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

Adam Csapo, Emer Gilmartin, Jonathan Grizou, JingGuang Han, Raveesh Meena, Dimitra Anastasiou, Kristiina Jokinen, and Graham Wilcock

***Abstract*—The paper presents a multimodal conversational interaction system for the Nao humanoid robot. The system was developed at the 8th International Summer Workshop on Multimodal Interfaces, Metz, 2012. We implemented WikiTalk, an existing spoken dialogue system for open-domain conversations, on Nao. This greatly extended the robot's interaction capabilities by enabling Nao to talk about an unlimited range of topics. In addition to speech interaction, we developed a wide range of multimodal interactive behaviours by the robot, including face-tracking, nodding, communicative gesturing, proximity detection and tactile interrupts. We made video recordings of user interactions and used questionnaires to evaluate the system. We further extended the robot's capabilities by linking Nao with Kinect.**

***Index Terms*—human-robot interaction, spoken dialogue systems, communicative gesturing.**

I. Introduction

The paper presents a multimodal conversational interaction system for the Nao humanoid robot. The system was developed at the 8th International Summer Workshop on Multimodal Interfaces, Metz, 2012. Our starting point was WikiTalk [1], a spoken dialogue system for open-domain conversations using Wikipedia as a knowledge source. By implementing WikiTalk on the Nao robot, we greatly extended the robot's interaction capabilities by enabling Nao to talk about an unlimited range of topics. In addition to speech interaction, we developed a wide range of multimodal interactive behaviours by the robot, including face-tracking, nodding, communicative gesturing, proximity detection and tactile interrupts. By integrating these mulitimodal behavours we made progress towards more natural communication possibilities between human users and robots.

As the basis for speech interaction, we implemented on Nao a spoken dialogue system, WikiTalk [1], that supports open-domain conversations using Wikipedia as a knowledge source. Earlier work with WikiTalk [2], [3] had used a robotics simulator. This paper describes the multimodal interactive behaviours made possible with a real robot.

The paper is structured as follows. Section II describes the system architecture. Section III explains the communicative gesturing that we developed for Nao, and its integration with speech interaction. Section IV presents a system evaluation based on questionnaires and video recordings of human-robot interactions. Section V describes the use of Kinect with Nao to further extend interaction functionality.

A. Csapo is with Budapest University of Technology and Economics.
E. Gilmartin and J. Han are with Trinity College Dublin.
J. Grizou is with INRIA, Bordeaux.
R. Meena is with KTH, Stockholm.
D. Anastasiou is with University of Bremen.
K. Jokinen and G. Wilcock are with University of Helsinki. e-mail: kristiina.jokinen@helsinki.fi, graham.wilcock@helsinki.fi.

II. System Architecture

An overview of the system architecture is shown in Figure 1. At the heart of the system is a conversation manager, which consists of a finite state machine, and a number of interactive extensions that store various parameters of the user's past interactions and influence the functionality of the state machine accordingly. The conversation manager communicates with a Wikipedia manager on the one hand (so as to be able to obtain appropriately filtered text from Wikipedia), and a Nao manager on the other (so as to be able to map its states onto the actions of the Nao robot).

In order to enable the Nao robot to react to various events while reading text from Wikipedia, the Nao manager is capable of registering events and alerting the appropriate components of the system when anything of interest (either on the inside or the outside of the system) occurs. Figure 1 shows three examples of event handling within the Nao Talk module (the class which implements this module is directly connected to the Nao robot and drives its speech functionality). Functions isSaying(), startOfParagraph(), and endOfSentence() are all called periodically by the Nao manager, and return True whenever the robot stops talking, reaches the start of a paragraph, or finishes a sentence, respectively. Whenever such events occur, the Nao manager can trigger appropriate reactions, for example, through the Gestures module.

