Multimodal Conversational Interaction with a Humanoid 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 Aldebaran Nao humanoid robot. The system was developed at the 8th International Summer Workshop on Multimodal Interfaces, Metz, 2012. Our starting point was a speech-based open-domain knowledge access system. By implementing this system on the robot, we greatly extended Nao's interaction capabilities by enabling the robot 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, to enhance naturalness, expressivity, user-friendliness, and add liveliness to the interaction.

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

Based on the above, the Nao robot with Nao WikiTalk can be regarded as a cognitive robot, since it can reason about how to behave in response to the user's actions. However, in the broader sense, the combination of Nao and WikiTalk is also viewed as a cognitive infocommunication system, as it allows users to interact via the robot with Wikipedia content that is remote and maintained by a wider community.

The paper is structured as follows. Section II explains the multimodal capabilities that we developed for Nao, including communicative gesturing and its integration with speech interaction. Section III describes the system architecture. 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.

II. Multimodal Capabilities

Human face-to-face interaction is multimodal, involving several input and output streams used concurrently to transmit and receive information of various types [3]. While propositional content is transmitted verbally, much additional information can be communicated via non-verbal and paralinguistic audio ('um's and 'ah's in filled pauses, prosodic features), and visual channels (eye-gaze, gesture, posture). These non-verbal signals and cues play a major part in management of turn-taking, communicating speaker and listener affect, and signaling understanding or breakdown in communication.

During interaction speakers and listeners produce bodily movements which, alone or in tandem with other audio and visual information, constitute cues or signals which aid understanding of linguistic information, signal comprehension, or display participants' affective state. Movements include shifts in posture, head movements, and hand or arm movements. We take 'gesture' to include head and hand or arm movements.

A. Gestures

Nao Wikitalk was designed to incorporate head, arm and body movements to approximate gestures used in human conversation. This section describes the motivation for adding gestures to Nao, and their design and synthesis. More comprehensive description of enhancing Nao with gestures and posture shifts can be found in [4].

Gestures take several forms and perform different functions. Following [5], we can distinguish commands and communicative gestures, and the latter can be categorized further as speech-independent (emblems -'ok' sign) or speech dependent (gestures accompanying speech). Speech dependent gestures may be iconic or metaphoric - "the fish was this big" with hands apart to show dimension, a palm-upward 'giving' gesture at start of narration. They may also be deictic (pointing to real or virtual objects) or beat gestures (simple

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.

Gesture	Purpose	Description
Open hand palm up	Presentation of new paragraph	The gestures mimics the offering of information to the subject.
Open hand palm vertical	Presentation of new information	Up and down movement to mark new piece 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

flicks which mark time on speech) [6]. Nods and eye gaze movements are also visual cues to turn-taking management and comprehension in speakers and listeners with listeners nodding feedback, and speakers using upsweeps and gazing at listeners to check understanding and invite contributions/feedback [7].

Nao Wikitalk allows the user to query Wikipedia via the Nao robot and have chosen entries read out by the robot. In a text-free environment the user needs to infer the structure of the article from the robot's output - Wikipedia entries are large blocks of text which can be very monotonous when simply read out by a synthetic voice, and comprehension could be enhanced by adding non-verbal cues to discourse level organization of the text. In Wikipedia relevant information is marked with hyperlinks to other entries. A system where the robot could signal these links non-verbally while reading the text would allow the user to further query the encyclopedia without recourse to explicit menus. Gesture and posture changes could also be used to help manage turntaking in Nao's dialogue, while the inclusion of gesture in Nao's conversational repertoire would also enhance expressivity and add liveliness to the interaction.

As a first step towards adding these functionalities to Nao, we identified a set of gestures which could be used to:

- Mark discourse level details such as paragraph and sentence boundaries.
- Indicate hyperlinks
- Help manage turntaking
- Add expressivity or liveliness

Table I provides an overview of the chosen gesture set.

B. Gesture synthesis

Gestures are performed as a sequence of actions, the most prominent of which is the key pose, which captures the essence of the gesture and conveys much of its communicative payload. The approach taken to gesture synthesis in Nao was to create an animation sequence which could start at any body pose, move to the key pose or action core, and then continue to a follow-up pose which would complete the gesture.

The gesture synthesis process began with the isolation of key poses in the gestures. These key poses were then created in the Nao manually and their parameters recorded using Nao's Choregraphe animation software. The key poses that we have defined for the purpose of this work are shown in Figures A to G in Figure 1. To illustrate, Figure C specifies the key pose for the open hand palm up gesture.

