CoCoMaps Project

CMLabs | IIIM

4-Way Turn-Taking (4WTT)

REPORT

DELIVERABLE: T12.D1
DATE: 10 March 2018

INTRODUCTION

This report describes our work on four-way turn-taking in the CoCoMaps project.

The CoCoMaps Project is a joint effort by CMLabs (UK) and the Icelandic Institute for Intelligent Machines (Iceland) to integrate an architecture for robot control and interaction with humans using natural communication.

The goal of 4-way turn-taking (4WTT) is to allow two people and two robots to interact in collaboration to achieve a common task through speech and action in situated dialogue. Achieving 4WTT involves a large portion of architectural features already demonstrated in demos 0, 1, 2 and 3, and are discussed here in light of detailing the principal architectural mechanisms for achieving 4WTT, as well as certain additions that are also presented in this report such as the direction a person's face is turned, and who is speaking.

This report pertains to Task 12 in the CoCoMaps project and is organised as follows: First we discuss the *requirements* for natural four-way turn-taking and dialogue, honing in on the key aspects that CoCoMaps has focused on. Then we describe the *Ymir Turn-Taking Module* (YTTM) and related modules that its processing depends on, followed by the relevant properties of the Task-Dialog Module (TDM) which handles high-level coordination in the CoCoMaps architecture.

Natural Multi-Party Dialogue

Our objective is to replicate certain aspects of human natural interaction for human-robot dialogue. While this has been a moving target in AI and robotics research for the past 30 years, it is still pursued by many companies and research centres. The present system approaches dialogue from the angle of available cues given modern available processing techniques, including prosody, human presence, who is speaking, and where participants are turned.

As 4WTT relies on the entire CoCoMaps system and is represents only an initial step in the direction of naturalness all milestones, KPIs and prior reports provide relevant measures for its progress.

COCOMAPS REQUIREMENTS for 4-WAY TURN-TAKING

The approach taken to the 4-way turn-taking system consists of an initial requirements analysis of human dialogue in light of potential for real-time information extraction of the kind that could realistically be implemented on the robots. In Demo-2 two robots and one human must interact to accomplish a task. The requirements analysis, along with somewhat detailed specifications for demos 2 and 3, laid the foundation for the ultimate focus of our 4-way turn-taking (4WTT) approach.

At the heart of any dialogue system are methods for taking turns – in human dialogue participants take turns providing semantic content, which primarily is conveyed via speech (words, sentences, and various communicative sounds). Humans do a vast amount of information processing on low-, mid- and high-level data to produce smooth turn-taking, content interpretation, interruption handling, and the various other features of human behaviour that characterise human communication. Replicating a useful subset of these has been, and still is, a multi-decade undertaking for the whole human-robot community. This project involves bringing some key features of human dialogue and interaction to commercially available robot platform running the Robot Operation System (ROS) – with human dialogue and available signal processing serving as a major source for prioritising how to simplify when it comes to its modelling.

A key difference between two-party interaction and 4-party interaction is that more than one person may attempt to speak at the same time, resulting in stricter requirements of tracking who has the floor (turn). Another is that utterances of any one participant may be addressed to the whole group, a subset of the group, or a single individual.

In multi-party interaction involving the execution of a partially-specified task with negotiable sub-tasks, participants must agree on who will do which part of the full task, and in what order. Humans rely on an extensive experience with such scenarios, which may extend only as far as cultural and individual differences allow for any scenario. In our approach the robots are provided with scripts that allow them to perform atomic (sub-)tasks, which can then be strung together into more complex tasks, by prescription, dynamically during interaction, or a combination of both.

In real-time interaction humans collect information for deciding what to do next by effectively and efficiently keeping track of who has the "floor" (turn) through signals at several levels of information abstraction, from the "low end" close to the raw signals ("I hear a voice - someone is speaking") to information that is highly abstracted and relies on large amounts of knowledge, such as e.g. social convention ("the current speaker is an important person that I choose not to interrupt"). Needless to say, even the most advanced state of the art human-robot communication uses signals that lie close to the low end of this spectrum.

