Evaluating Multimodal Human-Robot Interaction: A Case Study of an Early Humanoid Prototype

Gudberg K. Jonsson
Human Behavior Laboratory
University of Iceland
Skipholt 50, 105 Reykjavik, Iceland
gjonsson@hi.is

Kristinn R. Thorisson
School of Computer Science
University of Reykjavik
Menntavegi 1, 101 Reykjavík, Iceland
thorisson@ru.is

ABSTRACT

Multimodal natural behavior of humans presents a complex yet highly coordinated set of interacting processes. Providing robots with such interactive skills is a challenging and worthy goal and numerous such efforts are currently underway; evaluating the progress in this direction, however, continues to be a challenge. General methods for measuring the performance of artificially intelligent systems would be of great benefit to the research community. In this paper[1] we describe an approach to evaluating human-robot multimodal natural behavior. The approach is based on a detailed scoring and spatio-temporal analysis of the structure and patterning of live behavior, at multiple temporal scales, down to the decisecond level. The approach is tested in a case study involving an early virtual robot prototype, Gandalf, which is capable of real-time verbal and non-verbal interaction with people. Our analysis includes a comparison to a comparable human-human dyadic interaction scenario. Our main objective is to develop a methodology for comparing the quality and effectiveness of human-robot interaction between a wide variety of such systems. Early results indicate that our approach holds significant promise as a future methodology for evaluating complex systems that have a natural counterpart.

Author Keywords
T-patterns, human-robot interaction, multimodal dialogue, real-time, turn taking.

INTRODUCTION

In the fields of robotics and artificial intelligence work dealing with human-agent interaction, such as pet robot development for entertainment, humanoid robot applications [11], interactive teaching systems [4], and so on, has been gradually increasing. Domains where human behavior understanding is crucial (e.g., human-computer interaction, affective computing and social signal processing) require advanced pattern recognition techniques to automatically interpret the complex behavioral patterns of human-machine interaction. This is a challenging problem where many questions are still open, including the joint modeling of behavioral cues taking place at different time scales, the inherent uncertainty of machine-detectable evidence for human behavior, the mutual influence of people involved in interactions on each other, the presence of long-term dependencies and the important role of dynamics in human behavior understanding.

Earlier studies using the pattern detection approached described below have investigated the spontaneous play between the human and the AIBO robot and compared the temporal structure of the interaction with dog and AIBO in both children and adults [6]. The results indicated that both children and adults terminated T-patterns more frequently when playing with AIBO than when playing with the dog puppy, which suggest that the robot has a limited ability to engage in temporally structured behavioral interactions with humans. The authors argue that, as with other human studies, the results indicate that the temporal complexity of the interaction is good measure of the partner's attitude and conclude that more attention should be given to the robots' ability to engage in cooperative interaction with humans.

The RoboCup research community aims to developing artificial agents that will be able to mimic human behavioral patterns during soccer games and to meet this they have argued that there is a need for measuring and comparing the emerging behavior across populations (human vs. artificial) and to develop and standardize a particular pattern detection system that could be used by all research groups and that would further serve as a measure of efficiency of research improvements. For this purpose the t-pattern detection approach, described below, has been suggested with the argument that is has already been successfully applied to the analyses of human-animal and human-robot interactions, and real-life human soccer matches [2]. The

[1] This work was funded in part by the European Union FP7 program through the project HUMANOBS – Humanoids that Learn Socio-Communicative Skills Through Observation, grant number 231453.

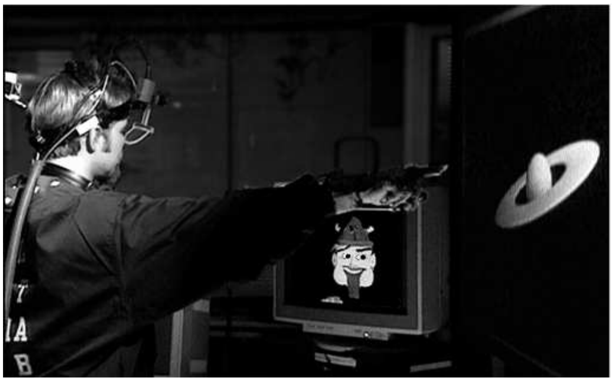

Figure 1. User interacting with Gandalf. Gandalf's behavior proceeds at natural human-human real-time dialogue speeds (notice glance in response to deictic gesture).

objective of the present work is to develop a general methodology for evaluating multimodal human-robot interaction. The approach is based on the analysis of the complex structure/patterns of verbal and non-verbal behavior in multimodal human-humanoid interaction. The interaction is compared to comparable human-human interaction. The study is a part of an ongoing and broader research concerning the development of a prototype for evaluating multimodal human-robot interaction where the goal is to develop new cognitive architectural principles to allow intelligent agents to learn socio-communicative skills by observing and imitating people.