A. Interactive extensions within the conversation manager

The history of the user's interactions is stored in a statistics dictionary within the conversation manager. Using a set of simple heuristics, it is possible to create more interesting dialogues between the user and the robot by:

- ensuring that the robot does not give the same instructions to the user in the same way over and over again
- varying the level of sophistication in terms of the functionalities that are introduced to the user by the robot (for example, in the beginning the robot may give simple instructions, allowing the user to practice and understand the basic functionalities of the system; while in the case of more advanced users, the system might suggest new kinds of use cases which may not have previously been known to the user)

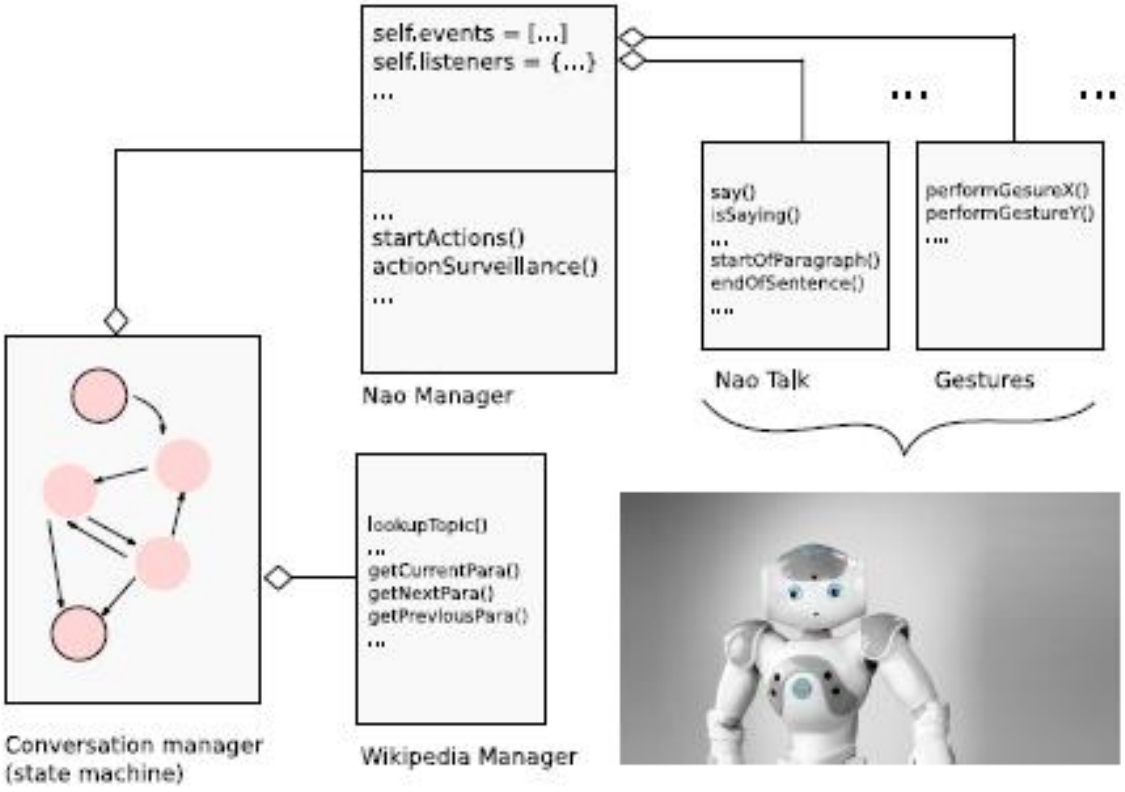

Fig. 1. Overall view of the system architecture.

B. Events and event listeners in the Nao manager

As mentioned earlier, the Nao manager component is capable of registering and listening to events that occur either on the outside of the system, or within the system. Examples of events within the system are currently related to the speech synthesis of the robot, and include:

- The start of new paragraph within the text
- The end of a sentence within the text
- The end of a logically coherent part of the text (for example, the end of a paragraph or a topic)
- The existence of a link within the text

Examples of events outside the system are related to the user's actions, and include:

- The user's proximity to the Nao robot's sonar sensors
- The user touching any of the 3 tactile sensors on the head of the Nao robot

The Nao manager can also be said to include implicit event listeners, which are an integral part of the Nao robot and need not be implemented explicitly by the developer. Examples of event listeners of this type include the Nao robot's capability to detect the presence of the user, track the user's head movements, or recognize the direction of a sound (e.g., when the user claps or makes other noises).

