Multimodal Open-domain Conversations with the Nao Robot

Kristiina Jokinen and Graham Wilcock

Abstract In this paper we discuss the design of human-robot interaction focussing especially on social robot communication and multimodal information presentation. As a starting point we use the WikiTalk application, an open-domain conversational system which has been previously developed using a robotics simulator. We describe how it can be implemented on the Nao robot platform, enabling Nao to make informative spoken contributions on a wide range of topics during conversation. Spoken interaction is further combined with gesturing in order to support Nao's presentation by natural multimodal capabilities, and to enhance and explore natural communication between human users and robots.

1 Introduction

In recent years, human-robot-interaction has been an active research area resulting in development of integrated platforms for interactive applications with various input and output technologies, as well as opportunities to study natural human-machine interaction with rich communicative capabilities. It is envisaged that future interactive systems will operate in ubiquitous computing contexts and smart spaces, and will be characterized by a flexible use of modalities, such as speech, gesturing, touch, gaze, and movement, so that users can readily understand how to get their task completed without needing to think about how the interaction should take place. Human-robot interactions should thus afford natural communicative capabilities and offer the user more intuitive and conversational ways for interaction [5].

Kristiina Jokinen and Graham Wilcock
University of Helsinki, Finland
e-mail: kristiina.jokinen@helsinki.fi, graham.wilcock@helsinki.fi

Novel technology enables robotic agents not only to record, analyse, and react to the changing environment, but also to be sensitive to users' presence, their communicative needs, and social norms. This is useful in different applications, and can be especially deployed in the realm of social robotics [3] which focuses on communicating robots, capable of interacting and cooperating with humans, and exhibiting relevant social behaviours in the context of human society and culture. Moreover, social robots exemplify conversational interfaces which do not necessarily aim at the most economical interaction in terms of numbers of contributions and turn-exchanges, but at interaction that allows associative and open-domain conversations which emphasise mutual rapport and construction of social bonds.

In this paper we discuss the design of human-robot interaction exemplifying aspects of social robot communication related to multimodal information presentation. The robot gets information about topics from Wikipedia and presents it to the user in speech. We explore how to present the new information to the user, and how to detect human contact and the partner's interest in continuing. In particular, multimodal observations are important in this context, assuming that much of the feedback behavior in conversations is conducted using gaze and gesture signals rather than explicit speech.

As a starting point for speech interaction we use WikiTalk [17], a system that supports open-domain conversations using Wikipedia as a knowledge source, previously developed using a robotics simulator [9, 10]. We show how to extend this model to a situated agent that is capable of multimodal communication, and refer to a prototype implementation on the Nao robot made at the eNTERFACE 2012 Summer Workshop in Metz (Figure 1).

Fig. 1 Human-robot interaction at eNTERFACE 2012.

The structure of the paper is as follows. We give an overview of the dialogue modelling framework and its application to the Nao robot in Section 2. In Section 3 we discuss two important research issues concerning interaction with the Nao robot: the flow of information and the use of gestures in information presentation. The implementation of the Nao WikiTalk prototype and its evaluation at the eNTERFACE 2012 Summer Workshop are briefly described in Section 4. Conclusions and future work are discussed in Section 5.

2 Communicative enablements in the context of human-robot interaction

The Constructive Dialogue Model (CDM) [5] assumes that the speakers are rational agents who interact in a cooperative manner. The speakers coordinate and control their interaction by providing feedback on the basic enablements of communication: Contact, Perception, Understanding (CPU), and Reaction. The enablements, set out by [1], are independent requirements for the communication to take place in the first place, and they operate on different levels. They can be said to model the agent's awareness of the communication as well as their involvement in the communicative interaction.

Contact refers to the fact that the agents need to be in a suitable proximity so that hearing and seeing, and in some cases also touching the partner, are possible, while Perception refers to the agent's conscious perception that the partner is sending meaningful symbols which are intended to be understood and evaluated as a communicative message in the current context. The enablements of Understanding and Reaction concern more intentional functioning of the agent: finding a semantic interpretation for the partner's action and producing one's own behaviour as a reaction to it.

Successful communication requires that the enablements are fulfilled, and the agents thus monitor each other and the communicative situation so they can react to problems and proactively avoid possible errors and misunderstandings. For instance, if the agent has lost contact, is not interested, or does not understand the partner's contribution, there is an obligation to make these CPU problems known to the partner, who, analogously, is obliged to adjust their communication to the level that addresses the problem situation.

