Integration of Gestures and Speech in Human-Robot Interaction

Raveesh Meena^{*}, Kristiina Jokinen,^{**} and Graham Wilcock^{**}

^{*} KTH Royal Institute of Technology, TMH, Stockholm, Sweden

^{**} University of Helsinki, Helsinki, Finland

raveesh@csc.kth.se, kristiina.jokinen@helsinki.fi, graham.wilcock@helsinki.fi

Abstract— We present an approach to enhance the interaction abilities of the Nao humanoid robot by extending its communicative behavior with non-verbal gestures (hand and head movements, and gaze following). A set of non-verbal gestures were identified that Nao could use for enhancing its presentation and turn-management capabilities in conversational interactions. We discuss our approach for modeling and synthesizing gestures on the Nao robot. A scheme for system evaluation that compares the values of users' expectations and actual experiences has been presented. We found that open arm gestures, head movements and gaze following could significantly enhance Nao's ability to be expressive and appear lively, and to engage human users in conversational interactions.

I. INTRODUCTION

Human-human face-to-face conversational interactions involve not just exchange of *verbal* feedback, but also that of *non-verbal* expressions. Conversational partners may use verbal feedback for various activities, such as asking clarification or information questions, giving response to a question, providing new information, expressing the understanding or uncertainty about the new information, or to simply encourage the speaker, through *backchannels* ('ah', 'uhu', 'mhm'), to continue speaking.

Often verbal expressions are accompanied by non-verbal expressions, such as *gestures* (e.g., hand, head and facial movements) and *eye-gaze*. Non-verbal expressions of this kind are not mere artifacts in a conversation, but are intentionally used by the speaker to draw *attention* to certain pieces of information present in the verbal expression. There are some other non-verbal expressions that may function as important signals to manage the dialogue and the information flow in a conversational interaction [1]. Thus, while a speaker employs verbal and non-verbal expressions to convey her communicative intentions appropriately, the listener(s) combine cues from these expressions to *ground* the meaning of the verbal expression and establish a *common ground* [2].

It is desirable for artificial agents, such as the Nao humanoid robot, to be able to understand and exhibit verbal and non-verbal behavior in human-robot conversational interactions. Exhibiting non-verbal expressions would not only add to their ability to draw attention of the users(s) to useful pieces of information, but also make them appear more expressive and intelligible which will help them build social rapport with their users.

In this paper we report our work on enhancing Nao's presentation capabilities by extending its communicative behavior with non-verbal expressions. In section II we briefly discuss some gestures types and their functions in conversational interactions. In section III we identify a set

of gestures that are useful for Nao in the context of this work. In section IV we first discuss the general approach for synthesis of non-verbal expressions in artificial agents and then present our approach. Next, in section V we discuss our scheme for user evaluation of the non-verbal behavior in Nao. In section VI we present the results and discuss our findings. In section VII we discuss possible extensions to this work and report our conclusions.

II. BACKGROUND

Gestures belong to the communicative repertoire that the speakers have at their disposal in order to express meanings and give feedback. According to Kendon, gestures are intentionally communicative actions and they have certain immediately recognizable features which distinguish them from other kind of activity such as postural adjustments or spontaneous hands and arms movement. In addition, he refers to the act of gesturing as *gesticulation*, with a *preparatory* phase in the beginning of the movement, the *stroke*, or the peak structure in the middle, and the *recovery* phase at the end of the movement [1].

Gestures can be classified based on their form (e.g., *iconic*, *symbolic* and *emblems* gestures) or also based on their function. For instance, gesture can complement the speech and single out a certain referent as is the case with typical *deictic* pointing gestures (*that box*). They can also illustrate the speech like *iconic* gestures do, e.g., a speaker may spread her arms while uttering *the box was quite big* to illustrate that the box was really big. Hand gestures could also be used to add rhythm to the speech as the beats. Beats are usually synchronized with the important concepts in the spoken utterance, i.e., they accompany spoken foci (e.g., when uttering *Shakespeare had three children: Susanna and twins Hamnet and Judith*, the beats fall on the names of the children). Gesturing can thus direct the conversational partners' attention to an important aspect of the spoken message without the speaker needing to put their intentions in words.