III. Gestures and Nao

In this section we provide an overview of the work related to synthesis of hand gestures and head movements on Nao robot in a multimodal interaction setup. We identify a set of hand gestures and head movements that Nao could use to improve the presentation of the information from Wikipedia. We describe the approach for synthesis of animated gestures on Nao. To evaluate the system, by asking human users to interact with the three versions of the system. We report the preliminary findings on the role of gestures in multimodal interaction with Nao.

A. Motivation

When humans engage in a conversation it is not a mere exchange of spoken words, but also that of other verbal and non-verbal expressions. Speakers use verbal feedback (ah, hmm, uhn, etc.) and non-verbal expressions (gestures and gaze) to convey their communicative intensions. On the other hand listeners take cues from these verbal and non-verbal expressions of the speaker to ground the meaning of the spoken words.

Non-verbal communicative expressions, such as hand gestures, are not mere artifacts in a conversation, but are often intentionally used by the speaker to convey aspects of information that is not explicitly conveyed through the words in an utterance. For example, speakers might use an open arm hand gestures along with the utterance ?the box was quite big? to emphasize that the box was really big. It has been established in the literature that a hand gesture with vertical palm and rhythmic up and down movement is used by the speaker to

Gesture	Purpose	Description
Open hand palm up	Presentation of new paragraph	Open palm The gestures mimics the offering of information to the subject.
Open hand palm vertical	Presentation of new information	A vertical palm, up and down movement to mark new peice of information.
Head nod down	Indicating end of sentence	Upon seeing links in a sentence. To mark new info.
Head nod up	Indicating surprise	On being interrupted.
Speaking to standing	Listening mode	Nao goes to standing pose and listens to speaker.
Standing to speaking	Speaking mode	Nao goes to speaking pose when speaking.

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

emphasize upon certain words or phrases in an utterance. Other non-verbal expressions such as gaze in combination with head movements are used by speakers to manage turn taking during a conversation. Speakers briefly shift their gaze at the listeners to indicate a possible change of turn, thereby allowing the listener to take over the conversational floor. However, a rising pitch towards the end of the utterance would instead suggest that the speaker want to keep the turn. Verbal and non-verbal communicative expressions thus play a vital role in expression of communicative intentions and managing the flow of a conversation.

It is desirable for conversational agent, such as a robot, to be able to understand and exhibit verbal and non-verbal behaviour in human-robot interaction (HRI) scenarios. This would not only add to their ability to express themselves and draw attention of the speakers to useful pieces of information, but make them appear more intelligent and social. While prosody plays a significant role in conveying new information, in this work we focus on the role of non-verbal expressions particularly hand gestures and head movements in conveying new information and managing the flow of conversation.

B. Objectives

In this work we explore the role of non-verbal behaviour in an open domain human robot interaction. We develop conversational abilities in Nao so that it can provide information about any topic of user's choice. On being asked to provide information about a certain topic, the system reads out the Wikipedia article (if it was found). Since the user doesn't have access to a computer monitor on which they could see the content of the article and the various embedded hyperlinks, it becomes essential for the system to somehow bring to the user's attention the presence of these links, which could be used as a new topic for further interaction.

We set off to use hand gestures and head movements for expression of the new information (hyperlinks). We also explore the role of these gestures in presentation of the structure of the article (as in paragraph change, end of sentence, etc.). These gestures have relevance at the discourse level. The list below summarizes the objectives that we set out to demonstrate on Nao:

- Mark/indicate discourse level details which are not explicitly conveyed by spoken words, such as paragraph opening, sentence ending, paragraph end, etc.
- manage the conversation through turn taking gestures (head nod up, head nod down), body poses (speaking to standing position, standing to speaking position) for ? giving the user the opportunity to browse to other parts of the article, and
- Indicate/draw user's attention to topics (the hyperlinks in an article).
- Add expressivity or life to Nao.

In order to achieve these objectives we identify a set of gestures and their purpose in the interaction and information presentation. Table I provides an overview of these gestures.

C. Approach

The communicative intention of any gesture is primarily conveyed by its key poses, which captures the essence of the action. For example, Figure C in Figure I specifies the key pose for open hand palm up gesture. Synthesizing a specific type of hand gestures on Nao basically requires an animated movement of joints from any current body pose to the respective key pose and a follow-up pose. It is the key pose, which coincides with the peak of the animation that conveys the central meaning of the animated action. The context, duration and energy of the movements would affect the interpretation of the gesture conveyed by key poses.