The Aldebaran (http://www.aldebaran-robotics.com) Nao robot has many sensors of different types that it uses to perceive its environment. For instance, it has four microphones and two cameras in its head that provide sound and vision about the environment, sonar sensors to check the distance of objects in its vicinity and tactile sensors on its head and body which are triggered when touched. By mapping the robot's sensor technology and its general processing capability to the basic enablements for communication, we can implement the concepts of a general communicative theory with the help

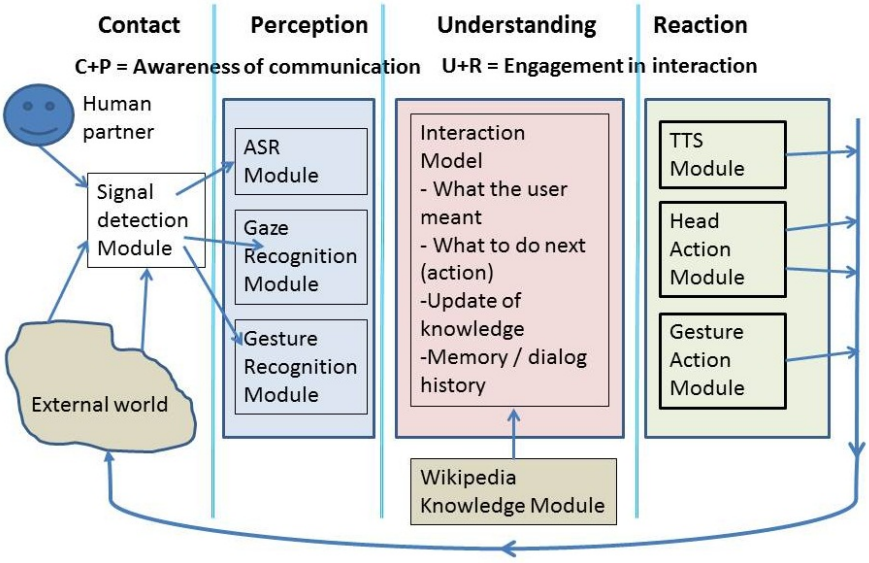

Fig. 2 Overall view of the CDM architecture in human-robot interaction.

of the robot's physical devices. Figure 2 depicts how CDM and the basic enablements are interpreted in the context of human-robot interaction.

Contact and Perception form the basis of the robot's awareness of the communicative situation. Contact, or the detection of the presence of another agent, can be operationalised through the robot's signal detection mechanism: the perception of a visual or auditive signal is a sign of another agent or an object being in contact range. Perception, on the other hand, can be implemented through various modality recognition engines such as speech, gaze and gesture recognition modules, which produce the first, signal-level analysis of the perceived auditive and visual signals.

The agent's engagement in the interaction requires conscious action that indicates the agent's active participation in the situation and willingness to cooperate with the partner. This includes Understanding, or further analysis of the signals through which the pragmatic meaning of the partner's contribution is constructed with respect to the agent's intentions and the current task, and Reaction, or the production of an appropriate response to the partner's message in the communicative situation. The Understanding component corresponds to the traditional dialogue manager that creates a semantic interpretation for the user's contribution and makes the decisions about what to do next. In the current implementation Understanding is a state-based module which coordinates WikiTalk conversations and the topic management with respect to the Wikipedia articles. The Reaction components refer to the robot's speech and motor control engines through which the execu-

tion of the robot's response takes place. They also render the communicative action with respect to the available modalities: speech and gesturing.

The agent's behaviour affects the external world, and the changes as well as the partner's reaction to them will function as a new input to the agent. The robot gets information about the changes via its sensors and start a new analysis and reaction according to the new input. Through the communicative cycle of action and observation, the agent can learn how its actions change the world, and formulate suitable interaction strategies to be used in different communicative situations. If the robot's functionalities include a learning algorithm, it is possible to implement adaptative behaviour.

3 Dialogue modelling with New Information

Dialogues are modelled as a series of dialogue states, representing the agent's beliefs of the external world and the current state of a communicative situation. Each transition from one state to another (or looping in the same state) corresponds to the communicative cycle described in Figure 2, and is characterised by the new information that is conveyed by the agent's action, and used to update the shared dialogue context. An important aspect of the interaction management is the flow of new information that is exchanged between the partners: it is important to show continuation with the topic or mark awkward shifts so as to maintain coherence of the overall dialogue.

