Evaluation of WikiTalk – User Studies of Human-Robot Interaction

Dimitra Anastasiou[1], Kristiina Jokinen[2], and Graham Wilcock[2]

[1] University of Bremen, Germany
[2] University of Helsinki, Finland
dimitra@d-anastasiou.com,
{kristiina.jokinen,graham.wilcock}@helsinki.fi

Abstract. The paper concerns the evaluation of Nao WikiTalk, an application that enables a Nao robot to serve as a spoken open-domain knowledge access system. With Nao WikiTalk the robot can talk about any topic the user is interested in, using Wikipedia as its knowledge source. The robot suggests some topics to start with, and the user shifts to related topics by speaking their names after the robot mentions them. The user can also switch to a totally new topic by spelling the first few letters. As well as speaking, the robot uses gestures, nods and other multimodal signals to enable clear and rich interaction. The paper describes the setup of the user studies and reports on the evaluation of the application, based on various factors reported by the 12 users who participated. The study compared the users' expectations of the robot interaction with their actual experience of the interaction. We found that the users were impressed by the lively appearance and natural gesturing of the robot, although in many respects they had higher expectations regarding the robot's presentation capabilities. However, the results are positive enough to encourage research on these lines.

Keywords: Evaluation, multimodal human-robot interaction, gesturing, Wikipedia.

1 Introduction

In human-robot interaction (HRI) not only speech, but also other modalities, such as gesture and nodding make the conversation more natural, effective, and user-friendly. However, the evaluation of such intelligent agents requires a comprehensive setup to define correlations between different modalities and their usage, and to investigate human interaction and its basis as an *affordable* model for HRI. *Affordance* is a concept that was brought to HCI by [1] and refers to the properties that suggest to the user the appropriate ways to use the artifact. The concept's use for spoken dialogue systems was suggested by [2]: when interactive systems *afford* natural interaction techniques their interfaces lend themselves to a natural use without users needing to reason how the interaction should take place to get the task completed. [2] also points out the *rationality* of system actions meaning that the system is not regarded as a simple reactive machine or a tool, but capable of acting appropriately in situations which are not directly predictable from the previous action.

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 32–42, 2013.

Concerning the evaluation of spoken dialogue systems, [3] stated that during the past years automatic evaluation and user simulations gained ground in order to enable quick assessment of design ideas without resource-consuming corpus collection and user studies. They distinguished between evaluation approaches (empirical vs. theoretical), conditions (laboratory vs. real usage), and goals, and categorized evaluation types in the following:

- Functional evaluation – Is the system functionally appropriate?
- Performance evaluation – How well does it perform the task it is designed for?
- Usability and quality evaluation – Are potential and real users satisfied with the performance, and how do the users perceive the system when using it?
- Reusability evaluation – Is the system flexible and portable?

In usability testing, questionnaires or subjective evaluations are used to learn from the users what a usable system is. To system designers, subjective evaluations may give more informative data of system functionality than objective performance measures, since they focus on the user's first-hand experience. For instance, recommendations for speech-based telephone services concern three different types of questionnaires:

- Those about the user's background at the beginning of an experiment.
- Those about the user's interactions with the system.
- Those about the user's overall impression at the end of an experiment.

The evaluation's goal and metrics are important factors to define how the different topics are translated into precise questions or statements. In our study we used two types of questionnaires: one at the beginning of the session to collect the user's expectations and one at the end of the experiment to collect the user's experience (cf. [10]).

The paper is laid out as follows: Section 2 presents the Nao WikiTalk application with focus on the gestures that were designed to enhance the robot's presentation and turn-management capabilities. Section 3 presents our user study and its setup, including instructions, the questionnaires and our methodology. Section 4 reports the results of our evaluation and Section 5 concludes the paper with some future prospects.

2 The Nao WikiTalk Application

WikiTalk [11] is a spoken dialogue system for open-domain knowledge access using Wikipedia as a knowledge source. Prototypes of WikiTalk were first developed using Windows Speech Engine and the Pyrobot robotics simulator [4]. The Nao WikiTalk version was implemented at the 8th International Summer Workshop on Multimodal Interfaces in Metz, France in July 2012. The Aldebaran Nao humanoid robot was used as the robot interlocutor in multimodal conversations. The work focused on different research issues, among others, on designing gestures, gaze tracking, integration of body motion with the spoken conversation system and also combination of suitable body language with Nao's own speech turns during the conversation. The main results of the joint work are reported in [5], while multimodal signaling is described in [6], and integration of gesturing and speech in [7].