The gestures that we are particularly interested in this work are Kendon's Open Hand Supine ("palm up") and Open Hand Prone ("palm down"). Gestures in these two families have their own semantic *themes*, which are related to offering and giving vs. stopping and halting, respectively. Gestures in "palm-up" family generally express offering or giving of ideas, and they accompany speech which aims at presenting, explaining, summarizing, etc. [1].

While much of the gestures accompany speech, some gestures may function as important signals that are used to manage the dialogue and the information flow. According to Allwood, some gestures may be classified as having turn-management function. Turn-management involves

TABLE I.
NON-VERBAL GESTURES AND THEIR ROLE IN INTERACTION WITH NAO

<i>Gesture</i>	<i>Function(s)</i>	<i>Placement and the meaning of the gesture</i>
Open Hand Palm Up	Indicating new paragraph	Beginning of a paragraph. The Open Hand Palm Up gestures has the semantic theme of offering information or ideas.
	Discourse structure	
Open Hand Palm Vertical	Indicating new information	Hyperlink in a sentence. The Open Hand Palm Vertical rhythmic up and down movement emphasizes new information (beat gesture).
Head Nod Down	Indicating new information	Hyperlink in a sentence. Slight head nod marks emphasis on pieces of verbal information.
Head Nod Up	Expressing surprise	On being interrupted by the user (through tactile sensors).
	Turn-yielding	End of a sentence where Nao expects the user to provide an explicit response. Speaker gaze at the listener indicates a possibility for listener to grab the conversational floor.
	Discourse structure	
Speaking-to-Listening	Turn-yielding	Listening mode. Nao goes to standing posture from the speaking pose and listens to the user.
Listening-to-Speaking	Turn-accepting	Presentation model. Nao goes to speaking posture from the standing pose to prepare for presenting information to the user.
Open Arms Open Hand Palm Up	Presenting new topic	Beginning of a new topic. The Open Arm Open Hand Palm Up gestures has the semantic theme of offering information or ideas.

turn transitions depending on the interlocutors action with respect to the turn: *turn-accepting* (the speaker takes over the floor), *turn-holding* (the speaker keeps the floor), and *turn-yielding* (the speaker hands over the floor) [3].

It has been established that conversational partners take cues from various source: the intonation of the verbal expression utterance, the phrase boundaries, pauses, and semantic and syntactic context to infer turn transition relevance place. In addition to these verbal cues, eye-gaze shifts is one non-verbal cue that conversational participants employ for turn management in conversation interactions. The speaker is particularly more influential than the other partners in coordination turn changes. It has been shown that if the speaker wants to give the turn, she looks at the listeners, while the listeners tend to look at the current speaker, but turn their gaze away if they do not want to take the turn, If the listeners wants to take the turn the listeners also looks at the speaker, and turn taking is agreed by the mutual gaze. Mutual gaze is usually broken by the listener who takes the turn, and once the planning of the utterance starts, the listener usually looks away, following the typical gaze aversion pattern [3].

III. GESTURES AND NAO

The task of integrating non-verbal gestures in the Nao humanoid robot was part of a project on multimodal conversational interaction with a humanoid robot [4]. We started with WikiTalk [5], a spoken dialogue system for open domain conversation using Wikipedia as a knowledge source. By implementing WikiTalk on the Nao, we greatly extended the robot's interaction capabilities by enabling Nao to talk about an unlimited range of topics. One of the critical aspects of this interaction is that since the user doesn't have access to a computer monitor she is completely unaware of the structure of the article and the hyperlinks present in there which could be a possible sub-topic for the user to continue the conversation. The robot should be able to bring the user attention to these hyperlinks, which we treat as the *new information*. While prosody plays a vital role in making emphasis on content words in this work we aim specifically at achieving the same with non-verbal gestures. In order to make the interaction smooth we

wanted the robot to coordinate turn taking. Here again we were more interested in the turn-management aspect of non-verbal *gestures* and *eye-gaze*. Based on these objectives we set the two primary goals of this work as:

Goal 1: *Extend the speaking Nao with hand gesturing that will enhance its presentation capabilities.*

Goal 2: *Extend Nao's turn-management capabilities using non-verbal gestures.*

Towards the first goal we identified a set of presentation gestures to mark topic, the end of a sentence or a paragraph, beat gestures and head nods to attract attention to hyperlinks (the new information), and head nodding as backchannels. Towards the second goal we put the following scheme in place: Nao will speak and observes the human partner at the same time. After presenting a piece of new information the user is expected to signal interest by making explicit requests or using backchannels. Nao should observe and react to such user responses. After each paragraph the human is invited to signal continuation (verbal command phrases like 'enough', 'continue', 'stop', etc.). Nao asks explicit feedback (may also gesture, stop, etc. depending on previous interaction). Table I provides the summary of the gestures (along with their functions and their placements) that we aimed to integrate in Nao.

IV. APPROACH

A. The choice and timing of non-verbal gestures

Synthesizing non-verbal behavior in artificial agents primarily requires making the choice of right non-verbal behavior to generate and the alignment of that non-verbal behavior to the verbal expression with respect to the temporal, semantic, and discourse related aspects of the dialogue. The content of a spoken utterance, its intonation contour, and the non-verbal expressions accompanying it together express the *communicative intention* of the speaker. The logical choice therefore is to have a composite semantic representation that captures the meanings along these three dimensions. The agent's domain plan and the discourse context play a crucial role in planning the communicative goal (e.g. should the agent provide an answer to a question or seek clarification).

However, an agent requires a model of *attention* (what is currently salient) and *intention* (next dialogue act) for extending the communicative intention with pragmatic factors that determine what intonation contours and gestures are appropriate in its linguistic realization. This includes the *theme* (information that is *grounded*) and the *rheme* (information yet to be grounded) marking of the elements in the composite semantic representation. The realizer should be able to synthesis the correct surface form, the appropriate intonation, and the correct gesture. Text is generated and pitch accents and phrasal melodies are placed on generated text which is then produced by a text to speech synthesizer. The non-verbal synthesizer produces the animated gestures.

As for timing of gestures the information about the duration of intonational phrases is acquired during speech generation and then used to time gesture. This is because gestural domains are observed to be isomorphic with intonational domains. The speaker's hands rise into space with the beginning of the intonational rise at the beginning of an utterance, and the hands fall at the end of the utterance along with the final intonational marking. The most effortful part of the gesture (the "stroke") co-occurs with the pitch accent, or most effortful part of pronunciation. Furthermore, gestures co-occur with the rhematic part of speech, just as we find particular intonational tunes co-occurring with the rhematic part of speech [6].

[6] presents various embodied cognitive agents that exhibit multimodal non-verbal behavior, including hand gestures, facial expressions (eye brow movements, lip movements) and head nods based on the scheme discussion above. In [7] a back projected talking head is presented that exhibits non-verbal facial expression such as lip movement, eyebrow movement, and eye gaze. The timing of these gestures is again motivated from the intonational phrase of the verbal expressions.

B. Integrating non-verbal behavior in Nao

The *preparation*, *stroke*, and *retraction* phases of a gesture may be differentiated by short holding phases surrounding the stroke. It is in the second phase—the stroke—that contains the meaning features that allows one to interpret the gestures. Towards animating gestures in Nao our first step was to define the stroke phase for each gesture type identified in TABLE I. We refer to Nao's full body pose during the stroke phase as the *key pose* that captures the essence of the action. Fig. A to G in TABLE II illustrates the key poses for the set of gestures identified in TABLE I. For example, Fig. A in TABLE II illustrates the key pose for the Open Hand Palm Up gesture.