The gestures were then created using Choregraphe's stop motion animation tools to interpolate the position of the robot's joints between the poses comprising the gesture. 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→Speaking→Open-hand Palm-up→Speaking*. In a similar fashion an emphatic beat gesture was synthesized as an interpolated animation of the sequence: *Speaking→Open-hand Palm-vertical→Speaking*. The sequence *Open-hand Palm-vertical→Speaking* could be animated in a loop for synthesizing rhythmic beat gestures for a sequence of new information. The gestures thus created could then be programmed into the robot for later performance.

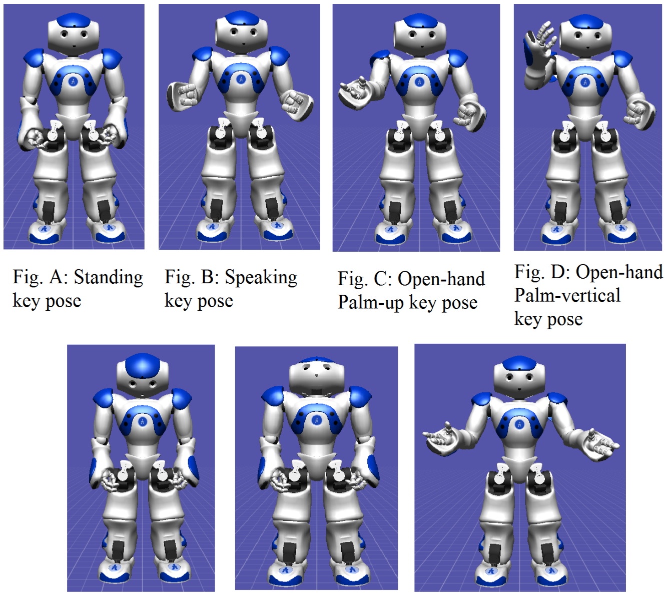
Fig. A: Standing key pose
Fig. B: Speaking key pose
Fig. C: Open-hand Palm-up key pose
Fig. D: Open-hand Palm-vertical key pose
Fig. E: Head down key pose
Fig. F: Head up key pose
Fig. G: Open arms open hand palm up key pose

Fig. 1. Key poses.

During the animation process it became evident that the animation software did not accurately reflect the timing of gestures when performed by the robot rather than onscreen. This reflects the mechanical limitations of the motors of the robot. In order to better control the timing of gestures and to add flexibility to the robot dynamics we obtained the corresponding Python code for each gesture and defined the gestures as parameterized functions. In this way gesture duration and speed could be finely controlled from the Wikitalk code rather than called as monolithic action sequences.

C. Synchronizing gestures with speech

The gesture sequences created for the Nao accompany speech. To create an illusion of coherence requires fine timing control and synchronization of the gesture with the relevant utterance - ideally aligning the gesture peak with the pitch accent of the marked word or phrase. A model for this sophisticated synthesis could not be 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.

In the system, gesture is controlled by a Gesture Manager (GM). The GM first identifies the relevant gesture for the planned utterance, using contextual details such as the status of discourse, the dialogue context and the contextual information in the article. The GM marks up the utterance to be spoken with tags containing information about the type of gesture that is to be triggered. The utterance and accompanying gesture are then created by the speech and the gesture synthesis components and sent to be executed by the robot.

The system currently includes gestures to mark discourse and structural features in the spoken text, and to signal the presence of new information at hyperlinks, both adding liveliness to the dialogue. We had intended to explore the turn taking mechanism in dialogue using gestures and gaze, but the Nao speech recognizer did not allow barge-in, in effect forcing the user to wait for a 'beep' before responding. Therefore, although the presence of a natural upsweep of the head at turn ceding by the Nao worked very well in prompting the user to speak, it was counterproductive in the Nao's current implementation as the user would speak 'before the beep' and thus before ASR had been enabled, confusing rather than enhancing the interaction. It was also noted that the motors were not always fast enough to produce gestures at the precise time indicated. Both of these problems are the result of engineering limitations, and it is highly likely that newer robots will offer improved performance, allowing a fuller range of gesture to be implemented in the system, and improving the timing of currently implemented gestures.

D. Face detection, tactile sensors, and non-verbal cues

As non-verbal information is vital in human face to face interaction, it is desirable for an anthropomorphic embodied conversational agent (ECA) to have facilities to synthesise and recognize non-verbal audio and visual information in addition to its speech synthesis and recognition modules. In this section we summarise the different methods and technologies that we studied for the Nao WikiTalk. The studies and experiments are discussed in more detail in [8].

The Nao platform provides several built-in technologies to enable non-verbal human-robot interaction. Using the Viola-Jones algorithm [9], Nao can detect faces and track people as well as detect the user's head movements like nodding and shakes. However, these capabilities interfere with other modules that send commands to the same motor, e.g. requests to nod, and the head movement appears "jerky" due to conflicting signals. We overcame this problem by deploying conflicting modules into separate threads.