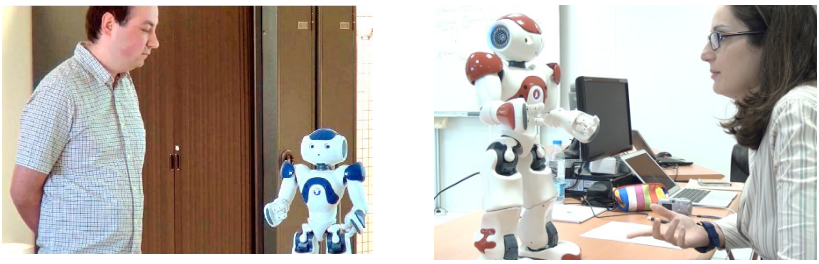

Fig. 1. Users interacting with the Nao robot

The Nao robot by Aldebaran Robotics (http://www.aldebaran-robotics.com) is a fully programmable humanoid robot, which has many sensors and actuators, and is of a convenient size and attractive appearance, with sophisticated embedded software (see Figure 1). Nao supports face and object recognition, speech recognition, text-to-speech, and whole body motion.

2.1 Gestures

Gesturing is a means of communication that can make HCI and HRI more natural, expressive, communicative, and user-friendly. A set of non-verbal gestures were designed in order to enhance Nao's presentation and turn-management capabilities [7]. These apply Kendon's [8] notion of gesture families. The *Open Hand Supine* ("palm up") and *Open Hand Prone* ("palm down") families have their own semantic themes related to giving ideas as well as presenting, explaining, summarizing vs. stopping and halting, respectively [9]. For the presentation capabilities, a set of presentation gestures were identified to mark the topic, the end of a sentence or paragraph, plus beat gestures and head nods to attract attention to hyperlinks (new information), and head nodding as backchannels (see more in [7]). For the turn-management capabilities, the following approach was applied: Nao speaks and observes the human partner; after each information chunk that Nao presents, the human is invited to signal continuation (phrases like 'continue' or 'stop'); Nao asks explicit feedback depending on user's turn; the robot may also gesture, stop, etc. depending on previous interaction. This shows the *rationality* of the system, i.e. situation-dependent appropriate actions of the robot.

3 User Study

In this section we discuss the study's setup (3.1) as well as the questionnaires about the expectations and experience of the participants (3.2), and our methodology (3.3).

3.1 Setup

The user study took place at the 8th International Summer Workshop on Multimodal Interfaces (eNTERFACE 2012) in Metz. We ran user studies to test the application with 12 participants (5 female and 7 male). All participants were members of other projects organized by the summer school. They were in the age group 20-40 and came from various countries, including France, Germany, Switzerland, Greece, and India. The participants interacted with Nao in three phases/tests (5-10 minutes each phase).

The evaluation follows the framework of [10]. The users first answer a questionnaire concerning their expectations about the application, and then, after their experience of using it, they answer the same questions concerning their actual experience. The questions are the same, but their linguistic formulation is adapted to suit the future expectations and the past experiences accordingly.

Before starting the first test, an experimenter explained to the participant the tasks to be done in all the tests, gave a consent form to sign, and also handed an instruction sheet which the participants could take with them in another room where they have the interaction with Nao. Their task was to interact with Nao in an open conversation asking for a topic from Wikipedia and to try out how well it can present interesting information. Our instructions regarding the topics were the following:

- Nao can talk about almost any topic.
- You can change to another (related) topic simply by saying the name of one of the things that Nao mentions.
- You can interrupt Nao any time, by touching the front button on top of its head.
- You can move around and try to catch Nao's attention from different angles.
- You can finish the interaction session by saying *thank you*.

The robot suggests some topics to start with, and the user can shift to related topics by speaking their names after the robot mentions them. The user can also switch to a totally new topic by spelling the first few letters. The instruction sheet listed the main user commands to Nao ('continue', 'repeat', 'enough', etc.), as well as the spelling alphabet (A = Alpha [AL FAH], B = Bravo [BRAH VOH] etc.).