One of our goals is to study how to present information in speech so as to help the partner to understand what the new information is. One of the most obvious mechanisms is to use prosody, and mark the new information by prosodic cues [16]. However, in this paper, the main focus is on non-verbal communication and how much dialogue control information can be conveyed by visual cues, i.e. gesturing, nodding, and body posture. The different research issues related to coherence and NewInfo presentation are briefly discussed in the subsections below.

3.1 Topic management and coherence of presentation

We use Wikipedia as the knowledge base, and follow the WikiTalk approach to enable open domain conversations with the Nao robot (see [10, 17] for more details about WikiTalk). The Wikipedia articles are considered as possible topics that the robot can talk about, while each link in the article is treated as new information that the user can shift their attention to, and ask for more information about. The paragraphs and sentences in the article are considered propositional chunks, or pieces of information that structure the topic into subtopics and form the minimal units for presentation, i.e. they

can be presented in one utterance by the robot. The challenge in presenting Wikipedia information is how to convey its structure to the users so that they can readily understand which are the new information links, and how to navigate in the topic structure smoothly. In other words, dialogue modelling should support interaction that affords natural information flow [5].

In dialogue management, topics are often managed by a stack, which is a convenient last-in-first-out mechanism to handle topics that have been recently talked about. However, stacks are a rather rigid means to describe the information flow in cases where the dialogues are more conversational and do not follow any particular task structure. We prefer topic trees [13], which enable more flexible management of the topics. The trees can be traversed in whatever order, while the distance of the jumps determines the manner of presentation of the information. Moreover, we use the concepts of Topic and NewInfo [5], where Topic refers to the particular issue (Wikipedia article) that the speakers are talking about, and NewInfos are the parts of the message (the hyperlinks) that are new in the context of the current Topic.

It must be emphasized that dialogue coherence, or the discourse relations between consecutive utterances that enable the listener to infer what their connection is, becomes rather straightforward: we can rely on the structure of Wikipedia to provide coherence for us. As the articles have already been written as coherent texts, and in particular the hyperlinks between the articles have been inserted so that they form coherent hypertexts, we can assume that the content of the topics and the NewInfo links that will be presented to the user form a coherent discourse. Interaction is then driven by the user's interests based on the content of the presentation rather than a particular task structure that would constrain the suitable topics. However, what is important in WikiTalk is to capture the partner's attentional state in such a way that the partner can focus attention on the available NewInfos.

3.2 Gesturing and presentation of NewInfo

Gestures and body movement play important roles in human communicative behaviour, and can be directly related to the information flow of interactions [11]. Hand movements are often used to visually explain the topic, e.g. the size or shape of an object, while beat gestures are usually associated with emphasis and rhythm of the speech, and deictic gestures with pointing or attention catching. We have experimented with various gestures to make the robot's presentation more expressive, especially to mark the NewInfo and to structure the WikiTalk presentation. Such gestures indicate to the partner how the conversation is to be understood and divided into communicatively important segments [6]. They are distinguished from iconic gestures and beats in that they aim to catch the partner's attention, and thereby they also control the dialogue flow. Kendon [11] calls them meta-discursive gestures.

Function	Body	Speed	Movement	Frequency	Orientation
Greeting	Hand	Fast	Complex (wave)	Repeated	Palm vertical
Start presentation	Both hands	Slow	Front, side	Single	Palm-up
Start topic	Left	Fast	Front, side	Repeated	Palm-up
Give feedback	Head	Fast	Down	Single/ Repeated	Nod
Elicit feedback	Head	Slow	Down	Single	Nod
Emphasise	Left	fast	Up-down	Single	Palm vertical
Beat	Left	fast	Up-down	Repeated	Palm vertical
Surprise	Head	slow	Up	Single	Nod

Fig. 3 Gesture taxonomy and parameters.

Much research concerns the use of gestures in interaction [11, 6]. Following Quek [15], we classify hand and arm movements into unintentional movements and gestures, and divide the latter into manipulative commands and communicative gestures. Manipulative commands are gestures that the user issues to control the robot. Sophisticated gesture recogniton would require a robust recognition algorithm, so we experimented with only one manipulative command: a stop gesture with fingers and palm vertical, that the user can use to stop Nao talking. However, hand recognition was too much dependent on the lighting and background, so in practise we did not include it in the interaction repertoire, but used tapping on the robot's head instead. See [4] for a review of the experiments on hand recognition.