We explored the use of sonar sensors and speech direction detection as conversation triggers. The robot can infer if there are users close by who may want to start a conversation.

Using sonar sensors, we recorded the distance between humans and robots in interactive situations, and could thus empirically test what is the optimal distance for human-robot interactions. In our setup, the best communication distance is about 0.9 meters.

Finally, we investigated different methods for interrupting the conversation, using tactile sensors and an object recognition method. The sensor on Nao's head was adopted as the most reliable method: when the user wants to interrupt Nao's speaking, he or she simply touches the robot on his head.

III. System Architecture

An overview of the system architecture is shown in Figure 2. 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 2 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 structure 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 the functionalities that are introduced to the user by the robot. For example, in the beginning the robot gives simple instructions, allowing the user to practice and understand the basic functionalities of the system; for more advanced users, the system suggests new kinds of use cases which may not have previously been known to the user.

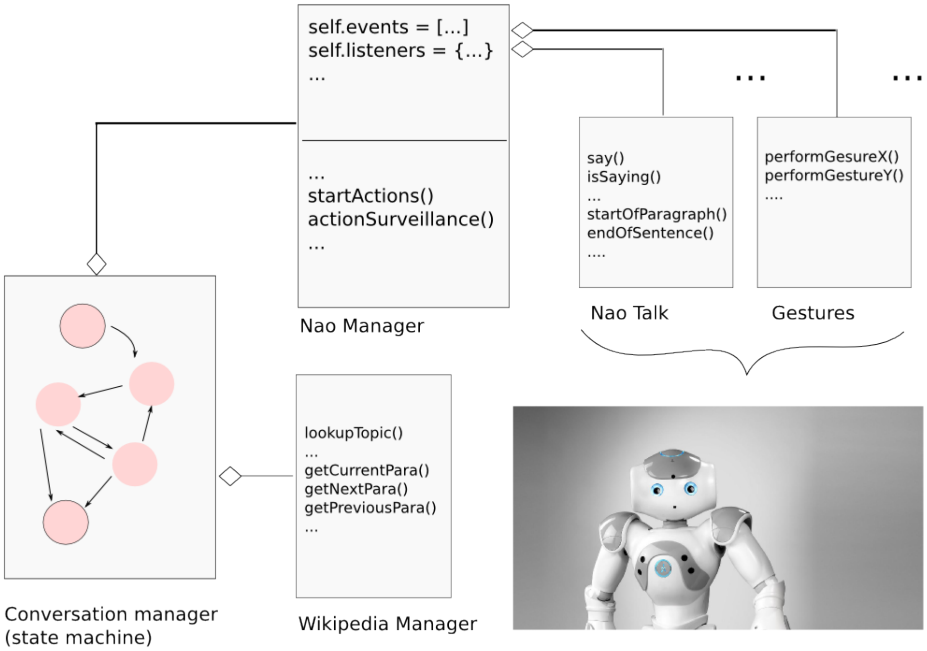

Fig. 2. Overall view of the system architecture.

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. Internal events related to speech synthesis 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

External events related to the user's actions 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).

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 5-minute interactions. The subjects were told that Nao can provide them information from Wikipedia.

We followed the evaluation scheme proposed in [10]. Users were first asked to fill a questionnaire, which was designed to gauge their expectations from the system. After the interaction with the system the users filled in another questionnaire that gauged 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 user filled in 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 in 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 as much as they wish, and try out how well it can present them 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. The user evaluation is discussed in more detail in [11].

V. Extending Nao with Kinect

Using Nao's own speech, sensing and acting capabilities makes the system easy to configure However we reached some of the limits of the Nao abilities, especially when it comes to detecting user behaviours Gesture recognition, gaze tracking

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.

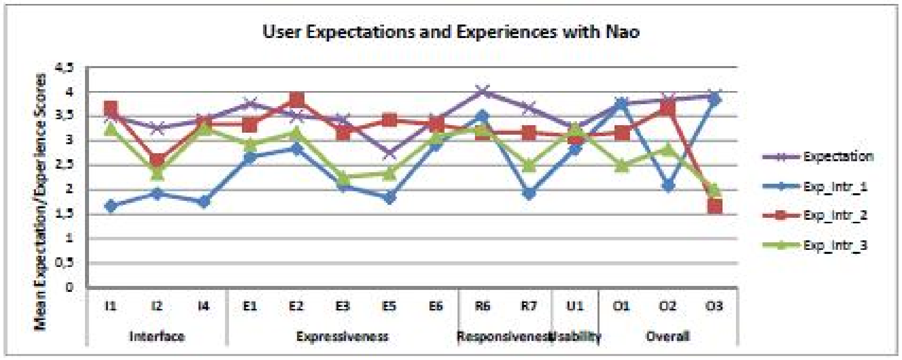

Fig. 3. User expectations and experiences with Nao.