The evaluation by each user was divided in three sessions, lasting about 5-10 minutes each. Each session involved one of three different system versions and accordingly, the users had to evaluate three different interactions with Nao. Table 1 summarises differences between the different system versions The first version did not include gesturing but only face tracking, while the two other versions differed in the number and variety of gestures and posture, to allow us to test the user reactions.

Table 1. Non-verbal gesture capabilities of Nao [8]

System version	*Exhibited non-verbal gestures*
System 1	Face tracking, always in the Speaking pose
System 2	Head Nod Up, Head Nod Down, Open Hand Palm Up, Open Hand Palm Vertical, Listening and Standing pose
System 3	Head Nod Up, Open Hand Palm Up and Beat Gesture (Open Hand Palm Vertical)

3.2 Questionnaires

Each participant filled in a questionnaire four times: first to give their expectations before starting their interaction with Nao, and then after each session to evaluate the system they had just interacted with. The questionnaire focused on the various aspects of the interaction and the robot's presentation. It included the following categories:

a) *Interface*: sound, hand and body movements, face tracking.
b) *Expressiveness*: instinctive, lively, natural way of communication.
c) *Responsiveness*: speed of reaction, appropriate responses, easy to follow.
d) *Usability*: easy to interrupt, easy to know what do next.
e) *Overall*: head tracking/movements, enjoyment in interaction.

Each category included 5-7 statements, specific to the category. In this way, more detailed and accurate information from the user could be collected. The questionnaire was formulated as statements, and the user estimated how much they agreed with each statement on a 5-point Likert scale, from *I strongly agree* to *I strongly disagree*. An example of the *Interface* category can be seen in Table 2.

Table 2. Part of *expectations* questionnaire

I expect to understand quickly the purpose of the sounds emitted by Nao.
I expect to notice if Nao's hand gestures are linked to exploring topics.
I expect to find Nao's hand and body movement distracting.
I expect to find Nao's hand and body movements creating curiosity in me.
I expect to find Nao's face tracking speed appropriate.

Experience was measured with an analogous questionnaire after each user test, with the same categories. A sample of the *Interface* category is in Table 3.

Table 3. Part of *evaluations* questionnaire

I understood quickly the purpose of the sounds emitted by Nao.
I noticed Nao's hand gestures were linked to exploring topic.
Nao's hand and body movement distracted me.
Nao's hand and body movements created curiosity in me.
Nao's face tracking speed was appropriate.

There were also multiple choice questions, like *What do you expect will be the most difficult part of interaction?* At the end there were open-ended questions with free text answers for users to give comments on issues that had not been taken into account.

3.3 Methodology

There were in total 48 completed questionnaires: 12 about expectations and 36 (12 users and 3 evaluation tests) about experience. For the expectations, we calculated the

average of each category (*Interface, Expressiveness,* etc). Regarding the experience, we calculated the average for each category, but also for each evaluation test (Evaluation 1, 2, and 3). More importantly, we compared the expectations with the experience for each category and each evaluation test; for each of the 30 statements individually as well as the average scores. Furthermore, we ran paired t-tests to find out whether there is a statistically significant difference between the expectations and evaluations.

4 Results

4.1 User Expectations

Figure 2 shows the average expectations for the five evaluation categories.

The highest expectations fall under *Usability*, the lowest in *Expressiveness*, indicating that the users expected interaction to be functional, but not very natural. The single lowest expectation was in *Expressiveness* with a mean value of 2.75, and in

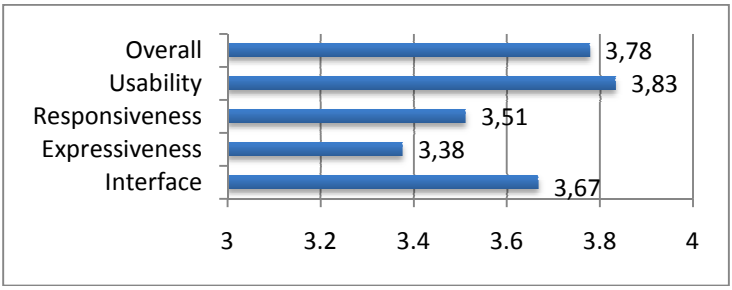

Fig. 2. Average of expectations in all categories (5-point Likert scale)

particular, the user did not expect Nao's gesturing to be natural. The highest expectation with a mean value of 4.25 was shared by the statements *I expect Nao's info to be correct* and *I expect to understand quickly the purpose of the sounds emitted by Nao*.

