstrate the importance for the robot to read an implicit communication signal as the establishment of mutual gaze to regulate the interaction. On the other hand we will assess under which conditions a contingent – or personalized – response could actually lead to a more efficient and/or a more pleasant interaction.

5. ACKNOWLEDGMENTS

The research presented here has been supported by the European CODEFROR project (PIRSES-2013-612555)

6. REFERENCES

[1] N. George and L. Conty, "Facing the gaze of others., *Neurophysiol. Clin.*, vol. 38, no. 3, pp. 197–207, Jun. 2008.

[2] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing In Human-Robot Conversations : How Robots Might Shape Participant Roles Using Gaze Cues," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, vol. 2, no. 1, pp. 61 – 68.

[3] J. G. Trafton, M. D. Bugajska, B. R. Fransen and R. M. Ratwani, "Integrating Vision and Audition within a Cognitive Architecture to Track Conversations," in *ACM/IEEE Iinternational Conference on Huma-Robot Interaction (HRI),* 2008, pp. 201 – 208.

[4] J. Ido, Y. Matsumoto, T. Ogasawara, and R. Nisimura, "Humanoid with interaction ability using vision and speech information," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 1316–1321.

[5] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The iCub humanoid robot: an open-systems platform for research in cognitive development.," *Neural Netw.*, vol. 23, no. 8–9, pp. 1125–34.

[6] M. S. Oder and J. Urgen, "The German Text-to-Speech Synthesis System MARY : A Tool for Research ," *Int. J. Speech Technol.*, vol. 6, pp. 365–377, 2003.

[7] C. M. Brown, "Human-computer interface design guidelines," Jan. 1988.

[8] D. E. King, "Dlib-ml : A Machine Learning Toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[9] K. Johnson and E. Street, "Response to intervention and precision teaching: creating synergy in the classroom," 2011