Figures A to G in Figure 2 provide the key poses that we have defined for the purpose of this work. The gestures were synthesized on the fly using the B-spline algorithm for interpolating the joint positions from one key pose to another. For example, the open hand palm up gesture for paragraph beginning was synthesized as an interpolated animation of the following sequence of key poses: *Standing* $\rightarrow$ *Speaking* $\rightarrow$ *Open-hand Palm-up* $\rightarrow$ *Speaking*. In a similar fashion an emphatic beat gesture was synthesized as an interpolated animation of the sequence: *Speaking* $\rightarrow$ *Open-hand Palm-vertical* $\rightarrow$ *Speaking*. The sequence *Open-hand Palm-vertical* $\rightarrow$ *Speaking* could be animated in a loop for synthesizing rhythmic beat gestures for a sequence of new information.

The key poses and the animated gestures were first handcrafted using the Choreograph software. In order to be able to have a control over the dynamics of these handcrafted gestures on the fly we obtained the corresponding python code and defined the gestures as parameterized functions. This way we could control the duration of the animation and the amplitude of joint movements.

D. Synchronizing gestures with speech

The types of gestures that we have focused in this work accompany speech. This requires the synthesis of gestures and speech to be synchronous, moreover, the peak of the gesture should coincide with the pitch accents in the spoken utterance.

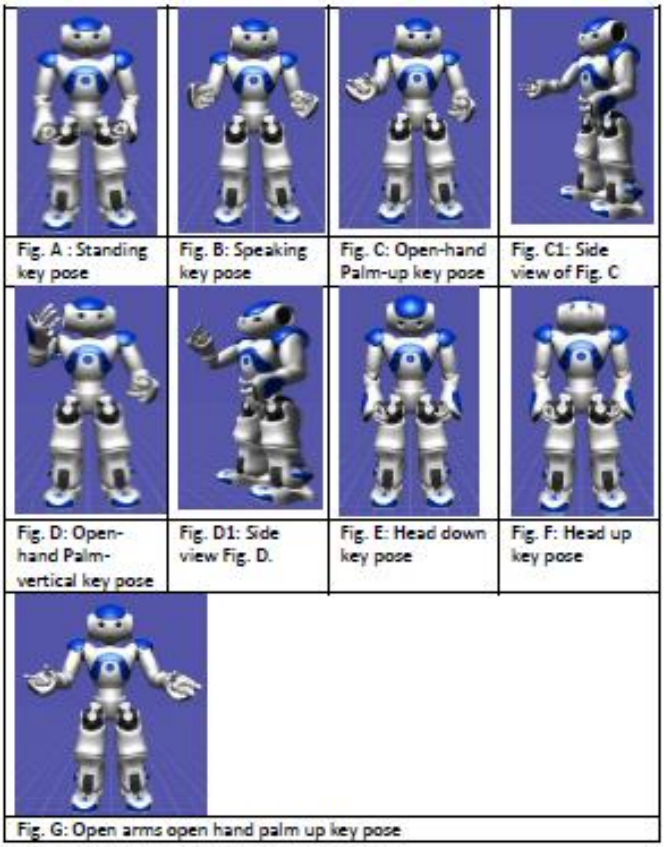

Fig. 2. Key Gestures.

A model for this sophisticated synthesis could not have been explored given the rather short duration of the workshop. Instead we took the approach of synthesizing gestures with rather generic parameters so that they would not be perceived completely out of place.

The gesture planning component (Gesture Manager- GM) in our system identifies the type of gestures to be synthesized. GM uses various contextual details such as the status of discourse, the dialogue context and the contextual information in the article, and identifies the type of gesture to be synthesized. The GM marks up the sentence to be spoken with tags containing information about the type of gesture that is to be triggered. The sentence was then sent to the speech synthesizer and the gesture synthesis components. We basically hoped that the alignment would look natural.