Humans use numerous signals for inferring the above information, and do so highly dynamically and opportunistically, depending on which ones are present. In human co-operative dialogue "errors" (e.g. speech overlap, long pauses, misunderstandings) are minimised through various effective methods (e.g. leaning closer to a speaker in a noisy environment) and fixing them when they occur (e.g. "Oh, I thought you meant..."). While it is probably a research program for the next few decades to capture all of that in robots, one

defining feature of human dialogue planning is the *opportunistic use* of available information to infer dialogue state, repair methods, and decisions about asking for additional information, clarifications, etc. Opportunism in signal interpretation is in fact a necessary capacity of any natural conversant as some cues may be present at certain times and not others due to the large degree of variability in human behaviour. Our publish-subscribe approach helps address that issue and was chosen in part for this reason.

An additional challenge is that everything that is difficult in 2-party conversation, e.g. speech recognition, meaning extraction, deciding what to say and/or do next, knowing whether a participant has left the discussion, etc., is exacerbated in the multi-party interaction case, due to noise, a larger source of signal generators, creating what essentially amounts to a more complex information-tracking and control problem. One task of any multi-party modelling effort must therefore be to know which corners to cut, and cut them in a way that doesn't preclude adding those corners back later.

What might then be some key features to try to capture in a multi-party human-robot interaction? A key and fundamental limiting assumption in CoCoMaps is that of cooperative interaction, that is, the interaction consists of sequential execution of tasks, some of which are specified by humans, some specified by robots, and some which are negotiated via dialogue and/or other kinds of communication, all executed in cooperative non-adversarial) manner. This reduces the complexity and scope of what we try to model in CoCoMaps, while retaining a perfectly valid subset of human interaction valid for human-robot interaction.

We further address the challenge of teaching robots human interaction skills, in a manner that meets and matches other requirements of the system, by focusing on six key aspects fundamental to the problem: (1) scarcity and availability of data necessary (but not necessarily sufficient) for successful interaction, (2) whom a person is addressing, (3) who has the floor, (4) presence and absence of participants, (5) how the robots coordinate between themselves, and (6) coherent coordination of content, interaction, and task execution, in the following manner:

1. *designing* and *implementing* a distributed system-wide dynamic publish-subscribe system that can make any and all data available to any and all processes at any time.
2. *extracting* and *integrating* the direction which a person's head is turned into the turn-taking mechanisms, based on real-time camera signal processing.
3. *extracting* and *integrating* who is speaking through a first approximation, based on real-time camera signal processing.
4. *extracting* and *integrating* who is present in the scene, based on real-time camera signal processing.
5. *designing* and *implementing* task and role negotiation mechanisms.
6. *designing a coordinating module* called Task-Dialog Manager (TDM) that integrates dialogue and task coordination, while separating interpretation proper via *a separate meaning-extractor* (MEx).

APPROACH & METHODOLOGY

This section outlines our approach to the above 6 items, and in addition describes briefly a key sub-system (called the YTTM) for controlling low-level turn-taking events and information.

System-Wide Distributed Pub-Sub

The CoCoMaps architecture addresses the modelling of a multi-party turn-taking control system by separating information extraction and processing into several levels of abstraction, in a modular way, where low-level signals (such as e.g. whether a participant is present, a speech signal is detected, whom a human speaker is addressing) are explicitly represented and marked as being at a lower level than e.g. the dialogue content generation and task control.

A system-wide publish-subscribe system can make any of such information entering and processed in the system available to any module at any point at runtime. While this makes runtime monitoring somewhat more difficult, the flexibility for programming and runtime modifications that this approach brings are significant, meeting the requirements of opportunistic processing required for integrated task-oriented dialogue. To make the approach work we follow a global policy for determining (at design time) which level of abstraction any particular piece of information belongs, and maintain a strict policy on naming intermediate partially-processed information. Together this results in a coherent tree-like structure that makes it feasible to construct a system with complex data flows that change in real-time at runtime. Our design for this is realised in the Psyclone 2.0 architecture.

