FLUIDITY: Defining, measuring and improving fluidity in human-robot dialogue in virtual and real-world settings

Julian Hough[1], Carlos Baptista de Lima[1], Frank Förster[2], Patrick Holthaus[2], Yongjun Zheng[2]

[1]School of Mathematics and Computer Science, Swansea University
[2]School of Physics, Engineering and Computer Science, University of Hertfordshire
Correspondence: julian.hough@swansea.ac.uk

Abstract

This paper summarizes the motivation, aims and objectives of the EPSRC-funded project FLUIDITY in simulated human-robot interaction with speech interfaces. Questions of defining the properties of fluid interaction and the communicative grounding mechanisms needed to achieve them are at the heart of the project.

1 Introduction

A key problem for current human-robot interaction (HRI) with speech interfaces is lack of fluidity. Although there have been significant recent advances in robot vision, motion, manipulation and automatic speech recognition, state-of-the-art HRI is slow, laboured and fragile. The contrast with the speed, fluency and error tolerance of human-human interaction is substantial. The FLUIDITY project[1] takes on this key challenge by developing the technology to monitor, control and increase the interaction fluidity of robots, such that they become more natural and efficient to interact with.

2 The challenge of fluidity for human-robot interaction with speech

In pick-and-place situations where a human responds to a spoken instruction like "put the remote control on the table" and a follow-up repair like "no, the left-hand table" when the speaker realizes the instructee has made a mistake, there is typically nearly no delay in reacting to the initial instruction, and adaptation to the correction is instant. FLUIDITY will give robots with speech understanding more seamless, human-like transitions from processing speech to taking physical action with *no delay*, permitting *appropriate overlap* between the two, and the ability to *repair actions in real time* as humans do (Hough et al., 2015a).

In human-human interaction, fluidity is achieved through humans being able to recognize the intentions of their conversational partner with low latency and using predictions (Tanenhaus and Brown-Schmidt, 2008; McKinstry et al., 2008), and in responding to speech, humans can begin moving in response to an instruction *before* the end of the instructor's utterance (Hough et al., 2015a). Current interactive robots do not exhibit these capabilities partly due to unsuitable control algorithms which demote fluid interaction quality over other concerns. FLUIDITY puts interaction fluidity and the rapid recovery from misunderstanding with appropriate repair mechanisms at the heart of interactive robots, aiming to develop state-of-the-art incremental spoken language understanding (SLU) and continuous multi-modal HRI control algorithms.

In an example pick-and-place scenario where a user communicates with a robot to move objects to different target locations using their voice, adapting from Hough and Schlangen (2016), the capability of current systems is shown in the interval diagrams in Fig. 1 in the 'non-incremental' mode (A). The interval blocks represent the user's speech and robot's actions over time from left to right.

In 'immediately successful' interactions (Fig. 1 top), the robot processes an instruction like "put the red phone on the table" and understands the user's intention correctly the first time, picking up the user's intended object. Due to the uncertainty caused by the robot's sensors (Kruijff, 2012), the robot needs confirmation from the user through utterances like "yes" or "go ahead" before completing the action to achieve its goal - in mode (A) this is safe, but cumbersome. In 'recovery from miscommunication' scenarios (Fig. 1 bottom) where the incorrect object is initially picked up and the user *repairs* the robot's actions with utterances such as "No! The other red phone." In mode (A), such an utterance cannot be recognized as a repair until the robot has stopped moving. Once the repair

[1]FLUIDITY in simulated human-robot interaction with speech interfaces. UKRI EPSRC grant: EP/X009343/1 project website: https://fluidity-project.github.io/.

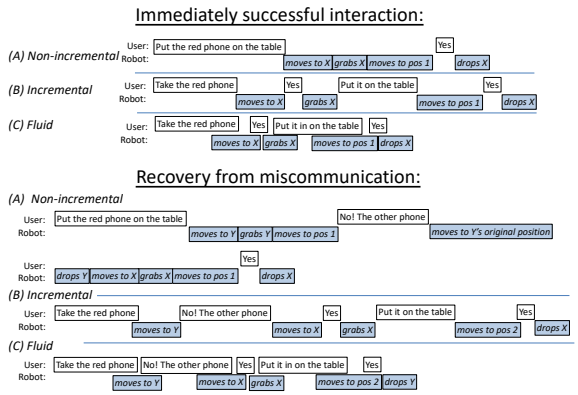

Figure 1: Fluidity in interaction from a non-incremental approach to speech processing (A) up to fluid processing. Modes (B) and (C) have incremental processing and in the fluid setting (C), robot actions start earlier as user feedback utterances can start earlier, as the robot constantly monitors and interprets relative to its actions.

is interpreted, not only must the current incorrect action be 'undone' but the new action must then be carried out in full, resulting in long periods of waiting. The ability to recognize intentions only from complete commands mapping to complete goals severely limits the fluidity of the interaction.

