Use of speech and motion cues for bootstrapping complex action learning in iCub

Emre Ugur *†, Jimmy Baraglia*, Lars Schillingmann*, and Yukie Nagai*
*Graduate School of Engineering, Osaka University, Osaka, Japan
†Dept. of Computer Science, Innsbruck University, Innsbruck, Austria

I. Introduction

Parental scaffolding is an important mechanism that speeds up infant sensorimotor development. Infants pay stronger attention to the features of the objects highlighted by parents, and their skills develop earlier than they would in isolation due to caregivers support. Parents are known to make modifications in infant-directed actions, called "motionese", which is characterized by a wider range of motion, repetitive actions, and longer and more pauses between movements. Inspired from motionese, we previously realized a robotic system [1] where the affordances and effect prediction capabilities that are learned in the previous stages of development are used to bootstrap complex imitation and action learning with the help a cooperative tutor through motionese. With this system, a robot could learn new skills via imitation learning by extracting the important steps from the observed movement trajectory, and then encoding them as subgoals that it can fulfill. Considering the affordances provided by the objects in the environment, it found and sequentially executed the actions that are predicted to generate the desired effects and achieve the subgoals; achieving the overall goal of complete imitation. We showed that motionese can be used to bridge the gap between the interacting agents with different movement capabilities, such as the human tutors and the arm-hand robot we employed. Furthermore, our experimental data indicated that naïve tutors who are not informed about the imitation mechanisms of the robot, changed their teaching strategy, and started to display motionese.

Motionese displayed by the real caregivers, on the other hand, is accompanied by rich set of scaffolding signals including social signals such as gaze and speech. In our previous robotic experiments, we only focused on detecting pauses in the motion and ignored any social signal. In this limited setting, we observed that adaptation of the naïve tutor to the imitation mechanisms of the robot was slow, and required many failed teaching attempts. In order to overcome this limitation, we propose to add social signals for scaffolding into our framework, and study the effect of robot-directed social signals in a real developmental robot system. In particular, we decided to use robot-directed speech in this study because of the ecological validity of the exaggerated intonational contours, long pauses, and shorter utterances in infant-caregiver interactions [2] in both biological and artificial systems. The acoustic information, typically in the form of narration, overlaps with action sequences, and can provide agents with a bottom-up guide to attend to the relevant/important parts of actions, and to find structure within them.

II. Contribution

The goal of this research direction is two-fold. First, we aim to extend our previous system that uses motionese, by adding perception of acoustic cues in order to obtain a more natural robot-tutor interaction system. Due to the richer set of scaffolding, the demonstrations can be more effectively segmented into subgoals that are in turn achieved by own repertoire of actions of the robot. Second, we plan to study the effect of combining speech and motion related cues in a human-robot interaction setting. We expect that processing these cues will not only make the robot a better imitator, but also leads to a reciprocal activation effect on use of both cues in teaching performance of the caregiver. In other words, we expect that the naïve tutors associate and utilize the motion and speech related scaffolding cues automatically to underline the subgoals and demarcate the boundaries of the action primitives.

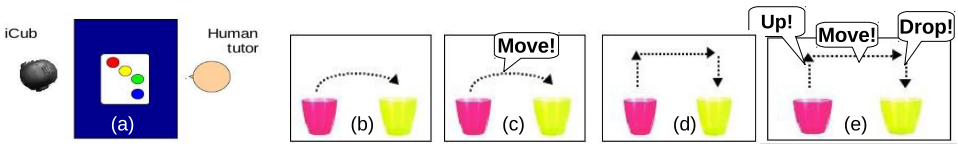

Fig. 1. Demonstration of simplified cup-insertion task with different motion and/or speech cues.

We use a simplified version of cup-insertion scenario as a case study in this abstract. Here, the robot is assumed that it already learned the action primitives such as grasp, lift, carry, drop and push; and the corresponding affordances such as graspability, rollability, pushability. Thus, the robot can make predictions for actions that involve single objects only; but it cannot predict the interaction dynamics between objects. In Fig. 1, cup-inserting is demonstrated without any scaffolding (b), with speech cues only (c), with motionese only (d), and finally with speech cues and motionese (e). If no motionese is displayed, the robot