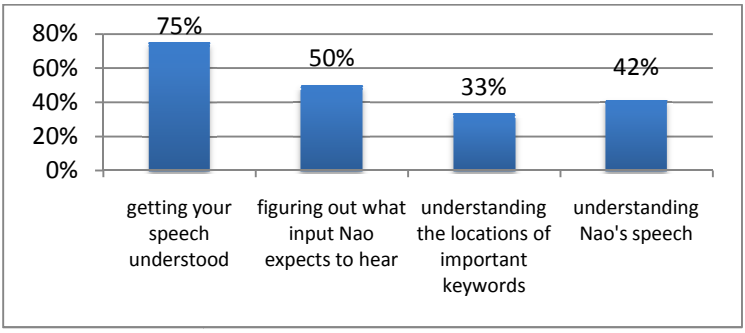

Fig. 3. Ranking of expectations about the most difficult part of interaction

Figure 3 shows the ranking of answers to the multiple choice questions concerning the most difficult parts of the interaction. Expectations about the difficulty of speech recognition (getting your speech understood) are significantly higher than those for speech synthesis (understanding Nao's speech), 75% and 42%, respectively, reflecting perhaps the user's prior knowledge of the problems in speech recognition. Participants were most confident (expected least problems) in that they would understand the locations of important keywords/topics that Nao can talk about (only 33% considered this a problem).

4.2 User Experience

Figure 4 compares the average of the expectations and of the evaluations.

As Figure 4 shows, in all categories (*Interface, Expressiveness,* etc.) the expectations were higher on average than the experience after the sessions. In categories *Interface, Expressiveness* and *Responsiveness,* System 2 was evaluated higher than the others (see Table 1 for the versions and gesture capabilities in each test). This gives support for the original goal by suggesting that the users appreciated and had a more positive experience of the interaction with full repertoire of gestures compared with the one with less expressive presentation. In *Usability*, System 3 was ranked slightly higher than the others, indicating perhaps that the users had become more familiar with how to interact with the system and thus experienced it more usable as well. In all categories, System 1 was ranked lowest, which supports our initial hypothesis that gesturing makes interaction with the robot more natural and expressive.

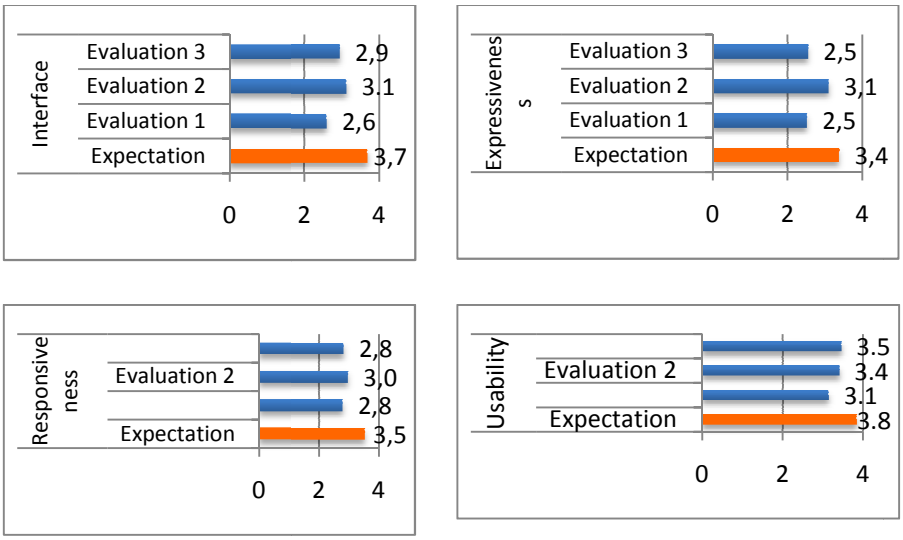

Fig. 4. Comparison between expectations and evaluations (5-point Likert scale)

