Multiparty Multimodal Social Dialogue with a Human-like Tutoring Agent

eNTERFACE2013 – Project Proposal

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze

Department of Speech, Music and Hearing
KTH Royal Institute of Technology, Stockholm, Sweden

{sameram, beskow, skantze}@kth.se

Abstract – Brief Description

Two of the major driving forces behind the developments of Embodied Conversational Agents (ECAs) are: 1) to provide human users with a natural interface that interacts under human terms (i.e. human-like interaction) that can understand natural interaction signal humans use when interaction with each other; and 2) their potential of having social and affective interaction skills that lack in traditional user interfaces of today, and are essential for the success of many interactional scenarios such as education and collaborative task-solving, commerce, and interpersonal relations [7,8,9].

However, arriving to operational systems that meet these goals, as has been shown in the long standing research on multimodal interfaces, is no easy task. Building these systems first depends on the success of several technologies that would allow the system to build real-time models of the on-going interaction with the system and its status, and second, these systems are intensively interdisciplinary, and require the expertise in several fields that are no always available to a single researcher or at a department.

This project will take an attempt at developing a state-of-the-art rich "embodied multimodal multiparty dialogue system" that targets the aforementioned goals. Believing that considerable research findings and advancements have been reached in sensing technologies that would allow the modeling of complex low-level human signals (such as speech, prosody, head movements, facial expressions), and from that, high-level multimodal fusion of these different signals and from different users, that would allow the system to generate context-adaptive, rich verbal and nonverbal output of social value.

The project will implement a dialogue setup consisting of 4 simultaneous speakers and one newly developed robot head (i.e. The Furhat robot head [1]), in a tutoring like scenario, where the users are competing to solve a task (e.g. play a game). The system will use different multimodal signals to capture input from all the speakers, and to build a situated model of the interaction that infers the state of attention and interaction of the different interlocutors. From these signals, the system will generate verbal, prosodic and head and facial output signals to regulate the interaction and provide feedback to the users and keep the users involved in the task.

Technical Project Description

This project aims at advancing the knowledge and modeling of social and interactional signals used in human conversation, especially in multimodal multiparty collaborative spoken dialogue scenarios. From that, the project also aims at advancing state-of-the-art implementations of multimodal dialogue systems, and the use of input/output real-time signals and behaviors in complex setups of a social nature (such as human-like tutoring systems), where a multitude of signals are extensively used and coordinated along with the stream of words.

The project suggests studying a multiparty collaborative dialogue setup where a group of humans interact with each other and with a dialogue system. The dialogue system will be represented by the Furhat back-project robot head [1] developed at the department (www.speech.kth.se/furhat). The robot head supports the synthesis of speech and facial movements (such as gaze, synchronized lip animation, facial gestures, head movements, etc.). The setup involves 4 persons (split into 2 teams), playing a game or solving a task that includes a physical setup (such as a board game). The members of each team collaborate together to solve the task, while competing with the other team. The dialogue system tutors the teams while solving the tasks using several real time strategies, e.g. using the state of the game board and the status of the players into account. To monitor the humans and the game, the system will use real-time face and head-pose tracking, a microphone array and multiple speech recognizers to track in real-time speaker activity, in addition computer vision techniques to track the status of a shared space of attention (e.g. board game), and to track the players hand gestures. The system will also possibly include other sensors that help in measuring the players' behavior and their involvement in the task, such sensors include armband cognitive load sensors and gaze trackers. The task of the system is to generate time-sensitive subtle signals to coordinate the interaction, such as feedbacks, interruptions, corrections and grounding, head-pose and gaze to coordinate turn-taking, and other verbal and non-verbal affect cues to regulate the conversational engagement of the players in the task.

To bring all these technologies together into an operational system, the project will use the Iris-TK dialogue management framework, also developed at the department [2], the team will provide short training sessions on the platform and its modules. All development on the platform modules and interaction control components will be put available online for the participants.

Figure 1 shows a preliminary chart of the setup with some of the capture devices.