Turn taking: While we had intended to explore the turn taking mechanism in dialogue using gestures and gaze, the inability of Nao speech recognizer to allow barge in refrained us from investigating the role of gestures and gaze in managing turn-taking in a dialogue. Instead the default Nao beeps, which indicate switching on and off of ASR were used a turn taking management.

IV. User evaluation

To evaluate the impact of the various gestures and body movements exhibited by Nao during an interaction, we conducted a user evaluation of the system. Subjects were asked to take part in three about 5-minute interactions. The subjects were told that Nao can provide them information from Wikipedia.

We followed the evaluation scheme proposed in [4]. Users were first asked to fill a questionnaire, which was designed to gauze their expectations from the system. After the interaction with the system the users filled in another questionnaire that gauzed their experience with the system. We evaluated the system along the following dimensions: Interface, Responsiveness, Expressiveness, Usability and Overall experience. Before their first interaction with the system each users fill a questionnaire about their expectations from the system. By doing so we subtly primed the user's attention to aspects of the conversation we wanted to evaluate. After each of the three interactions the users filled another questionnaire regarding their experience. For each question participants were asked to provide their response on a five point scale (where 1: Strongly disagree and 5: Strongly agree). Table II illustrates the questionnaire for evaluating the user expectations and experience on robot gestures and body movements.

Twelve users participated in the evaluation. All of them were participants of the 8th International Summer Workshop on Multimodal Interfaces, eNTERFACE-2012. The subjects were given instructions:

To talk to Nao, and play with it as much as they wish, and try out how well it can present you with interesting information. 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 with the interaction control.

Figure 3 provides an overview of user expectations and their experiences on the questions presented in Table II.

V. Extending Nao with Kinect

The work presented so far is embedded in the Nao robot. The use of Nao's own text-to-speech, speech recognition, sensing and acting capabilities makes the system easy to run from any computer with the Python Naoqi. However we reached some of the limits of the Nao abilities, especially when it comes to detecting behaviours of users interacting with the robot. Gesture recognition, gaze tracking or multiple interlocutors detection are skills beyond the embedded hardware and software of the Nao.

In order to enable more advanced interaction, we started to develop Kinect-based tools that can gather more precise data about the user's behaviour at the cost of an additional external device. Microsoft Kinect is an inexpensive non-invasive technology which by the means of a standard camera associated to a depth sensor is able to determinate the location of particular body joints in a 3D space. This section explains how it could be used to enhance the interaction with the Nao robot.

A. Application

Among the different potential applications of the Kinect in our system, we distinguish three categories : (1) information that helps the robot understand the behaviour of the user

System Aspect	Ref.	Expectation	Experience
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.
Interface	I3	I expect to find Nao's hand and body movement distracting.	Nao's hand and body movement distracted me.
Interface	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
Expressiveness	E2	I expect Nao will appear lively.	Nao appeared lively.
Expressiveness	E3	I expect Nao to nod at suitable times	Nao nodded at suitable times
Expressiveness	E5	I expect Nao's gesturing will be natural.	Nao's gesturing was natural.
Expressiveness	E6	I expect Nao's conversations will be engaging	Nao's conversations was engaging
Responsiveness	R6	I expect Nao's presentation will be easy to follow.	Nao's presentation was easy to follow.
Responsiveness	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	O2	I expect I will like Nao's gesturing.	I liked Nao's gesturing.
Overall	O3	I expect I will like Nao's head movements.	I liked Nao's head movements.

TABLE II
QUESTIONNAIRE FOR EVALUATING USER EXPECTATIONS AND EXPERIENCE WITH NAO.

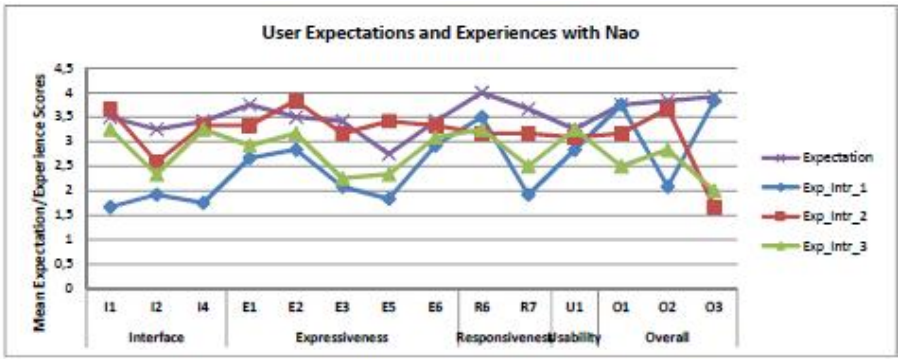

Fig. 3. User expectations and experiences with Nao.

and enhance the interaction, (2) information that helps us evaluating Human-Robot Interaction during user experiments and (3) tools that help us enhancing the behaviour of the robot.