MULTIMODAL HUMANOID SYSTEM

As an initial system to study we have chosen the Gandalf system - an early prototype of an artificial agent capable of broad multimodal real-time interaction with people [12,13]. This system was chosen both because of its breadth of behaviors and number of modes presented both in the input and the output, as well as the number and quality of video recordings available.

Gandalf is a communicative humanoid built in the Ymir framework at M.I.T. between 1992 and 1996 [12,13]. It represents a distributed, modular approach that can be used to create autonomous characters capable of full-duplex multimodal perception and action generation. Gandalf is capable of fluid turn-taking and unscripted, task-oriented dialogue; he perceives natural language, natural prosody and free-form gesture and body language, and conducts natural, multimodal dialogue with a human. Gandalf, an expert in the solar system, can perceive and interpret several thousand utterances related to the topic; two kinds of general manual gesture (deictic and iconic); contextual body, head and gaze direction; and speech prosody. He contextually generates several thousand utterances related the topic, many types of facial expressions, two types of manual gesture (beat and deictic), as well as head and gaze direction. Computer-naïve users' rated Gandalf highly on believability, language ability and interaction smoothness [12]: In less than a minute people communicate naturally and take turns efficiently. In evaluation questionnaires collected from users Gandalf's "interactivity" is rated somewhere between that of a dog and a real human. The original Ymir/Gandalf system ran on 8 networked workstations; more recent incarnations of Ymir have been used in increasingly complex dialogue systems [4] and run on humanoid robots, including Honda ASIMO [11].

CODING AND DATA ANALYSIS

Ten dyadic interactions between human user and Gandalf were transcribed and analyzed using the Theme software [7,8,9]. The results were compared to results obtained from several studies on human-human dyadic interactions involving interaction between doctor-patient; friends, strangers and students [5].

The interactions were transcribed using ThemeCoder. A category system for non-verbal behavior was adopted from McGrew [10], and verbal categories from Bromberg & Landré [1]. The transcribed records were then analyzed using Theme 5.0 [7,8,9]. The basic assumption of this methodological approach is that the temporal structure of a complex behavioral system is largely unknown, but may involve a set of particular type of repeated temporal patterns (T-patterns; [7,8,9]) composed of simpler directly distinguishable event-types, which are coded in terms of their beginning and end points (such as "boy begins deictic gesture" or "girl ends speaking"). The kind of behavior record (as set of time point series or occurrence times series) that results from such coding of behavior within a particular observation period (here called T-data) constitutes the input to the T-pattern definition and detection algorithms.

Within a given observation period, if two actions, A and B, occur repeatedly in that order or concurrently, and found more often than expected by chance, they are said to form a minimal T-pattern (AB), (assuming as h0 independent distributions for A and B, there is approximately the same time distance (called critical interval, CI) between them). Instances of A and B related by that approximate distance then constitute occurrence of the (AB) T-pattern and its occurrence times are added to the original data. More complex T-patterns are consequently gradually detected as patterns of simpler already detected patterns through a hierarchical bottom-up detection procedure (see a simple example in Fig. 2). Pairs (patterns) of pairs may thus be detected, for example, ((AB)(CD)), (A(KN))(RP)), etc. Special algorithms deal with potential combinatorial explosions due to redundant and partial detection of the same patterns using an evolution algorithm (completeness competition), which compares all detected patterns and lets only the most complete patterns survive. As any basic time unit may be used, T-patterns are in principle scale-independent, while only a limited range of basic unit size is relevant in each concrete study. This methodology is explained in detail in Magnusson 1996 paper [7].

The results for the analysis depend on two factors, the source material itself, and how well it is coded, and the

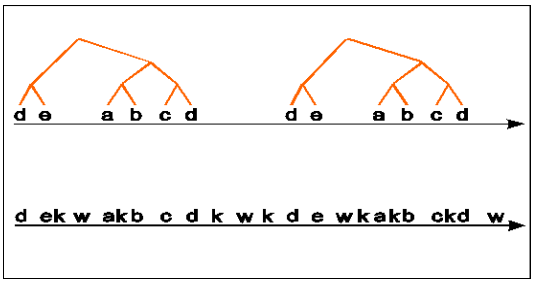

Figure 2. The lower part of this figure shows a simple real-time behavior record containing a few occurrences of several event types, a, b, c, d, & e, indicating their respective instances within the observation period. The upper line is identical to the lower one, except that occurrences of k and w have been removed. A simple t-pattern (abcd) then appears, that was difficult to see when the other events were present.

settings of particular analysis parameters. The coding was done by experienced coders with years of experience in using the Theme software[1]. We have chosen to code parameters that were a special target of the system design. As multimodal coordination is a key feature of the Gandalf system, we have included features from gaze, manual gesture, head movement, and speech. An analysis of the timing of events is of course part of the detailed comparison, as the T-patterns method analyzes spatio-temporal patterns. The following parameter settings were used for all analysis: Minimum number of occurrences set at 3 and significance level set at .0005 (other search values set at default: [7,8,9]).