For communicative gestures we use the idea of gesture families, following Kendon [11]. Gesture families consist of gestures which have similar form and meaning, such as Palm Down (Open Hand Prone), Palm Up (Open Hand Supine), Palm Sideways (Open Hand Vertical), and Index Finger Extended. They are associated with a semantic theme, for example gestures of the Palm Down family are often used in contexts where something is being denied, negated, interrupted or stopped, while those in the Palm Up family are used in contexts where the speaker is offering, giving or showing something or requesting the reception of something.

Figure 3 shows the gesture taxonomy and parameters in the Nao WikiTalk prototype. Gestures are organized into a gesture library, a collection of behaviours that Nao can choose from. The library consists of gesture families, each linked to a particular semantic theme. Selection of a gesture in a particular communicative context is based on the dialogue situation (give/elicit

feedback, inform, greet) and on the task (continue/change/stop the topic). More detailed description of the robot's gesturing can be found in [14].

We used Nao's Choregraphe tool to model a set of gestures which were then exported as Python code to be run by the dialogue manager. Gestures are parameterized, which helps to constrain the choice to a certain gesture family, and alternatives in the same gesture family are selected in a loop.

One of the main issues in gesture studies is the synchrony of gestures and speech. Speakers coordinate speaking and gesturing in an accurate manner so that the peak of the gesture is aligned with the speech and visually contributes to the information content of the message [11]. Theoretical questions are related to how simultaneous cognitive processing and planning of speech and gestures take place. Although mental modelling of communication is outside the scope of our research, the correct timing of beat gestures with the NewInfo is important for the natural presentation capabilities of the robot: gesturing provides 'silent' feedback to the partner, and prepares them to have the right stance to interpret the forthcoming message in the intended way.

We followed Kendon's three phases (preparation, nucleus and retraction) for temporal segmentation of our gestures, and experimented with methods for counting utterance words and timing their pronunciation. However, often gesturing is too slow, and more accurate synchrony would require access to speech synthesis timings in order to synchronize it with the motorics that implement robot gestures. Further research is needed to provide more accurate methods for the timing and synchrony of gesturing and speech.

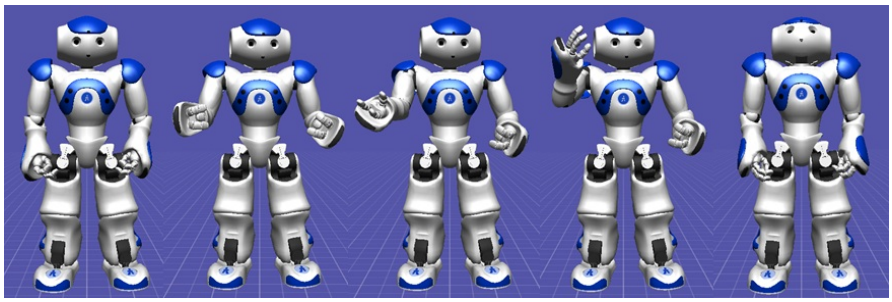

Fig. 4 Some key poses: standing, speaking, start presentation, emphasis, surprise.

Besides gesture and speech coordination, the whole posture of the robot is relevant in providing natural and intuitive presentation. Figure 4 presents some specific Nao key poses. The speaking posture with hands half way up differs from the standing position with hands by the side, and the robot has a particular palm-up presentation gesture, used at the beginning and end of presentations. Emphasis is expressed by single hand beat gestures, while surprise, e.g. after the user interruption, is expressed by a head nod up. In

the current version we cannot use pointing gestures due to the anatomy of Nao's hand: Nao can only have palm closed or palm open.

4 Implementation and evaluation

The implementation of the WikiTalk system on Nao and the development of the multimodal interaction modules were done during the eNTERFACE 2012 Summer Workshop in Metz. Further details are given in [2].

Technologies	Nao behaviours
Wikitalk	– Knowledge from Wikipedia – Segmentation of articles into paragraphs, sentences, and NewInfos
Face recognition	– Recognition of an object in the vision field – If human face, start to follow – Recognition of human contact and interest
Gesturing	1) User commands: Stop! 2) Nao's own communicative gestures: – Presentation of information with palm-up – Elicit of user feedback with nod down – Surprise at interruption with nod up – Emphasis of NewInfo by a beat
Dialogue managment	– Start a conversation – Continue with the same topic – Change to a new topic

Fig. 5 Basic technologies and behaviours implemented at eNTERFACE 2012.