1) Enhancing interaction: The face tracking option provide head orientation and position from which can be extracted an approximation of the gaze of the user. This information can be useful to detect if the user is bored during the interaction and trigger adapted robot behaviours, such as ending the topic, asking for a new topic... The skeleton tracking can be used to detect if a person enter or leave the room as well as the position of it in the room. That could trigger welcome and goodbye behaviour as well as focus the gaze of the robot in the direction of the user. (Note that Nao robots already include a face tracking ability but is limited to close range and proper light interaction, the Kinect is more robust to ambient condition and allow for a larger interaction area.) A gesture recognition module using data from the Kinect [5] would enable non-verbal communication between the human and the robot. In our current set-up, the robot quite often uses confirmation questions that can be boring for a user to verbally reply in the long run. The kind of recognizable gestures we could think of are nodding to say 'Yes' or 'No', arm movement to continue or stop the current topic. We could also use gesture data to focus the robot gaze towards the hands of the user when he performs a gesture. The multiple skeleton and face tracking Kinect abilities can even extend those option in a multi-users setting.

2) Tracking user behaviours: Similar data can be use to track the user behaviour during an interaction in order to get quantitative measurements of the gaze of the user, the user restlessness, the talking position ....

3) Enhancing the behaviour of the robot: Using the Kinect, one could also think of tele-operating the Nao robot, meaning that the gesture of a human standing in front of a Kinect is mapped to the body of the robot. This would decrease the amount of work needed to develop gestures for the robot. Instead of blind trial and error session using a graphical representation of the joint evolution in time, one could directly record a gesture by 'demonstrating' it to the robot. [6] investigates the creation of an affect space for the generation of emotional body language to be displayed by robots, in their case a Nao robot. The body posture where generated by the mean of motion capture data. This work focuses on static posture but can be extended to dynamic gesturing.

Finally, tele-operating the robot would make easier Wizard-of-Oz experiments where the robot gestures are remotely operated by an expert while a user experiment is running.

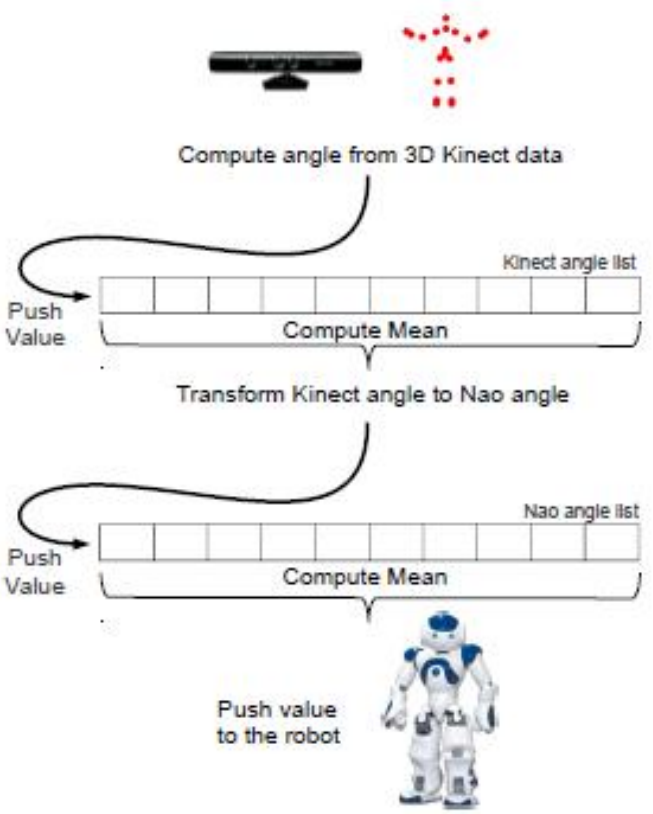

Fig. 4. Double mean filtering of the Kinect data.

