# Robot's Delight - A Lyrical Exposition on Learning by Imitation from Human-human Interaction

Dylan F. Glas,
Malcolm Doering, Phoebe Liu
Hiroshi Ishiguro Laboratories
ATR, Kyoto, Japan
+81 774 95 1405
{dylan, malcolm.doering,
phoebe}@atr.jp

Takayuki Kanda
Intelligent Robotics and
Communication Laboratories
ATR, Kyoto, Japan
kanda@atr.jp

Hiroshi Ishiguro
Intelligent Robotics Laboratory
Osaka University
Toyonaka, Osaka, Japan
ishiguro@sys.es.osaka-u.ac.jp

## Keywords

Learning by imitation; learning from demonstration; social human-robot interaction; multimodal interaction; dialog content; data-driven HRI; autonomous social interaction

## 1. Introduction

Now that social robots are beginning to appear in the real world, the question of how to program social behavior is becoming more pertinent than ever. Yet, manual design of interaction scripts and rules can be time-consuming and strongly dependent on the aptitude of a human designer in anticipating the social situations a robot will face.

To overcome these challenges, we have proposed the approach of learning interaction logic directly from data captured from natural human-human interactions. In comparison with teleoperation or web-based crowdsourcing, our approach has the benefit of capturing the naturalness and immersion of real interactions, but it faces the added challenges of dealing with sensor noise and an unconstrained action space.

In the form of a musical tribute to The Sugarhill Gang's 1979 hit "Rapper's Delight", this video presents a summary of our technique for capturing and reproducing multimodal interactive social behaviors, originally presented in [1], and preliminary progress from a new study in which we apply this technique to an android for interactive spoken dialogue.

## 2. Learning Multimodal Interaction

In the first study [1], we had participants role-play in a camera shop scenario. Using a position tracking system and speech recognition from smartphones, we captured data from 178 interactions to train the robot to act as the shopkeeper. To reduce the dimensionality of the problem, we introduced several abstractions: we used unsupervised clustering of the position data to define discrete locations and trajectories in the space, we clustered utterances by lexical similarity to identify typical utterance actions, and we applied proxemics models from HRI research to characterize relative positioning.

We then represented the shopkeeper's actions as a discrete combination of utterance, location, and proxemic formation which could be reproduced by a robot. For each observed customer action (motion and/or speech) we trained a classifier to predict one of these actions, using a vectorization of the joint state of both participants as an input. In the online system, whenever a customer action was identified, the observed joint robot-human state was fed into the predictor to select a robot action to execute. Experiments with users showed the system to be capable of responding to the questions, requests, and movements of participants role-playing customers.

## 3. Modeling Interaction Structure

The next study investigated ways to apply a similar approach using ERICA, a highly-humanlike android robot, in a travel agent scenario. While we first expected the learning problem to be simpler without the necessity of modeling locomotion and proxemics, we found just the opposite, as physical location in the camera shop had helped to indicate the topic of the interaction. Ambiguous questions like "how much does it cost?" can be answered easily if the robot can sense which camera the customer is standing at, but not in the absence of spatial cues.

To address this problem, some model of topic or interaction context is necessary. We chose to avoid language processing approaches, instead focusing on patterns of action occurrence in general. We found that topic-specific utterances frequently occur in "runs" in interactions, and we developed a method to estimate topic based on temporal proximity of utterance actions.

By enumerating sub-sequences of the captured interaction data and computing co-occurrence metrics such as lift and support (metrics from association rule analysis), we were able to form clusters of actions which show a strong correspondence to topic. Including these topic clusters in action prediction improved accuracy in predicting the answers to ambiguous questions.

This work is still in progress, and we are currently preparing to run experiments with real users to evaluate its effectiveness in live human-human interactions.

## 4. Acknowledgments

This work was supported by the JST ERATO Ishiguro Symbiotic Human Robot Interaction Project. Thanks to Jason Lim for his help in designing the robot dance sequences.

## 5. References

[1] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Data-Driven HRI: Learning Social Behaviors by Example from Human-Human Interaction," *IEEE Transactions on Robotics,* vol. 32, pp. 988-1008, 2016.

*HRI '17 Companion*, March 06-09, 2017, Vienna, Austria
ACM 978-1-4503-4885-0/17/03.
http://dx.doi.org/10.1145/3029798.3036646