In our approach we model the preparatory phase of a gesture as comprising of an intermediate gesture, the *preparatory pose*, which is a gesture pose halfway on the transition from the current Nao posture to the target *key pose*. Similarly, the retraction phase is comprised of an intermediate gesture, the *retraction pose*, which is a gesture pose half way on the transition between the target *key pose* and the follow-up gesture. The complete gesture was then synthesized using the *B-spline* algorithm [8] for interpolating the joint positions from the *preparatory pose* to the *key pose* and from the *key pose* to the *retraction pose*.

TABLE II.
NON KEY POSES FOR VARIOUS GESTURES AND HEAD MOVEMENTS.

Fig. A : Open Hand Palm Up	Fig. A1: Side view of Fig. A	Fig. B: Open Hand Palm Vertical	Fig. B1: Side view of Fig. B.

Fig. C: Head Nod Down	Fig. D: Head Nod Up	Fig. E: Listening key pose	Fig. F: Speaking key pose

Fig. G: Open Arms Open Hand Palm Up			

It is critical for the key pose of a gesture to coincide with the pitch accent in the intonational contour of the verbal expression. During trials in the lab we observed that there is always some latency in Nao's motor response. Since gestures can be chained and the preparatory phase of the follow-up gesture unifies with the retraction phase of the previous gesture, considering the Listening key pose (Fig. E TABLE II), the default standing position for Nao, as the starting pose for all gestures, increased the latency, and was often unnatural as well. We therefore specified the Speaking key pose (Fig. F TABLE II) as the default follow-up posture. This approach has the practical relevance of not only reducing the latency but also that the transitions from the Listening key pose to Speaking key pose (presentation mode) and vice versa served the purpose of turn-management. Synthesizing a specific gesture on Nao then basically required an animated movement of joints from any current body pose to the target gestural key pose and the follow-up pose.

As an illustration, the Open Hand Palm Up gesture for paragraph beginning was synthesized as an B-spline interpolation of the following sequence of key poses: *Standing* → *Speaking* → *Open Hand Palm Up preparatory pose* → *Open-Hand Palm Up key pose* → *Open-hand Palm Up retraction pose* → *Speaking*.

Beat gestures, the rhythmic movement of Open Hand Palm Vertical gesture, are different from the other gestures as they are characterized by two phases of movement: a movement into the gesture space, and a movement out of it [6]. In contrast to the pause in the stroke phase of other gestures, it is the rhythm of the beat gestures that is intended to draw the listeners' attention to

the verbal expressions. A beat gestures was synthesized as an B-spline interpolation of *Speaking key pose* $\rightarrow$ *Open Hand Palm Vertical key pose* $\rightarrow$ *Speaking key pose*, with no *Open Hand Palm Vertical preparatory* and *retraction poses*. This sequence of key poses was animated in loops for synthesizing rhythmic beat gestures for drawing attention to a sequence of new information.

We handcrafted the *preparatory*, *key* and *retraction* poses for all the animated gestures using the Choregraphe® (part of Nao’s toolkit). Choregraphe® offers an intuitive way of designing animated actions in Nao and obtained the corresponding C++/Python code. This enabled us to develop a parameterized gesture function library of all the gestures. We could then synthesize a gesture with varying duration of the animation and the amplitude of joint movements. This approach to define gestures as parameterized functions obtained from templates is also used for synthesizing non-verbal behavior in embodied cognitive agents [6] and facial gestures in back projected talking heads [7].

C. Synchronizing Nao gestures with Nao speech

Since much of gestures that we have focused in this work accompany speech we wanted to align the *key pose* of a target gesture with the content words bearing new information. To achieve this we should have extracted the intonational phrases information from Nao’s text-to-speech synthesis system. However, back then, we were unable to obtain the intonational phrase information from Nao’s speech synthesizer. Therefore we took the rather simple approach of finding the average number of words before which the gesture synthesis should be triggered such that the key pose coincides with the content word. This number is calculated based on a gesture’s duration (of the template) and the length of the sentence (word count) to be spoken. Based on these two we approximated (online) the duration parameter of the gesture to be synthesized. In similar fashion we used the punctuations and structural details (new paragraph, sentence end, paragraph end) of a Wikipedia article to time the turn-management gestures. Often, if not always, the timing of these gestures was perceived okay by the developers in lab.