The basic technologies, including WikiTalk, face recognition, gesturing and dialogue management, listed in Figure 5, were integrated into the Nao system. The robot's functionalities include segmenting articles into paragraphs, sentences and NewInfos, detecting a human face, using communicative gestures, as well as managing the dialogue. For instance, to track the human face and thus draw conclusions about the user's interest, a simple face tracker was implemented. At the start of the interaction, when Nao detects a face, it makes contact by saying 'Hello, I can tell you many interesting things, what would you like to hear?'. If the face disappears for more than a few seconds at any time during the dialogue, Nao pauses and explicitly asks if the user wishes to hear more. If there is no answer, Nao stops and waits. More details about using non-verbal cues in human-robot interaction are given in [4].

The Nao Wikitalk prototype was evaluated with 12 students and staff at the eNTERFACE 2012 workshop. The evaluation is reported in [2] and [14].

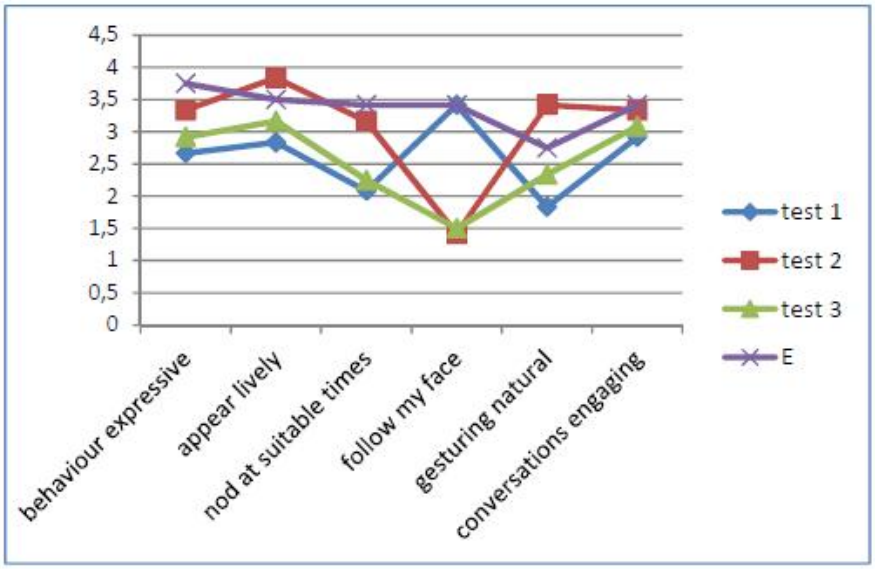

Fig. 6 User expectations (-X-) and experience on some aspects of the 3 system versions.

Following the evaluation scheme proposed in [8], the users were asked to fill in a questionnaire twice, first to capture their expectations of the system before interacting with it, and then to measure their experience of the system after their interaction with it. The subjects tested three different versions of the system: with gaze tracking only (no hand gestures), with small gesturing (head nods and presentation postures) and with full gesturing, along the dimensions of Interface, Responsiveness, Expressiveness, Usability and Overall Experience. Averaged results of some of these aspects between user expectations and experience are shown in Figure 6.

Version 2 (small gesturing) exceeded user expectations in appearing more lively and gesturing more naturally than expected. However, in all versions the behaviour was not as expressive as expected and the timing of nodding was not suitable. Perhaps 'expressiveness' requires more facial expressions than is possible with Nao and the scores were thus low. It is also likely that asynchrony of gesturing was the reason why Version 3 with 'big' visible gestures was rated less natural than Version 2 with 'small', less visible gesturing. Concerning engagement in the interaction, however, Version 2 seemed to fulfill user expectations, and the other versions also ranked highly, which supports the view that interactions with robots are generally engaging.

The evaluation data is available for further research purposes by request from the authors. The data is unique in that it is a fairly large, systematic collection of open-domain multimodal human-robot interactions.

5 Discussion and future work

The paper describes interaction with the Nao robot from the point of view of Constructive Dialogue Modelling and demonstrates how the framework can be applied to the Nao WikiTalk application. Nao's interaction capabilities are greatly extended by the multimodal aspects related to gesturing, and by enabling it to make informative spoken contributions on a wide range of topics during conversations. As far as we know, this is the first multimodal human-robot conversational interaction system that is open-domain.