Facing-Direction

The system already performs face detection and face recognition and from the raw data coming back from the facial recognition software we are able to estimate the approximate gaze direction sufficiently well for use in a dialogue context. We work with a small number of discrete face angles and map the current measurement to one of these.

Who is Speaking

Again from the face recognition software we are able to extract information about the position and state of the mouth. By performing a time series analysis of these we have been able to create a detector which estimates which of the faces in the image are most likely the current speaker.

Humans Present

We get the number of humans present from the face detector module and we get their identities from the face recognition module.

Task- and Role Negotiation

The robots negotiate the assignment of roles and tasks via the CCMCatalog using the RoleNegotiator and TaskNegotiator modules. More information about these can be found in the document *Final Implementation of the CCM Architecture* (**deliverable T8.D2**).

Task-Dialog Manager (TDM)

Among the complexities of handling dialogue and task execution in a coordinated manner is keeping track of (a) relevant needed content that is already available, (b) content that is needed and provided in the dialogue, (c) roles that need to be assigned, and (d) task assignments. Instead of designing a large number of separate pipelines through which data and control flows, which certainly would meet the requirement of a coherent and manageable system, in CoCoMaps communication between agents and collaborative task execution is coordinated by a central module Task-Dialog-Manager (TDM). The TDM uses a mix of techniques to provide planning capability for plans that are represented in the form of sequential AND-OR trees (stored as JSON structures), allowing for hierarchies composed of sub-tasks with AND or OR relationships. Each robot has its own TDM module and they communicate and negotiate with each other via the CCMCatalog.

Meaning Extractor (MEx)

The MEx receives content from the speech recogniser and, via control signals from the TDM and YTTM, marks it as belonging to a particular turn. The main role of the MEx is to turn the speech content (and, in the future, other real-time content such as deictic gestures, facial expressions, and more) into a viable action – whether question, command, operation on the environment, or some other event relevant and appropriate for the task at hand. The meaning is contextualized by the current task (and sub-task) as tracked by the TDM, and thus the TDM and MEx together implement context-dependent meaning interpretation.

YTTM

The TDM relies on the YTTM (Ymir Turn-Taking Module)[1] to integrate low-level signals related to who has turn. For Demo-0 we implemented a new version of the YTTM from scratch, following the Psyclone pub-sub specification, and integrated it with CoCoMaps. The YTTM has been further improved as the CoCoMaps architecture has expanded for subsequent demos.

Our implementation descends from prior (simulated) demonstrations of its principles for two-party dialogue. The most recent incarnation before CoCoMaps proceeded to demonstrate multi-party turn-taking, serving as a proof of concept that the YTTM could be extended to handle more than two dialogue participants. However, this demonstration did not include content interpretation, planning or coordination.

[1] The YTTM is based on the *Ymir turn-taking model* (Thórisson 2002, Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action, in B. Granström, D. House, I. Karlsson (eds), Multimodality in Language and Speech Systems, 173-207. Dordrecht, The Netherlands: Kluwer Academic Publishers).

MEASUREMENTS

Based on several measurements of the performance of the system and its subsystems, the following numbers were obtained.

Table 1.
Summary of performance measurements relevant to 4WTT.

EVENT	**Success rate** (%)	**Speed Ave.** (msecs)	**SD**	**Useful time** (msecs)	**SD**	**Wasted effort**	**SD**
Role Negotiation Time and effort measurement from one robot deciding that a role needs to be assigned until the negotiation has been completed.	100	0.42	0.087	0.42	-	0	-
Task Negotiation Time and effort from one robot deciding that a task needs to be carried until the negotiation has been completed about which robot has accepted the task.	100	0.0684	0.038	0.42	-	0	-
Turn-Taking Smoothness % turns with no overlaps and <2,5 sec pauses between turns	97	2120	362	182 (830)	830	492	363

Role & Task Negotiation

As the numbers indicate, role negotiation (which robot is 'communicator' and which one is 'task executor') is fast and reliable. Negotiation of tasks occurs between the robots, as they decide which one has to accomplish which given task. The negotiation mechanism, and their supporting processes, operates reliably and efficiently.