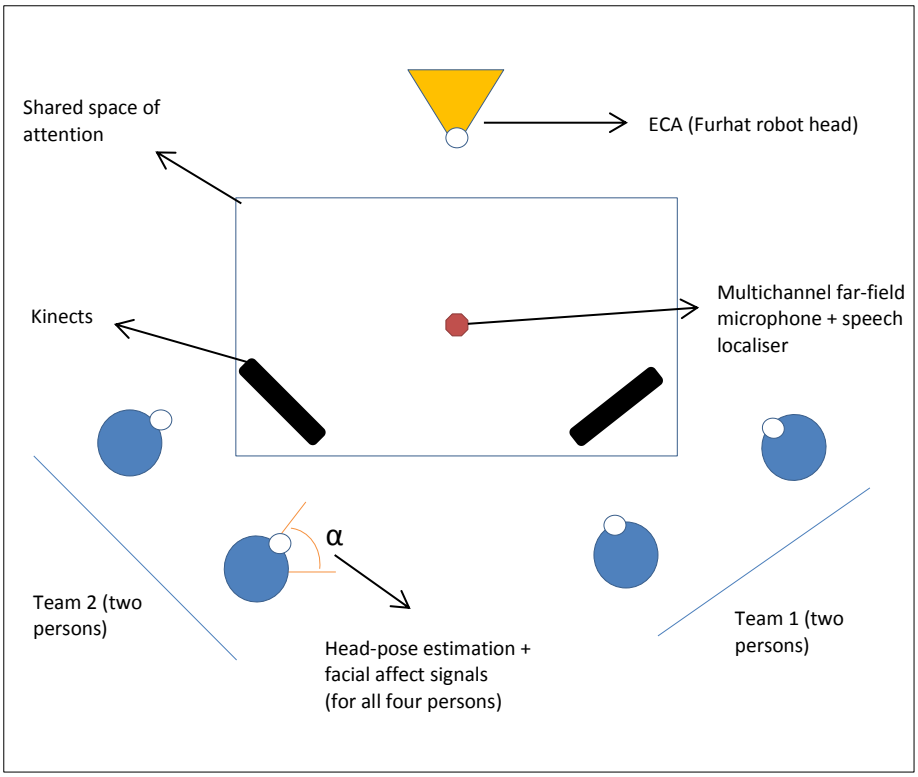


Figure 1 - *A Birdseye view of the situated setup and some of its modules.*

- <u>System input</u>

Low level signals

- Speaker and speech activity detection.
- Simultaneous speech recognition from all far-field microphones.
- Real-time head pose estimation of all speakers.
- Hand gesture recognition.
- Facial expressions detection (e.g. smiles).

High level signals

- Real-time attention modeling of each speaker.
- Modeling of turn-taking behavior in-team and across-team.
- Detection of overlaps (collaborative overlap, competitive overlap).
- Modeling involvement of all users in the task.
- Measuring performance of the team using verbal spoken input, hand gestures, and the status of the shared space of attention.

- System output

Low level signals

- Audio-visual speech output.
- Backchannel and feedback spoken tokens.
- Eye and head movements.
- Facial expressions.
-

High level signals

- Dialogue management.
- Gaze movements (eye and head movements) to regulate turn taking.
- Control of conversational involvement of the different persons, in-team and across-team.
- Feedback (positive and negative) on the team performance in the task.
- Interruptions to resolve conflict and regulate turn-taking between the users.
- Affect signals (e.g. facial mimicry).

Tools & Toolkits

- ***Furhat robot head (Figure 2b)***

The Furhat robot head is a state-of-the-art humanoid head that supports the generation of lip-synchronized synthesized speech, head movements using a 3 degrees-of-freedom neck, and highly natural gaze movements, and facial gestures and expressions. The head uses a novel technique of back-projecting an animation of a face model using a micro-projector, bringing of the flexibility, resolution and speed of computer animation into a physically situated robot head.

For more details, check www.speech.kth.se/furhat, and refer to [1]. The head is provided with a software interface that allows full control of all the head capabilities using a simple but powerful XML messaging protocol (For more on this protocol, refer to [2]).

- ***Microcone multichannel microphone and speech localiser (Figure 2a)***

Microcone[1] is a newly developed 6 channel microphone (over 360 degrees) that provides high quality far-field speech input, along with activation values for the different microphones, allowing for the detection of multiple speakers, and hence overlaps and speaker location. The Microcone will also be provides with an API that can also be controlled using the same XML messaging protocol.