Looking at individual statements, the highest scores were as follows. In *Interface,* the highest ranked statement in all three tests was *I understood quickly the purpose of the sounds emitted by Nao* (4.2, 4.3 and 4.2, respectively); this statement even exceeded the expectations in the second test slightly (4.2). In *Expressiveness*, the statements *Nao appeared lively* and *Nao's gesturing was natural* scored high in System 2 (3.8 and 3.4, respectively), and also exceeded user expectations (2.7 for both). These are positive results concerning our original goal of making the interaction more expressive by using gesturing. In *Responsiveness*, users ranked *Nao was slow to react* high in each test (3.7; 3.2 and 3.2, respectively), and also *It was easy to stop Nao speaking* (3.0, 3.6 and 3.4, respectively). It is interesting that the statement *Nao was able to change topic when I wanted* scored high in System 2 (3.4) and quite high also in System 3 (3.2), but not so in System 1 (2.8); it is likely that expressive gesturing also added to the user's positive experience of controlling the interaction. In *Usability,* the statement *I knew what to do when Nao stopped talking* was ranked the highest in System 3 (3.8).

4.3 Comparing Experience with Expectations

Figure 5 presents an overview of the evaluation categories. We can deduce from the curves that user's experience was similar to their expectations: *Usability* was ranked high and *Expressiveness* low. However, in *Expressiveness,* System 2 with full set of gestures was ranked much higher than the other two systems. In general, System 2 was experienced as the best of the three system versions, and scored closer to the expectations.

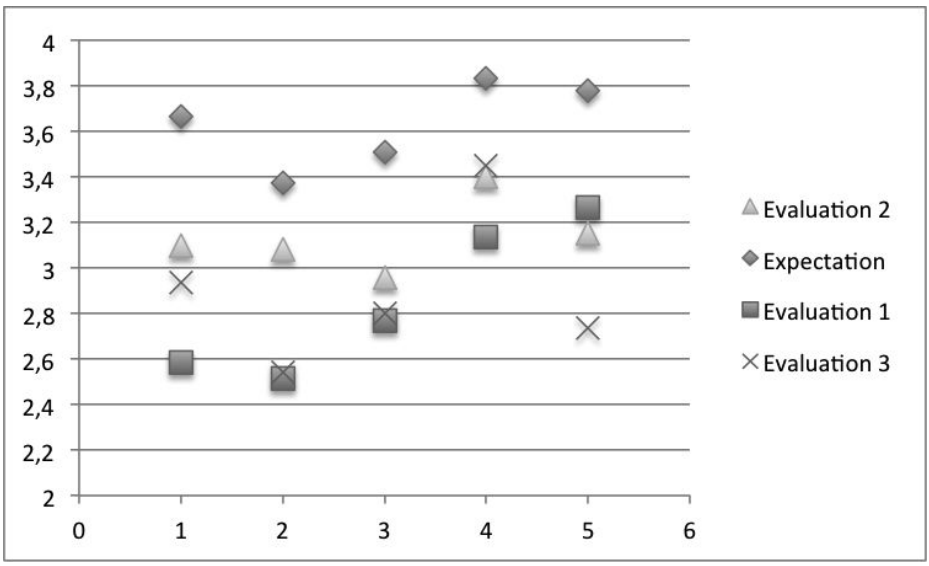

Fig. 5. Overall category and general overview

To test for a statistically significant difference between expectations and experience, we ran paired t-tests (alpha 0.05) on each evaluation category and the results are shown in Table 4. The boldfaced differences are significant on the level $p < 0.05$. We notice that with respect to System 1, the user's expectations in all categories are significantly higher than their experience. With System 3, expectations are significantly higher than the experiences for *Expressiveness* and *Responsiveness*. However, with System 2, experience was significantly lower only with *Responsiveness,* while in other categories the user's experience did not differ significantly from their expectations. It is interesting that the differences in *Expressivenss* of system 1 and *Responsiveness* concerning the Systems 2 and 3 are significant also on the stricter level of 0.01, marking these differences the most prominent ones among the system versions.