FIGURE 1 provides an overview of Nao’s Multimodal Interaction Manger (MIM). On receiving the user input, Nao Manager instructs the MIM to process the User Input. MIM interacts with the Wikipedia Manager to obtain the content and the structural details of the topic from Wikipedia. MIM instructs the Gesture Manager to use these pieces of information in conjunction with the Discourse Context to specify the gesture type (refers to the

TABLE III.
NON-VERBAL GESTURE CAPABILITIES OF THE MIM INSTANTIATIONS

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)

Gesture Library). Next, the duration parameter of this gesture is calculated (Gesture Timing) and used for placing the gesture tag at the appropriate place in the text to be spoken. While the Nao Text-to-Speech synthesizer produces the verbal expression, the Nao Manager instructs the Nao Movement Controller to synthesize the gesture (Gesture Synthesizer).

V. USER EVALUATION

We evaluated the impact of Nao’s verbal and non-verbal expressions in a conversational interaction with human subjects. Since we wanted to also measure the significance of individual gesture types, we created three versions of Nao’s MIM with each system exhibiting a limited set of non-verbal gestures. TABLE III summarizes the non-verbal gesturing abilities of the three systems.

For evaluation we followed the scheme [9] of comparing users’ expectations before the evaluation with their actual experiences of the system. Under this scheme users were first asked to fill a questionnaire that was designed to measure their expectations from the system. Subjects then took part in three about 10-minute interactions, and after each interaction with the system the users filled in another questionnaire that gauged their experience with the system they had just interacted with.

Both the questionnaire contained 31 statements, which were aimed at seeking users’ expectation and experience feedback on the following aspects of the systems: *Interface*, *Responsiveness*, *Expressiveness*, *Usability* and *Overall Experience*. TABLE IV shows the 14 statements from the two questionnaires that were aimed at evaluating Nao’s non-verbal behavior. The expectation questionnaire served the dual purpose of priming user’s attention to system behaviors that we wanted to evaluate. Participants provided their response on the Likert scale from one to five (with five indicating strong agreement).

Twelve users participated in the evaluation. They were participants of the 8th International Summer Workshop on Multimodal Interfaces, eNTERFACE-2012. Subjects were instructed that Nao can provide them information from Wikipedia and that they can talk to Nao, and play with it as much as they wish. There were no constraints or restrictions on the topics. Users could ask Nao to talk about almost anything. In addition to this, they were provided a list of commands to help them familiarize themselves with the interaction control. All the users interacted with the three systems in the same order: System 1, System 2 and then System 3.

I. RESULTS

The figure in TABLE V presents the values of the expected and observed features for all the test users. The x axis corresponds to the statement id (*S.Id*) in TABLE IV.

FIGURE 1: NAO’S MULTIMODAL INTERACTION MANAGER

TABLE IV.
QUESTIONNAIRES FOR MEASURING USER EXPECTATIONS AND REAL EXPERIENCE WITH NAO.