The evaluation of the prototype system implemented at the eNTERFACE 2012 Summer Workshop in Metz shows that the system can engage humans in interaction which is lively and fairly natural. The combination of speech communication with gesturing supports natural interaction between human users and robots, and enhances possibilities for successful application of the technology to various other types of tasks such as educational applications, tourist guiding, and game interfaces.

At the summer workshop we also explored other multimodal features, in particular gaze-tracking and motion capture. Gaze-tracking is important in order to manage smooth turn-taking [7, 12] and to get feedback about the partner's interest in the topic. As humans direct their gaze towards objects of interest, it is useful if the robot can infer where the partner's attention is focussed, and if they are still interested in what it is presenting. Further experiments are planned to model agents' awareness and focus of attention, and to explore the notion of conversational engagement. Finally, we also experimented with motion capture technology using Kinect as one of the robot's inputs. Preliminary work on this is presented in [2].

Acknowledgments

We would like to thank Adam Csapo, Emer Gilmartin, Jonathan Grizou, JingGuang Han, Raveesh Meena and Dimitra Anastasiou for implementing and evaluating Nao WikiTalk and the multimodal interaction capabilities on the Nao robot at eNTERFACE 2012 in Metz in July 2012.

References

1. Allwood, J.: Linguistic Communication as Action and Cooperation: A Study in Pragmatics. Gothenburg Monographs in Linguistics 2. University of Gothenburg (1976)
2. Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., Wilcock, G.: Multimodal conversational interaction with a humanoid robot. In: Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012). Kosice (2012)

3. Fong, T., Nourbaksh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems **42**, 143–166 (2003)
4. Han, J., Campbell, N., Jokinen, K., Wilcock, G.: Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. In: Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012). Kosice (2012)
5. Jokinen, K.: Constructive Dialogue Modelling: Speech Interaction and Rational Agents. John Wiley & Sons (2009)
6. Jokinen, K.: Pointing gestures and synchronous communication management. In: A. Esposito, N. Campbell, C. Vogel, A. Hussein, A. Nijholt (eds.) Development of Multimodal Interfaces: Active Listening and Synchrony, pp. 33–49. Springer (2010)
7. Jokinen, K., Harada, K., Nishida, M., Yamamoto, S.: Turn-alignment using eye-gaze and speech in conversational interaction. In: Proceedings of 11th International Conference on Spoken Language Processing (Interspeech 2010). Makuhari, Japan (2010)
8. Jokinen, K., Hurtig, T.: User expectations and real experience on a multimodal interactive system. In: Proceedings of Ninth International Conference on Spoken Language Processing (Interspeech 2006). Pittsburgh, USA (2006)
9. Jokinen, K., Wilcock, G.: Emergent verbal behaviour in human-robot interaction. In: Proceedings of 2nd International Conference on Cognitive Infocommunications (CogInfoCom 2011). Budapest (2011)
10. Jokinen, K., Wilcock, G.: Constructive interaction for talking about interesting topics. In: Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012). Istanbul (2012)
11. Kendon, A.: Gesture: Visible action as utterance. Cambridge University Press (2004)
12. Levitski, A., Radun, J., Jokinen, K.: Visual interaction and conversational activity. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality. Santa Monica, USA (2012)
13. McCoy, K.F., Cheng, J.: Focus of attention: Constraining what can be said next. In: C. Paris, W. Swartout, W. Mann (eds.) Natural Language Generation in Artificial Intelligence and Computational Linguistics), pp. 103–124. Kluwer Academic Publishers (1991)
14. Meena, R., Jokinen, K., Wilcock, G.: Integration of gestures and speech in human-robot interaction. In: Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012). Kosice (2012)
15. Quek, F.: Toward a vision-based hand gesture interface. In: Proceedings of the Virtual Reality System Technology Conference, pp. 17–29. Singapore (1994)
16. Swerts, M., Geluykens, R.: Prosody as a Marker of Information Flow in Spoken Discourse. Language and Speech **37**, 21–43 (1994)
17. Wilcock, G.: WikiTalk: A spoken Wikipedia-based open-domain knowledge access system. In: Proceedings of the COLING 2012 Workshop on Question Answering for Complex Domains, pp. 57–69. Mumbai, India (2012)