Speech, gaze and gesturing – multimodal conversational interaction with Nao robot

Graham Wilcock and Kristiina Jokinen
University of Helsinki

The general goal of this project is to learn more about multimodal interaction with the Nao robot, including speech, gaze and gesturing. As a starting point for the speech interaction, the project has a more specific goal: to implement on Nao a spoken dialogue system that supports open-domain conversations using Wikipedia as a knowledge source. A prototype of such an open-domain conversation system has already been developed using Python and the *Pyrobot* robotics simulator (Wilcock and Jokinen, 2011). Preliminary work has been done with the Nao *Choreographe* software which supports Python, but *Choreographe* does not include Nao's speech recognition or speech synthesis components.

We are also interested in multimodal communication features for the robot, especially gaze-tracking and gesturing. These need to be integrated with the spoken conversation system. The robot needs to know whether the human is interested or not in the conversational topic, and the human's gaze is important for this. The robot should also combine suitable gestures and body language with its own speech turns during the conversation.

2. Project's objectives (max 1 page)

The main goal of the project is to extend the Nao robot's interaction capabilities by enabling Nao to make informative spoken contributions on a wide range of topics during conversation. The speech interaction will be combined with gaze-tracking and gesturing in order to explore natural communication possibilities between human users and robots. If eye-tracking equipment is available we will use it, but if not we will provide simulated gaze information so that the role of gaze-tracking can be integrated in the interaction management.

- To learn basic techniques for spoken dialogue modeling and conversational chatting, especially related to topic management and presentation of new information
- To learn complex issues related to synchrony and unification of multimodal communication models, especially speech and gaze.
- To implement simple spoken conversational interactions with a robot agent
- To integrate some of the available speech, face, gaze, and gesture recognition technologies into the human-robot interaction
- To explore possibilities of natural intuitive human-robot interaction
- To learn to implement and develop conversational models for the Nao robot
- To learn techniques and theories for useful future interactive applications

3. Background information (max 1 page)

Human-robot interaction is an area where much work and interest has recently been focussed. Robots allow development and evaluation of integrated technological platforms for various input and output modalities, and they also seem to come close to the break-through application that would support and justify the complex theories of natural (language) communication. The rich interaction capability that spoken dialogue management has brought with itself has made ordinary users as well as expert researchers skeptical about the possibilities for natural language-based communication: in internet-based service application spoken natural language has not been favoured because of the less robust technology in recognizing human speech. However, conversational interfaces, which do not aim at completing a particular task in the most efficient manner, have brought natural language interfaces onto the stage again: it is not only the most economical interaction that is preferred but the one that allows associative and free-flowing conversations. Also service providers have been pushed to change the ideal of the useful and usable behavior: they are expected to gather information from the user's preferences and to provide information that is new and interesting to the user. In this project we focus on human-robot interaction related to communication that is meant to be chatty and interesting. In particular the robot needs to give explanations about its own actions and what it is doing. This kind of interaction is important in the context of "socially interactive robots", where the robot needs to provide a natural interface for interacting with users. Multimodal interfacing is important in this context as well, with the hypothesis that much of the feedback behavior is conducted using gaze and gesture signals rather with explicit speech, and for a smooth interaction it is important that the robot agent is able to recognize some of these signals and to react to them in an appropriate manner. We thus assume a holistic view of communication where shared understanding is constructed through the interaction among the interlocutors and between the interlocutors and the environment.

Please see the list of references in section 8 for some of the background literature. It also contains some research that will be used as the basis for the lectures and research work. A more complete list of references will be given later, containing references to other work on human-robot interaction (e.g. work at Waseda and Doshisha Universities in Japan, distance robot control work in Budapest, and others)

4. Detailed technical description (max 3 pages):

(a) Technical description;

The general goal of this project is to learn more about multimodal interaction with the Nao robot, including speech, gaze and gesturing. As a starting point for the speech interaction, the project has a more specific goal: to implement on Nao a spoken dialogue system that supports open-domain conversations using Wikipedia as a knowledge source. A prototype of such an open-domain conversation system has already been developed using Python and the *Pyrobot* robotics simulator (Wilcock and Jokinen, 2011; Jokinen and Wilcock, 2011).

We are also interested in multimodal communication features for the robot, especially gaze-tracking and gesturing. These need to be integrated with the spoken conversation system. The robot needs to know whether the human is interested or not in the conversational topic, and the human's gaze is important for this. If eye-tracking equipment is available we will use it, but if not we will provide simulated gaze information so that the role of gaze-tracking can be integrated in the interaction management. The robot should also combine suitable gestures and body language with its own speech turns during the conversation. This requires a model of when to gesture and what kind of gestures to use (hands, head, body).