<i>System aspect</i>	<i>S.Id.</i>	<i>Expectation questionnaire</i>	<i>Experience questionnaire</i>
Interface	I2	I expect to notice if Nao's hand gestures are linked to exploring topics.	I noticed Nao's hand gestures were linked to exploring topic.
	I3	I expect to find Nao's hand and body movement distracting.	Nao's hand and body movement distracted me.
	I4	I expect to find Nao's hand and body movements creating curiosity in me.	Nao's hand and body movements created curiosity in me.
Expressiveness	E1	I expect Nao's behaviour to be expressive.	Nao's behaviour was expressive.
	E2	I expect Nao will appear lively.	Nao appeared lively.
	E3	I expect Nao to nod at suitable times.	Nao nodded at suitable times.
	E5	I expect Nao's gesturing will be natural.	Nao's gesturing was natural.
Responsiveness	E6	I expect Nao's conversations will be engaging	Nao's conversations was engaging
	R6	I expect Nao's presentation will be easy to follow.	Nao's presentation was easy to follow.
	R7	I expect it will be clear that Nao's gesturing and information presentation are linked.	It was clear that Nao's gesturing and information presentation were linked.
Usability	U1	I expect it will be easy to remember the possible topics without visual feedback.	It was easy to remember the possible topics without visual feedback.
Overall	O1	I expect I will like Nao's gesturing.	I liked Nao's gesturing.
	O2	I expect I will like Nao's head movements.	I liked Nao's head movements.
	O3	I expect I will like Nao's head tracking.	I liked Nao's head tracking.

Measuring the significance of these values is part of the ongoing work, therefore we report here just the preliminary observations based on this figure.

Interface: Users expected Nao hand gestures to be linked to exploring topics (I1). They perceived their experience with System 2 to be above their expectations, while System 3 was perceived somewhat closer to what they had expected. As System 1 lacked any hand gestures the expected behavior was hardly observed. Users expected Nao hand and body movement to be distracting (I3). However, the observed values suggest that it wasn't the case with any of the three interactions. Among the three, System 1 was perceived the least distracting which could be due to lack of hand and body movements. Users expected Nao's hand and body movement to cause curiosity (I4). This is in fact true for the observed values for System 2 and 3. Despite the gaze following behavior in System 1 it wasn't able to cause enough curiosity.

Expressiveness: The users expected Nao to be expressive (E1). Among the three systems, the interaction with System 2 was experienced closest to the expectations. System 2 exceeded the users' expectation when it comes to Nao's liveliness (E2). Interaction with System 3 was experienced more lively than interaction with System 1 suggesting that body movements could add significantly to the liveliness of an agent that exhibit only head gestures. Among the three systems, the users found System 2 to meet their expectations about the timeliness of head nods (E3). Concerning the naturalness of the gestures System 2 clearly beats the user expectations while System 3 was perceived okay. Users found all the three interactions very engaging (E6).

Responsiveness: The users expected Nao's presentation to be easy to follow (R6). The gaze following gesture in System 1 was perceived the easiest to follow. System 2 and 3 were able to achieve this only to an extent. As to whether gesturing and information presentation are linked (R7), the interactions with System 2 were perceived closer to the users' expectations.

Usability: Users expected to remember possible topics without visual feedback (U1). For all the three systems, the observed values were close to expected values.

Overall: The Nao gestures in System 1 were observed to meet the users' expectations (O1). The head nods in System 2 were also perceived to meet the users' expectations (O2), and the gaze tracking in System 1 was also liked by the users (O3). The responses to O2 and O3 indicate that the users were able to distinguish head nods from gaze following movements of the Nao head.

In all, the users liked the interaction with System 2 most. This can be attributed to the large variety of non-verbal gestures exhibited by System 2. System 2 and System 3 should benefit by incorporating the gaze following gestures of System 1. Among the hand gestures, open arm gestures were perceived better than beat gestures. We attribute this to the poor synthesis of beat gestures by the Nao motors.

II. DISCUSSION AND CONCLUSIONS

In this work we extended the Nao humanoid robot's presentation capabilities by integrating a set of non-verbal behaviors (hand gestures, head movements and gaze following). We identified a set of gestures that Nao could use for information presentation and turn-management. We discussed our approach to synthesize these gestures on the Nao robot. We presented a scheme for evaluating the system's non-verbal behavior based on the users' expectations and actual experiences. The results suggest that Nao can significantly enhance its expressivity by exhibiting open arms gestures (they serve the function of structuring the discourse), as well as gaze-following and head movements for keeping the users engaged.