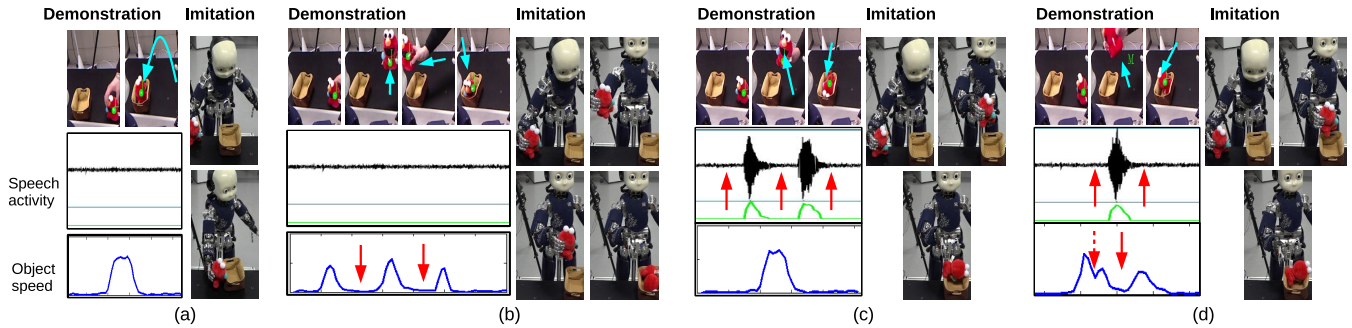

Fig. 2. Demonstration of stacking task and robot's imitation performance. Red arrows show the detected subgoals. Please see the robot video at http://emreugur.net/ICDL-EPIROB-2015/

can only extract the initial and final states of the demonstrated action, encode this task with the goal of moving the red object to right - and as the red object affords pushability - it may attempt to bring the red object to the goal position by simply pushing it to the right, where the red object will push the yellow object away rather than being inserted in it. However, when important steps are highlighted by for example acoustic cues and pauses as in (e), the robot can extract subgoals represented in its perceptual space and find a behavior sequence from its behavior repertoire to imitate the action correctly.

III. Experiments and Preliminary Results

Our experimental setup is composed of an iCub robot for manipulation, a Kinect sensor for visual perception, a microphone for acoustic perception, and a number of objects that are placed on the table as shown in Fig. 2. The robot's workspace consists of several objects and a table where the robot can track objects in real-time using OpenCV library and also can extract object position using the depth image of the Kinect. The robot interacts with the objects using three actions, namely grasp, carry, and release that are assumed to be learned before [1]. Grasp action brings the robot hand to the center of the object, encloses it, and moves the hand $5cm$ upwards in order to avoid any collision with the table in the subsequent actions. Carry action moves the hand to an arbitrary position within the robot's workspace, and release action opens the hand after movement.

Subgoals in the demonstration are detected by segmenting either object motion trajectory or speech activity. Object motion segmentation is performed by observing the speed of the object and creating segments if a pause is detected. Speech activity segmentation is achieved using the ESMERALDA [3], which segments utterances based on a lower and an upper signal energy threshold. The two thresholds define a hysteresis which helps to segment unvoiced utterance beginnings and endings correctly.

In order to verify our system, we performed a number of experiments where an expert tutor demonstrated the cup stacking task, the robot attempted to imitate these demonstrations by sequentially emulating the subgoals it detected. Fig. 2 gives snapshots from four different trials. In the first trial (a), the tutor demonstrated the stacking task without any pause or speech, thus the robot could not detect any subgoal. It tried to bring the object directly to the observed final position without lifting it. The movement was obstructed by the wall of the container and the imitation attempt failed. In (b), the tutor inserted pauses in a rectangular shaped trajectory without producing any speech. As shown, the robot was able to correctly segment the trajectory, find the subgoals based on the pauses detected in the motion of the object, and sequentially emulate the detected subgoals by executing actions from its own repertoire. In (c), the tutor did not introduce any pause while moving the object, but pronounced the verbs "up" and "down" during his demonstration. Speech segmentation was able to segment the voice activity accordingly and find the subgoals; which were sequentially emulated by the robot as shown. Finally in (d), the tutor picked up the object, shook it, and put it inside the container. While shaking, he pronounced "here!", which was detected by the voice segmentation system. Based on the subgoals detected by the voice segmentation or motion segmentation, the robot was able to imitate the observed action. Note that, depending on the thresholds (both in motion and speech processing), different number of segments can be generated.

IV. Conclusion and Future Work

We realized an imitation system that finds the subgoals in the demonstrations based on motion and speech cues, and sequentially emulates these subgoals using robot's own repertoire of actions. In future, in order to reveal the teaching strategies of the humans, and to obtain the most effective human-robot teaching setup, we plan to systematically analyze the role and effectiveness of different cues and parameters in our parental scaffolding setting with experiments that involve naïve tutors.

References

[1] E. Ugur, Y. Nagai, and E. Oztop, "Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills," *Robotica*, 2014.
[2] R. J. Brand and S. Tapscott, "Acoustic packaging of action sequences by infants," *Infancy*, vol. 11, no. 3, pp. 321–332, 2007.
[3] G. A. Finkco, "Developing hmm-based recognizers with esmeralda," in *Text, Speech and Dialogue*. Springer, 1999, pp. 229–234.