As we do not yet have access to a Nao robot, preliminary work has been done with the Nao *Choreographe* software which supports Python, but *Choreographe* does not include Nao's speech recognition or speech synthesis components.

(b) Resources needed: hardware and software (have a look on the eNTERFACE'12 website for a full description of the available hardware)

- Nao robot, including speech recognition and speech output components (language is not a major issues although of course English would work best for us!)

- Eye-tracker, if possible

5. Work plan and implementation schedule (max 1 page): A tentative timetable detailing the work to be done during the workshop

Week 1: Planning and introduction, basic issues in multimodal interaction modeling and robot programming
Week 2: First interactive model (speech + Nao)
Week 3: Multimodal extensions (gaze + Nao)
Week 4: Final evaluation

6. Benefits of the research (max 1 page):

- learning to program the Nao robot
- learning to integrate speech and gaze in the interaction model for Nao
- learning about human-robot interaction
- possible future research projects

7. Profile of the team:
(a) Leader (with a 1-page max CV)
Co-leaders: Graham Wilcock and Kristiina Jokinen

Kristiina Jokinen is Adjunct Professor and Project Manager at University of Helsinki and leads the 3I (Intelligent Interaction and Information Systems) Research Group. She is Adjunct Professor of Interaction Technology at University of Tampere and Visiting Professor at University of Tartu, Estonia. She has played a leading role in many academic and industrial research projects, and served in several programme and review committees. She is EACL-2012 Workshop chair, and was invited speaker at the IWSDS-2011 conference. She is Secretary-Treasurer of SIGDial. Her research focuses on natural language communication, spoken dialogue systems, human-human and human-machine interaction, multimodality (gesturing, eye-gaze), and corpus collection. She has published many papers and three books: Constructive Dialogue Modelling - Speech Interaction and Rational Agents (John Wiley), Spoken Dialogue Systems (with M.McTear; Morgan & Claypool) and New Trends in Speech-based Interactive Systems (edited with F.Chen; Springer).

Graham Wilcock is Adjunct Professor and University Lecturer in Language Technology at University of Helsinki. He is a member of Prof. Jokinen's 3I (Intelligent Interaction and Information Systems) Research Group. He was awarded an IBM Innovation Award for work on Unstructured Information Analytics in 2008. His book *Introduction to Linguistic Annotation and Text Analytics* (Morgan & Claypool, 2009) has regularly been in the Amazon best-sellers lists for Natural Language Processing books.

(b) Staff proposed by the leader (with 1-page max CVs).

You may propose some members of your future team. If possible, try to avoid having too many people from your group: part of the benefit of eNTERFACE is to let people meet and share experiences from different places, and possibly in different languages

(c) Other researchers needed (describing the required expertise for each, max 1 page)

Local help with Nao robot set-up (we understand that e.g. Dr Jean-Baptiste Tavenier is an expert on the API)

8. References: This is very important to mention the relevant literature for future members of the team so that they can prepare themselves.

Jokinen, K. (2009). *Constructive Dialogue Modelling – Speech Interaction and Rational Agents*. John Wiley & Sons, Chichester, UK.

Jokinen, K. and McTear, M. (2009). *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Jokinen, K. (2007). User Interaction in Mobile Navigation Applications. In Meng, L., A. Zipf, and S. Winter (eds.) *Map-based mobile services - usage context, interaction and application*, Springer series on Geoinformatics and Cartography. Springer-Verlag Berlin/Heidelberg. pp. 168-197.

Jokinen, K., M. Nishida and S. Yamamoto (2012). Modelling eye-gaze behaviour for interaction management. ACM TiiS Journal .

Jokinen, K. and S. Scherer (2012). Embodied Communicative Activity in Cooperative Conversational Interactions - studies in Visual Interaction Management. *Acta Polytechnica. Journal of Advanced Engineering*

Jokinen, K. and G.Wilcock (2011). Emergent Verbal Behaviour in Human-Robot Interaction. In Baranyi, P. et al (ed.): *Proceedings of 2nd International Conference on Cognitive Infocommunications (CogInfoCom 2011*), Budapest.

Wilcock, G. and K. Jokinen (2011). Adding Speech to a Robotics Simulator. *Proceedings of the Paralinguistic Information and its Integration into Spoken Dialogue Systems Workshop*, Granada, Spain.

9. Additional information (mandatory): Indicate there any other information you consider useful for the evaluation of the proposal.

At the University of Helsinki, the project is part of the research within the 3I - Intelligent Interactive Information Systems Research Group (3IRG) (www.helsinki.fi/~kjokinen/3I).

We also have excellent contacts with the Hungarian Academy of Sciences, and cooperation exists to conduct joint research in the field of relation between behavioural sciences and cognitive infocommunications. Jokinen is in the area chair of interaction modeling for the CogInfoComm –initiative and workshop series initiated by Dr. Baranyi.