Synthesizing sophisticated movements such as beat gestures would require a more elaborate model for gesture placement and smooth yet responsive robot motor actions. In this work we handcrafted the gestures ourselves, using Choregraphe®. We believe other approaches in the field such as use of motion capture devices or Kinect could be

TABLE V.
USER EXPECTATIONS (uExpect'n) AND THEIR EXPERIENCES (ueSys1/2/3) WITH NAO.

used to design more natural gestures. Also we didn't conduct any independent perception studies for the synthesized gestures to gauge how human users perceive the meaning of such gestures in context of speech. Perception studies similar to the one presented in [3], [8] should be useful for us.

We believe the traditional approach of gesture alignment using the phoneme information would have given better gesture timings. We also need a better model for determining the duration and amplitude parameters for the gesture functions. Exploring the range of these parameters in the lines of [10] on exploring the *affect space* for robots to display Emotional Body language would be an interesting direction to follow.

As to whether the users were able to remember the new information conveyed by the emphatic hand gestures has not been verified yet. This requires extensive analysis of the video recordings and has been planned as future work. Moreover, previous research has shown that hand gestures and head movements play a vital role in turn management. We could not verify whether Nao's gestures also served this kind of role in interaction coordination (Goal 2, p.2), but we believe that non-verbal gestures will be well suited for turn-management, especially to be used instead of the default beep sound that Nao robot currently employs to explicitly indicate turn changes. However, our findings suggest that open arm hand gestures, head nods and gaze following can significantly enhance Nao's ability to engage users (Goal 1, p.2), verified by the positive difference between the user's experience and expectations of the Nao's interactive capability.

ACKNOWLEDGMENT

The authors thank the organizers of eNTERFACE 2012 at Supelec, Metz, for the excellent environment for this project.

REFERENCES

- [1] K. Jokinen, "Pointing Gestures and Synchronous Communication Management," in *Development of Multimodal Interfaces: Active Listening and Synchrony*, vol. 5967, A. Esposito, N. Campbell, C. Vogel, A. Hussain and A. Nijholt, Eds., Heidelberg, Springer Berlin Heidelberg, 2010, pp. 33-49.
- [2] H. H. Clark and E. F. Schaefer, "Contributing to Discourse," *Cognitive Science*, pp. 259-294, 1989.
- [3] K. Jokinen, H. Furukawa, M. Nishida and S. Yamamoto, "Gaze and Turn-Taking Behavior in Casual Conversational Interactions," in *ACM Transactions on Interactive Intelligent Systems, Special Issue on Eye Gaze in Intelligent Human-Machine Interaction*, ACM, 2010.
- [4] A. Csapo, E. Gilmartin, J. Grizou, F. Han, R. Meena, D. Anastasiou, K. Jokinen and G. Wilcock, "Multimodal Conversational Interaction with a Humanoid Robot," in *Proceedings of 3rd International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, Slovakia, 2012.
- [5] G. Wilcock, "WikiTalk: A Spoken Wikipedia-based Open-Domain Knowledge Access System," in *Question Answering in Complex Domains (QACD 2012)*, Mumbai, India, 2012.
- [6] J. Cassell, "Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems," MIT Press, 1989.
- [7] S. Al Moubayed, J. Beskow, G. Skantze and B. Granström, "Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction," in *Cognitive Behavioural Systems. Lecture Notes in Computer Science.*, A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann and V. C. Müller, Eds., Springer, 2012.
- [8] A. Beck, A. Hiole, A. Mazel and L. Canamero, "Interpretation of Emotional Body Language Displayed by Robots," in *Proceedings of 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE'10)*, Firenze, Italy, 2010.
- [9] K. Jokinen and T. Hurtig, "User Expectations and Real Experience on a Multimodal Interactive System," in *Proceedings of Interspeech 2006*, Pittsburgh, Pennsylvania, US, 2006.
- [10] A. Beck, L. Canamero and K. A. Bard, "Towards an Affect Space for Robots to Display Emotional Body Language," in *Proceedings of the 19th IEEE International Symposium on Robot and Human Interactive Communication (Ro-MAN 2010)*, Principe di Piemonte - Viareggio, Italy, 2010.