Generation of robot gestures driven by speech

Robotics MSc project proposal

Examiner: Hedvig Kjellström, hedvig@kth.se

Conversational agents in the form of virtual agents or social robots are rapidly becoming wide-spread. Humans use non-verbal behaviors to signal their intent, emotions and attitudes in human-human interactions. Conversational agents therefore need this ability as well. In this project we focus on physical robots, and one type of non-verbal communicative behavior, gestures.

Previous systems for gesture production were typically rule-based and could not represent the range of human gestures. Recently the gesture generation field has shifted to data-driven approaches, with large success for virtual agents. However, modern systems are most often trained with datasets containing human speech and gesture recordings. This poses a challenge when applying them to physical robots, as both the speech and embodiment of a physical robot is less human-like than that of a virtual agent.

This project will address this issue in 2 steps:

1. Create a model that can generate human gestures from synthetic speech. This will be done by retraining a state-of-the-art gesture generation model [1] on synthetic speech instead of human speech, by realigning synthetic speech with human speech using Dynamic Time Warping for every recording in the dataset and retraining the model on the resulting data [2].
2. Add a step that generates robot gestures instead of the output human gestures. You will work with either Pepper or NAO, and the approach will be to retarget the motion from the human embodiment to that of a humanoid robot, e.g. using the approach developed by Vijayan et al [3].

A successful project will result in a system that can enable a humanoid robot to express itself with both speech and accompanying gestures.

The project requires very good programming skills. If performed in a successful manner, the results of the project will be publishable in an international peer-reviewed conference.

References

[1] T. Kucherenko, D. Hasegawa, G. Eje Henter, N. Kaneko, and H. Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *International Conference on Intelligent Virtual Agents*, 2019.

[2] N. Sadoughi, L. Yang, and C. Busso. Meaningful head movements driven by emotional synthetic speech. *Speech Communication* 95:87-99, 2017.

[3] A. E. Vijayan, S. Alexanderson, J. Beskow, and I. Leite. Using Constrained Optimization for Real-Time Synchronization of Verbal and Nonverbal Robot Behavior. In *IEEE International Conference on Robotics and Automation*, 2018.