or multiple interlocutors detection are currently 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 means of a standard camera and 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 Kinect in our system, we distinguish three categories : (1) information that helps the robot understand the behaviour of the user and enhance the interaction, (2) information that helps us evaluate human-robot interaction during user experiments and (3) tools that help us enhance 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 enters or leaves the room as well as their position 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 the face tracking ability already included with Nao robots is limited to close range and proper light interaction, the Kinect is more robust to ambient condition and allows for a larger interaction area.) A gesture recognition module using data from the Kinect [12] would enable non-verbal communication between human and 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 they perform a gesture. Kinect's multiple skeleton and face tracking abilities can even extend this to a multi-users setting.

2) Tracking user behaviours: Similar data can be used 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 and so on.

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 sessions using a graphical representation of the joint evolution in time, one could directly record a gesture by 'demonstrating' it to the robot. [13] investigates the creation of an affect space for emotional body language to be displayed by robots. The body postures were generated by means 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.

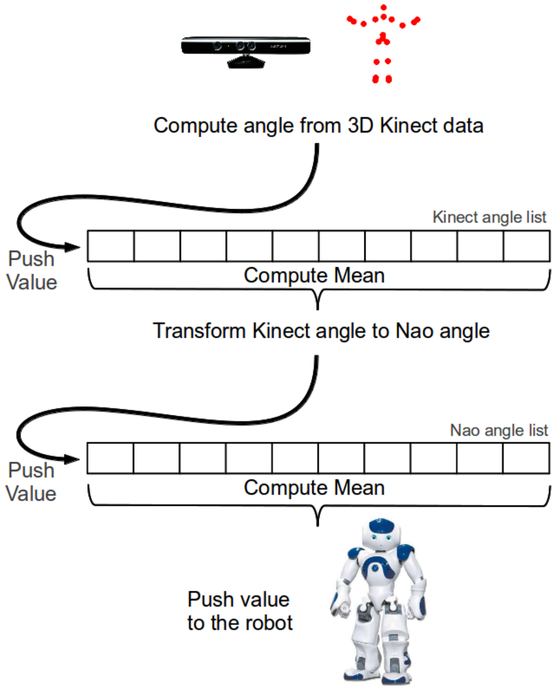

Fig. 4. Double mean filtering of the Kinect data.

B. Teleoperating Nao upper body using Kinect

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.

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 depends on the points chosen and the positive direction defined. In our case we use 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 from the Kinect, angles are computed and pushed into a list. The mean from this list is used to compute the corresponding Nao angle which is pushed into a second list. The mean of this Nao angle list is used to control the robot. The best buffer size was chosen by empirical tests.

If empty or incomplete data are received from the Kinect (person left the room, Kinect obstruction), an empty value is pushed into the Kinect angle list. This simple method allows a smooth and yet reactive filtering. In addition, we set fraction_of_max_speed to 0.5. This avoids the robot reaching its current goal before receiving a new one (i.e. avoid shaky movements) and has been evaluated by empirical tests.

Acknowledgment

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

References

[1] K. Jokinen and G. Wilcock, "Constructive interaction for talking about interesting topics," in *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2010)*, Istanbul, 2012.

[2] G. Wilcock, "WikiTalk: A spoken Wikipedia-based open-domain knowledge access system," in *Question Answering in Complex Domains (QACD 2012)*, Mumbai, India, 2012.

[3] J. Allwood, "Bodily Communication – Dimensions of Expression and Content," in *Multimodality in Language and Speech Systems*, B. Granström, D. House, and I. Karlsson, Eds. Kluwer Academic Publishers, Dordrecht, 2002, pp. 7–26.

[4] R. Meena, K. Jokinen, and G. Wilcock, "Integration of gestures and speech in human-robot interaction," in *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, 2012.

[5] F. Quek, "Toward a vision-based hand gesture interface," in *Proceedings of the Virtual Reality System Technology Conference*, Singapore, 1994, pp. 17–29.

[6] A. Kendon, *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[7] C. Navarretta, E. Ahlsén, J. Allwood, K. Jokinen, and P. Paggio, "Feedback in Nordic first-encounters: a comparative study," in *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2010)*, Istanbul, 2012.

[8] J. Han, N. Campbell, K. Jokinen, and G. Wilcock, "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.

[9] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[10] K. Jokinen and T. Hurtig, "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.

[11] D. Anastasiou, K. Jokinen, and G. Wilcock, "Evaluation of WikiTalk - user studies of human-robot interaction," Proceedings of 15th International Conference on Human-Computer Interaction (HCII 2013).

[12] 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.

[13] 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.