Improvement is possible in mode (B), an *incremental* mode, taking inspiration from (Kempson et al., 2001; Schlangen and Skantze, 2011; Purver et al., 2011; Eshghi et al., 2015; Hough et al., 2015b; Kennington and Schlangen, 2015; Madureira and Schlangen, 2020) and others in computational semantics focused on incrementality. Here, while turn-taking still happens in a half-duplex fashion with no overlap between human speech and robot motion, opportunities for confirmation or repair arise after shorter bursts of speech. This is possible by the robot predicting parts (increments) of the user's overall goal as speech arrives into the system word-by-word, such as predicting the target object to be picked up before predicting the target location. The 'recovery from miscommunication' scenarios show the benefit of incremental processing in situations of repair, as partially incorrect action plans can be corrected early and substantially reduce task completion time.

In the fluid mode (C), speech processing is also incremental, however the system goes *beyond incrementality*, allowing *full-duplex interaction* where concurrency of human speech and robot motion is reasoned with appropriately using *continuous intention prediction*. The robot can begin moving as soon as it is sufficiently confident about the user's goal and it can interpret confirmations and repairs during its movement appropriately, allowing it to complete correct actions more quickly and change course immediately in the middle of its initially selected action if corrected, leading to faster task completions in both scenarios. We also predict the more fluid the interaction, the more this behaviour will be perceived as natural, intelligent and likeable, building from Hough and Schlangen (2016).

3 Aims and objectives

The FLUIDITY project will investigate the *automatic measurement and improvement of fluidity in HRI*. With respect to Fig. 1, the aim is to move away from interaction as it happens in current systems in the non-incremental mode (A) to modes (B), incrementally, and finally, (C), fluidly.

The project will also address the difficulty of developing interactive models with real-world robots. A key outcome, currently under development, is a toolkit for building and testing interactive robot models with human participants in a Virtual Reality (VR) HRI environment, concretely, the simulation of the University of Hertfordshire Robot House[2] with the *Fetch Mobile Manipulator*[3]. The environment will be used to collect Wizard-of-Oz data with participants as the basis for training our SLU and interaction management/control models and of interest to both dialogue and HRI researchers. To achieve fluid interaction, the project will use the data to give a robot with speech understanding capabilities the following abilities:

1. predict the user's intention from their speech and confidence in that prediction as quickly and accurately as possible when sufficiently confident, investigating DS-TTR (Purver et al., 2011; Eshghi et al., 2015) and incrementalized deep learning models (Madureira and Schlangen, 2020) for the SLU.
2. monitor its own motion and estimate the earliest point that its own intention has become recognized by, or 'legible' to the user in the sense of (Dragan et al., 2013), whilst moving.
3. use abilities 1 and 2 in parallel to control its interactive behaviour appropriately, including repairing goals, to allow fluid interaction in both the virtual and real-world settings.

[2] https://robothouse.herts.ac.uk/

[3] https://www.zebra.com/us/en/products/autonomous-mobile-robots.html

References

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, London, UK. ACL.

Julian Hough, Iwan de Kok, David Schlangen, and Stefan Kopp. 2015a. Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL)*, pages 86–94.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015b. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings IWCS 2015*, London, UK. ACL.

Julian Hough and David Schlangen. 2016. Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles. ACL.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell, Oxford.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the ACL. ACL.

Geert-Jan M Kruijff. 2012. There is no common ground in human-robot interaction. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–374.

Chris McKinstry, Rick Dale, and Michael J Spivey. 2008. Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1):22–24.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th IWCS*, Oxford, UK.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1).

Michael K Tanenhaus and Sarah Brown-Schmidt. 2008. Language processing in the natural world. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):1105–1122.