Turn-Taking Smoothness

We measure the efficiency of the Turn-Taking system by the average time difference between the reception of the human speech input and the decision taken by the system to act upon it.

- Proportion of turn-taking events with one or more wasted cues: 31%
- Proportion of failed turn-taking events (robot did not take turn): 25.6%
- Success rate (robot took turn): 74.4%

Table 2.
Summary of performance measurements related to 4WTT:
Facing-Direction, Human-Present and Dialog Understanding.

Measure	Ave % Correct	Description of Measure
Facing-Direction The measurement of correctness versus false estimates.	~92	Based on largest mean absolute error (MAE=4,4) as over the full practical (60º) angle for which the measurement is possible
Who is Speaking Event-triggered analysis of mouth movements in video images.	59	The ability of the robots to detect who is speaking, using real-time information from cameras.
Dialog 'Understanding' % turns resulting in correct robot action / event.	66	The ability of the robots to say / do the right thing during the dialog.

Facing-Direction

The calculations for measuring which direction a person's face is turned show relatively high accuracy. However, this number must be moderated by the quality of person detection, because direction is not calculated unless a person is first detected.

Who is Speaking

The system's ability to detect who is speaking is reasonable, although there is room for improvement. This could be done by supplementing this measure with e.g. directional information from the two robot-mounted microphones, body language analysis and perhaps other measures as well.

Dialog Understanding

We get an average of 66% of understood turns. Since dialog understanding relies not only on speech recognition but a host of other data and processes, this number should in fact be somewhat lower than the speech-to-text transcription number. The fact that it isn't is in large part due to our ability to interpret interaction contextually, via the chosen design of the Task-Dialog Module (TDM) and Meaning Extractors (MEx).

CONCLUSION

The presented data shows that we have successfully fulfilled the Milestones and Deliverables related to 4-way turn-taking, the numbers indicating that the KPIs have been met. Although there is room for improvement, the measurements clearly show that:

- Several major and key aspects of 4-way turn-taking have been implemented and integrated into the architecture.
- The current implementation has been demonstrated and recorded to perform reasonably well, with clear demonstration of its principles.
- The architecture, and the modular distributed methodology on which it builds, not only shows integration of target features but points in clear direction with respect to further development and scaling.

DISCUSSION AND FUTURE WORK

Multi-party multi-modal collaboration in human interaction and dialogue is a very important social and practical tool which humans use to communicate and work in groups. It is, however, very hard to define a gold standard as the nature of turn-taking varies significantly between human cultures and different people will find different behaviours more or less understandable and/or acceptable. The software development community still has only taken baby steps towards meeting some of its features discussed in this report, but each step is significant in reaching the goal of allowing computers and robots to interact with humans in a more socially adequate and natural way.

To be sure, there are still some low-hanging fruits yet to be exploited for the purpose of naturalness in the CoCoMaps system, such as e.g. detecting nods and head shakes, while others, such as integrating deictic (and other) gestures and facial expressions into real-time dialogue in a useful and efficient manner may require improvements in semantic interpretation and computer vision. Using conversents' line-of-sight to track more closely (and precisely) to whom each participant is speaking is high on the list of features to improve in the near future.

Even small improvements in the of performance of many modules in the CoCoMaps system would greatly and significantly impact the holistic overall behaviour and "feel" of the system in interaction, as quality tends to flow towards the lowest common denominator. Improvements in the quality of the YTTM and sensor-based information extraction would have the biggest impact as these provide information on which virtually all events in the system rely.

Be this as it may, to serve as a foundation for further development the current state of the CoCoMaps architecture is readily demonstrable by the collected evidence and we are quite excited to see this tangible potential for further work in the coming months and years.