[1] http://www.dev-audio.com/products/microcone/

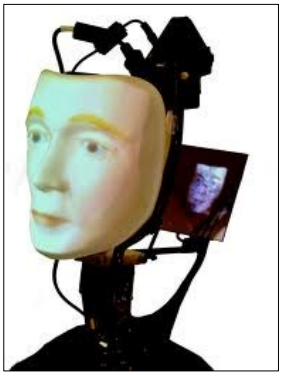
- ***Kinect for multi-person real-time gesture recognition (Figure 2c)***

  Kinect will be used for audio beam-forming using the embedded microphone array, in addition to hand gesture recognition using the depth data.

- ***Tobii stand-alone real-time gaze tracker (Figure 2d)***

  The Tobii X1[2] is a reliable standalone gaze tracker. There is a possibility to use the eye tracker to record gaze behaviour of a subject during a recording, where this subject is replacing the system. Using the data of the eye tracker, a model of head and eye movement can be built for the autonomous generation of these movements by Furhat.

a) Microcone multichannel mic.



b) The Furhat robot head.

c) Microsoft Kinect depth camera and array mic.

d) Tobii X1 gaze tracker.

e) Fraunhofer SHORE™ face tracker.

Figure 2 – *Photographs of the different hardware/software tools planned to be used in the project*

2 http://www.tobii.com/Global/Analysis/Downloads/User_Manuals_and_Guides/Tobii_X1Light_EyeTracker_UserManual.pdf

- ***Fraunhofer real-time multi-face tracker (Figure 2e)***

  Fraunhofer offers a highly reliable multiple faces, realtime tracker called SHORE™[3] [5] that provides the location and rotation of the faces visible in a webcam in realtime. The software also provides other non-verbal measures of the faces in the image, such as facial expressions, age and gender estimates. We have tested this at the lab, and developed a module that would provide information about the visible faces in the image in the shape of XML messages, compatible with the XML protocol. This tracker (or other face tracker provided by another possible member of the team), will be used in combination with the speech localiser, in order to detect speakers head rotation, and from that, infer the addressee of the subjects when producing speech (i.e. towards the system, towards a team member, or towards the other team). The location of the faces can also be used to control the head rotation and eye gaze of the head to establish eye contact with the subjects, even when they change their location in relation to the robot head.

- ***Speech recognition software (either Microsoft ASR, or Nuance).***

Team - Why you really want to be in this project!

This project aims at building a system that is highly interdisciplinary, this means that people with different skills, either working on single modalities, fusion of different modalities, and generation of natural behavior are needed.

If you are a researcher that matches one or more of the following criteria, then you would most definitely love to join this project!

- You are developing a single sensory modality (such as ASR, Prosody modeling, face tracking, gaze modeling, etc.) and you would like to unleash its powers in a wider context and test it in a real-life application.

- You are doing research on aspects of conversational paralinguistics and have made observations and would like to use your observations/hypotheses/models to make human-machine dialogue more natural.

- You are working on dialogue management and want to take advantage of complex natural input and output modules that would allow for more complex dialogue.

- You realize how limited single modalities are, and are working on modality fusion to build more robust and rich interaction with the user.

- You work on human-robot or human-agent user experience and want to evaluate this proposed system.

[3] http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore.html

- You are working on natural nonverbal generation in human-robot interaction and want to collect data, analyze, or implement these models in a human-like robot head.

- You are hands-on, application driven, skilled at programming, and want to work with an interdisciplinary team to build a system that goes beyond state-of-the-art multimodal and multiparty dialogue systems.

Work-plan

This is a tentative schedule for the four-week project. Slight changes will most-certainly take place depending on the finalized setup of the project and the skills of the different project participants.

Week	Description
Week1	- Team building - Tutoring on the different technologies involved - Finalizing project setup and objectives - Modular design of the system - Designing a communication protocol between different modules
Week2	- Data recording and analysis - Implementation of the different modules - First draft report
Week3	- Testing of individual modules - Integration of different tasks
Week4	- Setting up a demonstrator - Evaluation – user study - Final report

Work packages / Deliverables

- Multimodal Dialogue Management
- Visual Input – visual situation modeling (e.g. game-board modeling)
- Array Speech Recognition (Kinect Microphone Array)
- Prosodic and cognitive load modeling
- Multimodal Conversational engagement
- Gaze and head-pose coordination and synthesis
- Hand and finger tracking/pointing gestures recognition
- Real-time multi-face and head pose tracking

Project leaders & Participants

Department for Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden.