RESULTS OF ANALYSIS

A high number of temporal patterns were detected in the Gandalf data set involving turn-taking. The number, frequency and complexity of detected patterns indicate that behavior was highly synchronized in all situations. This synchrony was found to exist on different levels, with highly complex time structures that extended over considerable time spans where some of the patterns occurred in a cyclical fashion. The Gandalf data is most similar to human-human scenarios that involve interviews and interrogations. The reason is that Gandalf is designed to be a passive "guide to the solar system" - the dialogue proceeds exclusively via human-initiated dyadic interactions. Many of the patterns seen are the same as those found in the human-human data from interviews/interrogations. Patterns involving coordinated gaze and gesture were found to be similar to human-human dialogue but some were found exclusively in the Gandalf scenarios and others in both human-human and Gandalf-human dyads. For those patterns the content and order of the turn-taking events was the same but differences were found in the interval between even types. Preliminary results indicate that the interval between question and answer was at least 15% less in friendship dyads than in the Gandalf scenarios. The interval between the turn-taking events was found to be more similar between the Gandalf scenarios and interview and stranger dyads.

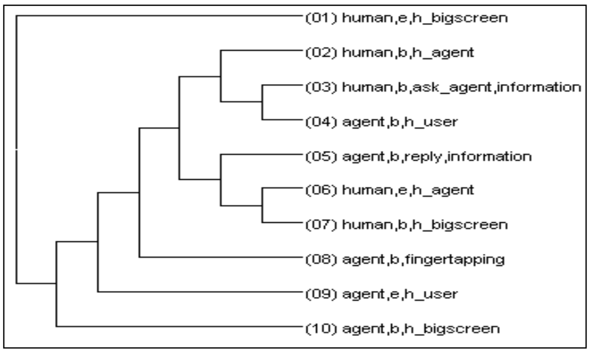

Figure 3. Example of a complex pattern found in the human-agent dyads with 10 events occurring in the same order with significantly similar time interval between each event type occurrence: 1. Human ends looking at big screen. 2. Human begins to look at agent. 3. Human begins to ask agent for information. 4. Agent begins to look at user. 5. Agent begins to reply. 6. Human ends looking at agent. 7. Human begins to look at big screen. 8. Agent begins finger tapping. 9. Agent ends looking at user. 10. Agent begins to look at big screen.

On average the Gandalf scenario patterns seem to be somewhat slower than human-human, which can likely be explained by a longer average duration between critical elements of the turn taking system, especially the speech recognition[2]. A closer comparison to human-human data suggests that the Gandalf data has the highest similarity to dyads involving interaction between (human) friends, higher than doctor-patient, dialogue between strangers. Turn-taking in the Gandalf data reaches a “mean level” of patterning quickly, as is the case in dyadic interaction between friends [5]. In all Gandalf dyads we find examples of a “patterning-growth” period, were the patterns in many cases developed into highly complex structures, comparable to human-human in complexity.In the Gandalf-human dyads the human initiates over 99% of patterns detected and seems to controls the growth and synchronization or “beat” of the interaction. In that sense the human-agent interaction has some similarities with doctor-patient interaction. The duration and patterning of manual gesture in relation to speech and turn-taking, seems to be highly involved in the patterning of the turn-taking system in all situations analyzed. In human-human dyads we find a richer repertoire of gestures – higher number of different gestures used, even though the functionality seems to be the same.

[1] The inter-observer reliability of coding has been assessed in prior work and was not measured here; our prior results indicate that scores was 0.74 for all classes of behavior, but over .85 for “looking behavior” and “verbal behavior”.

[2] Gandalf's speech recognition is the largest bottleneck, taking on average around 1.5 to 2 seconds to process the utterances. Other processes typically take only a fraction of that to produce output, such as e.g. Gandalf's ability to gaze in the direction of pointing, but sometimes a processes producing visible behaviors are serially dependent.

We analyzed the coordination of head, hand and gaze in relation to speech and turn-taking. When looking at simple and short turn-taking patterns results indicate that similar structures are detected across all dyads analyzed. More complex patterns are though detected in the human-human dyads than in the Gandalf scenario, partly explained by the limited repertoire of behavior displayed by Galdalf. When looking at head-turn and gaze, apart from different interval detected between event in the Gandalf scenarios and human-human, we also find that the structure of Gandalf's looking behavior is more similar to humans with moderate and high self-esteem and extraverts [5] than those with low self-esteem, even though the frequency of "looking at partner" behavior is less. Human gaze patterns of people with low self-esteem and/or introverts are of a lower frequency and duration than that of those who have high self-esteem and/or are classified as extroverts [5].

