Towards Real-time Coordination in Spoken Human-Robot Interaction

Gabriel Skantze

KTH Royal Institute of Technology
skantze@kth.se

When humans interact and collaborate with each other, they have to coordinate their behaviours. One of the most fundamental behaviours that needs to be coordinated is the order in which they speak. Since it is difficult to speak and listen at the same time, they need to take turns speaking, and this turn-taking has to be coordinated somehow. To achieve fluent spoken interaction between humans and machines (such as social robots), it is essential that we understand how this coordination is accomplished. Studies on human-human interaction have shown that humans use multi-modal signals, expressed in the face and voice, such as gaze and intonation. Thus, to engage in spoken interaction, social robots should be able to continuously generate and understand these signals. Since social robots are embodied and physically situated, they have a richer repertoire of multi-modal signals, than for example voice assistants in smart speakers. This facilitates more sophisticated coordination, such as multi-party interaction with several users. In multi-party interaction, the coordination of turn-taking becomes more complicated, since the interlocutors not only have to understand when someone yields the floor, but also who is expected to speak next. In such settings, the gaze of the robot and the users becomes an even more important coordination signal.

In this presentation, I give an overview of several studies that we have done to model turn-taking in dialogue. First, I will show how humans in interaction with a human-like robot make use of the same coordination signals typically found in studies on human-human interaction, and that it is possible to use multi-modal sensors and machine learning to automatically detect and combine these cues to facilitate real-time coordination. Second, I will show how a human-like robot face and voice can be used to display turn-taking signals – such as gaze aversion, breathing, facial gestures and hesitation sounds – and that humans react naturally to such signals, without being given any special instructions. By displaying such cues, the robot can for example claim the floor without being interrupted, and it can influence who will be the next speaker. In a multi-party interaction, it means that the robot may regulate the turn-taking to increase the speaking time of non-dominant speakers. Finally, I will present recent work on how Recurrent Neural Networks can be used to train a predictive, continuous model of turn-taking from human-human interaction data. I will show how such a general model can be applied to a number of different tasks, including pause, backchannel and overlap detection, and I will discuss how it could potentially be used to control the verbal and non-verbal signals displayed by the robot.

References

Skantze, G. (2017). Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of SigDial*. Saarbrucken, Germany.

Skantze, G. (2017). Predicting and Regulating Participation Equality in Human-robot Conversations: Effects of Age and Gender. In *Conference on Human-Robot Interaction (HRI2017)*. Vienna, Austria.

Skantze, G. (2016). Real-time Coordination in Human-robot Interaction using Face and Voice. *AI Magazine*, 37(4), 19-31.