B. Teleoperating Nao upper body using Kinect

Microsoft Kinect provides joint positions in a three dimensional space relative to the center of the sensor while the Nao joints are angle based controlled.

In order to teleoperate the robot we need to extract useful angle values from the joint positions as well as to filter out the noise in the data received by the Kinect. In this section we will refer to the name of the joints and angles as respectively referenced in the Kinect SDK and in the Nao documentation.

1) Extracting useful data: In order to map data from the Kinect to the Nao, we need to extract the corresponding angles from the skeleton points gathered though the Kinect. Two aspects have to be considered, (1) the angle measure have to be independent to any other movement of the human and (2) angles should correspond to one degree of freedom of the robot. As gathered data are points in a three dimensional space, we have to choose the plane where points will be projected for the angle measurement.

2) Mapping: Depending on the reference and positive and negative direction, angles extracted from the Kinect data have to be shifted and/or inverted as well as min/max constrained to match with the particular Nao angle reference. This mapping depend on the points chosen and the positive direction defined. It can be done in different ways according to the application. In our case we are using a simple linear mapping from Kinect angle to Nao angle. A non linear mapping could also be used to have more precise movement in certain range.

3) Filtering: Data from Kinect are noisy. In order to get a smooth mapping from human gestures to robot movements, the noise has to be cancelled. Removing noise will add a delay between data acquisition and actual movement on the robot.

As shown in Figure 4, we use two mean filters in a row. For every new data received from the Kinect, angles are computed and are pushed into a list. Then the mean of the value in this list is used to compute the corresponding Nao angle which is pushed into a second list. Finally the mean of this Nao angle list is used to control the robot. The choice of the buffer size is a trade-off between delay in execution and smoothing of the trajectory, in our case the lengths of the Kinect and Nao angle lists have been chosen by empirical tests.

In case empty or incomplete data are receive from the Kinect (person left the room, Kinect obstruction), an empty value is pushed into the Kinect angle list. This value is not taken into account when computing the mean. This simple method allows a smooth and yet reactive filtering.

In addition, we have set the fraction_of_max_speed variable to 0.5. This setting avoids having the robot reaching its current goal before receiving a new one (i.e. avoid shaky movements) and has been evaluated by empirical tests.

VI. Conclusion

The paper presents a multimodal conversational interaction system for the Nao humanoid robot. The system was developed at the 8th International Summer Workshop on Multimodal Interfaces, Metz, 2012. We implemented WikiTalk, an existing spoken dialogue system for open-domain conversations, on Nao. This greatly extended the robot's interaction capabilities by enabling Nao to talk about an unlimited range of topics. In addition to speech interaction, we developed a wide range of multimodal interactive behaviours by the robot, including face-tracking, nodding, communicative gesturing, proximity detection and tactile interrupts. We made video recordings of user interactions and used questionnaires to evaluate the system. We further extended the robot's capabilities by linking Nao with Kinect.

Acknowledgment

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

References

[1] G. Wilcock, "WikiTalk: Wikipedia-based open-domain conversations with robots," submitted.

[2] K. Jokinen and G. Wilcock, "Emergent verbal behaviour in human-robot interaction," in *Proceedings of 2nd International Conference on Cognitive Infocommunications (CogInfoCom 2011)*, Budapest, 2011.

[3] ——, "Constructive interaction for talking about interesting topics," in *Proceedings of Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, 2012.

[4] K. Jokinen and T. Hurtig, "User expectations and real experience on a multimodal interactive system," in *Proceedings of 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Pittsburgh, USA, 2009.

[5] K. Lai, J. Konrad, and P. Ishwar, "A gesture-driven computer interface using Kinect," in *Image Analysis and Interpretation (SSIAI 2012)*, 2012, pp. 185–188.

[6] A. Beck, L. Canamero, and K. Bard, "Towards an affect space for robots to display emotional body language." in *RO-MAN, 2010 IEEE*, 2010, pp. 464–469.