Samer Al Moubayed[4] is a postdoctoral researcher at the Department of Speech, Music and Hearing, at KTH, Stockholm, Sweden. His background is in computer science, with specializations in artificial intelligence – speech and language technologies. He received his PhD from KTH in 2012, for his studies and developments on human-robot face-to-face interaction. Samer has been part of several EU projects, including H@H, MonAMI, and IURO. His main work and interest are embodied dialogue systems, multimodal synthesis, and nonverbal social signal processing

Jonas Beskow[5] is an associate professor in speech technology and communication, with main research interests in the area of audio-visual speech synthesis, talking avatars and virtually- and physically embodied conversational agents. He has participated in numerous EU projects related to multimodal speech technology in human-machine interaction and accessibility applications, including PF-STAR, SYNFACE, CHIL, MonAMI, HaH, IURO and LipRead. He is currently the principal investigator of two nationally funded projects in the area of sign language and gesture in face-to-face interaction. He is involved in two start-up companies in the domain of talking avatars, and is one of the developers of the open source speech processing tool WaveSurfer.

Gabriel Skantze[6] is a senior researcher (Docent) in speech technology at the Department of Speech Music and Hearing, KTH, Stockholm, Sweden. He holds a PhD in speech communication from KTH. During his studies he specialized in error handling and miscommunication in spoken dialogue systems. His current research focus is on real-time models of spoken dialogue and empirical studies of human-robot interaction. He has participated in numerous EU projects related to dialogue systems and robotics, including CHIL, MonAMI and IURO. He is currently the principal investigator of a nationally funded project in the area of multimodal incremental dialogue processing.

In addition to other external participants, three PhD students from the Department for Speech, Music and Hearing will join the project and work on it for the whole period of the summer school:

Martin Johansson – Dialogue modeling and management.
Kalin Stefanov – Gesture recognition.
Catharine Oertel – Attention control and conversational involvement in dialogue.

Requirements

The project will require a dedicated, relatively large (preferably round) table, four dedicated chairs, 2-3 camera-stands (tripods), electricity hubs and extension cables, in a relatively quiet room.

[4] http://www.speech.kth.se/staff/homepage/index.html?id=sameram
[5] http://www.speech.kth.se/staff/homepage/index.html?id=beskow
[6] http://www.speech.kth.se/staff/homepage/index.html?id=skantze

References – Reading Material

The references marked with a *, are highly recommended to be read prior to the start of the project.

* [1] Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B.(2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. *In Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), Cognitive Behavioural Systems*. Lecture Notes in Computer Science. Springer.

* [2] Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA.

* [3] Al Moubayed, S., Skantze, G., Beskow, J., Stefanov, K., & Gustafson, J. (2012). Multimodal Multiparty Social Interaction with the Furhat Head. In Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI. Santa Monica, CA, USA

[4] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an Emerging Domain. Image and Vision Computing, 31(1):1743–1759, 2009.

[5] Kueblbeck, C. and Ernst, A. 2006. Face detection and tracking in video sequences using the modified census transformation. *Journal on Image and Vision Computing*, vol. 24, issue 6, pp. 564-572, 2006, ISSN 0262-8856

* [6] Bohus, D. & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *Proc. of ICMI'10*. Beijing, China

[7] Clifford, N., Steuer, J. & Tauber, E. (1994). Computers are social actors. *CHI '94: Proc. of the SIGCHI conference on Human factors in computing systems,* ACM Press, pp. 72–78

[8] Cohen, P. (1992). The role of natural language in a multimodal interface. In proc*. of User Interface Software Technology (UIST '92) Conference*, Academic Press, Monterey, CA, pp. 143–149.

[9] Cohen, P. & Oviatt, S. (1995). The role of voice input for human–machine communication. *Proc. of the National Academy of Sciences*, vol. 92(22), pp. 9921–9927.

* [10] Hjalmarsson, A & Oertel, C. (2012). Gaze Direction as a Back-Channel Inviting Cue in Dialogue. *Workshop on Real-time Conversational with Virtual Agents (RCVA),* Santa Cruz, CA. U.S.A.

* [11] Skantze, G., & Gustafson, J. (2009). Attention and interaction control in a human-human-computer dialogue setting. In *Proc. of SigDial 2009*. London, UK.

* [12] Oertel, C., Cummins, F., Edlund, J., Wagner, P., & Campbell, N. (2012). D64: a corpus of richly recorded conversational interaction. Journal of Multimodal User Interfaces.