Table 4. Paired t-tests between user expectations along the different evaluation dimensions

Eval	**Interface**		**Expressiveness**		**Responsiveness**		**Usability**	
	t(4)	p	t(5)	p	t(7)	p	t(4)	p
1	3.34	**0.029**	5.09	**0.004**	2.9	**0.022**	3.4	**0.027**
2	1.21	0.290	0.77	0.475	4.5	**0.003**	2.3	0.079
3	1.80	0.146	3.26	**0.022**	5.1	**0.001**	1.6	0.174

We also compared the different evaluation versions with each other. In general, interactions with the different systems were not ranked very different from each other, but statistically significant differences could also be found. Interactions with System 2 were significantly more *expressive* than those with System 3, supporting the claim that a full repertoire of multimodal signals makes the interaction more natural and expressive. Evaluations 2 and 3 were significantly more *usable* than evaluation 1 and evaluation 2 was significant even on a tighter level of 0.01.

4.4 Free User Comments

In the evaluation questionnaires, apart from the statements, there was also a text box to fill in any comments that participants may have that were not covered in any of the statements. There were 23 comments in total and they fall into three categories: 9 comments were about speech or sound interaction; 3 were about interaction based on gestures and 3 were about the selection of topics; the remaining were without useful information (thanks, no comment, etc.). As far as speech interaction is concerned, it was commented that it was very laborious and should be faster and more accurate.

Many comments were related to the topic selection. Some noteworthy comments are that it was difficult to identify the topics that Nao mentions and participants wished that they had a list of topics. Another participant said that it looked like a constrained topic. In fact, participants could select another topic, but this was obviously unclear to them. Some of them commented on the sound interaction, i.e. it would be better not to have to wait for a beep before accepting human input (turn-taking), to reduce delay in the interaction, and that they had to speak close to make the robot understand. Regarding gestures, they mentioned that the arm movements came in arbitrary places during the conversation and should be clearer when they occur.

5 Conclusion and Future Prospects

In this paper we have described the evaluation of a robot application, Nao WikiTalk, and presented results comparing the user's expectations and experiences with respect to three different versions of the robot behaviour. The results show that System 2 with most human-like and affordable presentation of information was most highly valued and exceeded the user expectations in two respects: lively appearance and natural gesturing. The current prototype version supports multimodal interaction technology and provides a platform for experimenting with different interaction possibilities. In the future we plan to enhance Nao WikiTalk further with respect to its communicative capabilities, using more expressive and accurately timed gestures, and we will also focus on issues related to topic management and coherence of interaction.

Human-robot interaction is a fast growing and interesting research area, inviting deeper investigations into interaction between humans and intelligent agents. Exciting research topics include multiparty interaction where the robot is one of several interactive participants, and extending interactive situations into virtual environments. This research supports the development of services and applications that can improve daily life by providing more natural access to digital information.

Acknowledgments. We would like to thank the other project members Adam Csapo, Emer Gilmartin, Jonathan Grizou, Frank Han, and Raveesh Meena for their collaboration during the project. We would also like to thank the organisers of eNTERFACE-2012 for providing us with an excellent working environment.

References

1. Norman, D.A.: The Psychology of Everyday Things. Basic Books, New York (1988)
2. Jokinen, K.: Rational communication and affordable natural language interaction for ambient environments. In: Lee, G.G., Mariani, J., Minker, W., Nakamura, S. (eds.) IWSDS 2010. LNCS, vol. 6392, pp. 163–168. Springer, Heidelberg (2010)
3. Jokinen, K., McTear, M.: Spoken Dialogue Systems. Synthesis Lectures on Human Language Technologies, vol. 2(1). Morgan & Claypool (2009)
4. Jokinen, K., Wilcock, G.: Constructive interaction for talking about interesting topics. In: Proceedings of Eigth Language Resources and Evaluation Conference (LREC 2012), Istanbul, pp. 404–410 (2012)
5. 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, pp. 667–672 (2012)
6. Han, J., Campbell, N., Jokinen, K., Wilcock, G.: Integrating 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), pp. 679–683. Kosice (2012)
7. 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), pp. 673–678. Kosice (2012)

8. Kendon, A.: Gesture. Visible Action as Utterance. Cambridge University Press, Cambridge (2005)
9. Jokinen, K.: Pointing gestures and synchronous communication management. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) Second COST 2102. LNCS, vol. 5967, pp. 33–49. Springer, Heidelberg (2010)
10. Jokinen, K., Hurtig, T.: User expectations and real experience on a multimodal interactive system. In: Proceedings of 9th International Conference on Spoken Language Processing (Interspeech 2006), Pittsburgh (2006)
11. 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, Mumbai, pp. 57–69 (2012)