METHODOLOGY: RESULTS

The preliminary investigations described have reinforced the authors' belief that T-pattern identification has good potential as an effective research tool in AI/Robotics. One reason is the fact that that the T-pattern detection correctly identified the turn-taking patterns by the system, and allowed us to compare them to comparable naturally occurring data at a fine level of detail. In addition, the method detects more complex structures than has been possible before and, as in the present study, it detected patterns indicating when the human-agent turn-taking pattern were about to fail. One potential concern is that the results are strongly affected by the choice of parameter settings for the T-pattern analysis, which could make comparisons between researchers more difficult. In the present study we were careful to set these parameters at the same settings for all analysis. In the future it should be investigated to link them to the actual data, to make cross-evaluations also possible.

CONCLUSION

Our approach for studying human-robot interaction shows great promise. It is especially relevant when - as we do here - it is possible to compare the results to real human-human data of comparable circumstances. The results are in line with prior work using T-patterns and the Theme software for analyzing spatio-temporal patterns. More experience must of course be collected on the use of these tools, and they must be applied to a broader range of systems. We believe that the approach presents the way towards the development of a set of methods and indexes that will improve quantitative and qualitative comparison between large dataset of different artificial intelligence systems. The identification of complex and repeated patterns, which are not identifiable through simple observation, has great benefits for the development of an index aimed to advance the continued development of artificial agents and robots.

REFERENCES

1. Bromberg, M. & Landré, A. Analyse de la Structure Interactionnelle et des Stratégies discursive dans un talk-show. Psychologie Francaise, *38,* 2 (1993), 99-109.
2. Hadzibeganovic, T., Van den Noort, M.W.M.L., Cannas, S.A., Bosch, M.P.C., Jonsson, G.K., & Magnusson, M.S. The missing neurocognitive and artificial general intelligence bases of RoboCup research: What still needs to be done before 2050? Proceedings of the 3rd Austrian RoboCup Workshop 2008. Villach, Austria.
3. Johnson, W. L., Vilhjalmsson, H. and Marsella, M. Serious Games for Language Learning: How Much Game, How Much AI? 12th International Conference on Artificial Intelligence in Education, 2005).
4. Jonsdottir, G. R. and Thórisson, K. R. Teaching Computers to Conduct Spoken Interviews: Breaking the Realtime Barrier With Learning. IVA '09, 446–459; LNAI 5773.
5. Jonsson, G.K. Personality and Self-Esteem in Social Interaction. In From Communication to Presence: Cognition, Emotions and Culture towards the Ultimate Communicative Experience. Edited by Giuseppe Riva et al. IOS Press, 2006. ISBN 1-58603-662-9.
6. Kerepesi, A., Kubinyi, E., Jonsson, G.K., Magnusson, M.S., Miklósi, Á. Behavioural comparison of human-animal (dog) and human-robot (AIBO) interactions. *Behavioral Processes*, *73* (2006), 92-99.
7. Magnusson, M.S. "Hidden Real-Time Patterns in Intra- and Inter-Individual Behavior: Description and Detection." *European Journal of Psychological Assessment*, *12*, 2 (1996). 112-123.
8. Magnusson, M.S. Discovering hidden time Patterns in Behavior: T-patterns and their detection. *Behavior Research Methods, Instruments & Computers*, *32* (2000), 93-110.
9. Magnusson, M.S. Structure and Communication in Interaction. In G. Riva, M.T. Anguera, B.K. Wiederhold, F. Mantovani (eds.) 2006. From Communication to Presence: Cognition, Emotions and Culture Towards the Ultimate Communicative Experience. Útgefandi: IOS Press October 1, 2006.
10. McGrew, W.C. An Ethological Study of Children's Behaviour. New York: Lawrence Erlbaum Associates, 1972.
11. Ng-Thow-Hing, V., K. R. Thórisson, R. K. Sarvadevabhatla, J. Wormer, T. List. The Cognitive Map architecture for facilitating human-robot interaction in humanoid robots. *IEEE Robotics & Automation*, *16*, 1 (2009), 55-66.
12. Thorisson, K. R. Communicative Humanoids: A Computational Model of Psycho-Social Dialogue Skills. Unpublished Ph.D. Thesis, Media Laboratory, Massachusetts Institute of Technology, 1996.
13. Thorisson, K. R. A Mind Model for Multimodal Communicative Creatures and Humanoids. International *Journal of Applied Artificial Intelligence*, *13*, 4-5 (1999), 449-486.