Reactive Gaze Control for Natural Human-Robot Interactions

Yasser Mohammad and Toyoaki Nishida
Graduate School of Informatics
Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
yasser@ii.ist.i.kyoto-u.ac.jp, nishida@i.kyoto-u.ac.jp

***Abstract*—Nonverbal behavior during human-human close encounters is critical to the accomplishment of natural interaction. For this reason, humanoid robots trying to achieve natural interactions with humans should be able to understand and synthesis nonverbal behavior in a way that mimics the human use of it. One of the most important situations during natural human-robot interactions is the explanation scenario in which the human is explaining a task to the robot using natural verbal and nonverbal behavior. This situation occurs frequently in many HRI applications and is critical to the success of the Robots as Knowledge Media project suggested by the authors. In this paper the implementation of a humanoid robot that can show human like gaze control during explanation settings based only on reactive processing is presented. The software of the robot is based on the EICA architecture designed to combine autonomy with interactivity in the lowest level of the system. The details of the implementation and analysis of the naturalness of behavior and the effect of noisy input is presented in this paper.**

***Index Terms*—Gaze Control, Reactive HRI**

I. INTRODUCTION

The robot envisioned in this work is a knowledge transfer agent that can *listen* to expert explanations of a given task and then *conveys* this knowledge to novice humans when they have a need for such knowledge transfer ([9]). More details can be found in [5], [7].

To achieve this goal, the robot needs to give the expert a natural explanation experience during knowledge acquisition. In this paper a minimalist implementation of such a natural listener is presented in the knowledge blind case (see [7] for details) in which the robot do not understand the verbal content of the explanation.

In [3] Kidd and Breazeal compared a robot and a computer-graphic agent and found that subjects felt the robot to be more informative and credible than the computer-graphic agent for communication concerning real-world objects like manipulating color objects on a table. Many other experiments showed the effectiveness of utilizing facial expression, gaze and gestures for communication purposes [8]. Mutual Attention was discussed by many researchers as a mutual body movement essential for natural communication in HRI [13], [14] and [9].

Some researchers focused on natural listening behavior in human robot interactions. In [11] Ogawa and Watanabe studied the concept of speech driven embodied robots using two robots that play the role of the listener and the speaker. Two models were incorporated: one is a listener's action model in which nodding, blinking and the motions of head, arms and body are estimated by the hierarchical moving average (MA) model of the burst-pause (ON-OFF) of speech to the nodding; the other is a speaker's action model in which the motions of head, arms and body are estimated by its own MA model of the ON-OFF of speech to the head motion. By the sensory evaluation and behavioral analysis in human-robot interaction, the effectiveness of this InterRobot's interaction was demonstrated. Although [11] focused on the temporal aspect of natural listening, [2] focused on both temporal and spatial cooperativeness in a route guiding scenario using communication units and customizable firing rules for them.

In [10] a Bayesian network was employed to implement natural listening behavior in a demonstration scenario similar to the explanation scenario demonstrated in this paper. The authors designed four communication modes for the speaker and used nonverbal cues to decide the current communication mode. This decision is then employed to guide the nonverbal behavior of the robot in a probabilistic framework. In this paper a similar behavior is investigated that utilizes very simple state machines that depend directly on the row sensor information and combined using the EICA architecture to generate complex natural behavior comparable to the reported human-human natural listening in close encounters.

The following section briefly explains the EICA architecture used to develop the software of the robot followed by a detailed explanation of the current implementation in section III. In section IV preliminary results showing the applicability of the proposed approach are presented then the paper is concluded.

II. EICA ARCHITECTURE

The software of the robot presented in this paper is built using the EICA architecture proposed by the authors to combine autonomy with natural interactivity. The theoretical foundations of the EICA (Embodied Interactive Control Architecture) can be found in [4] and [6]. Only a brief account of this foundation and implementation details of the architecture are given here followed by a discussion of its features that make it suitable for implementing natural listening behavior. As shown in Fig. 1, the EICA system consists of a set of type specifications that can be realized either by software objects or hardware dedicated circuits, processors, or microcontrollers. The set of customizable active components are:

978–1–4244–1676–9/08/$25.00 ©

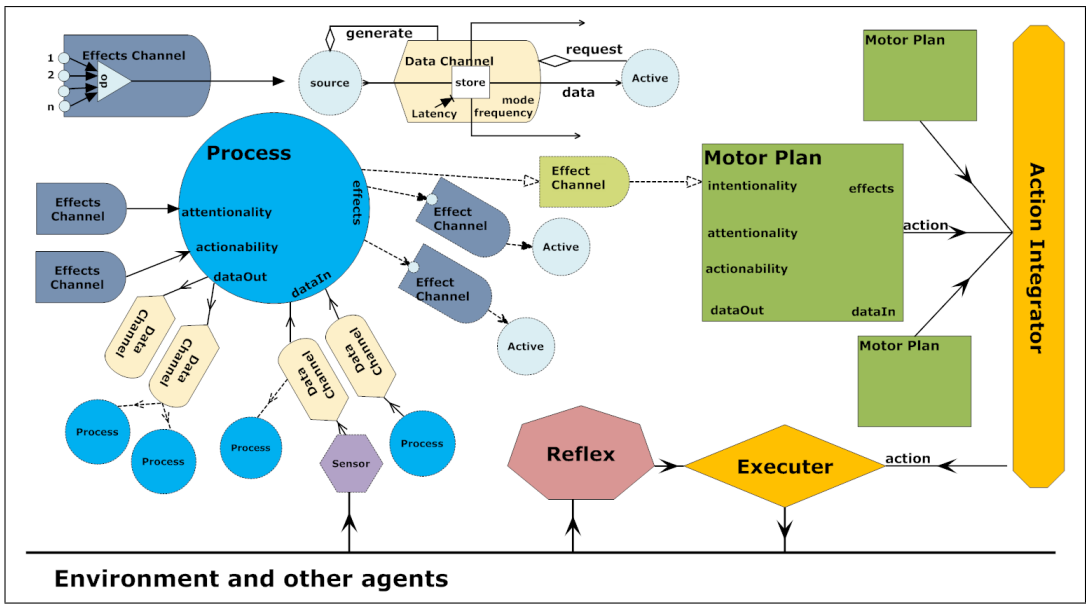

Fig. 1. L_0EICA

- *Processes*: Active components that run continuously. Every process has an *attentionality* attribute that determines how frequent it checks its inputs. Processes can create, drop, communicate, and change the attentionality of other processes or motor plans. Processes can also manage the intentionality of registered motor plans. Normal processes cannot generate actions that affect the robot's behavior directly.
- *Reflexes*: Special type of processes that have fixed maximum attentionality and that can produce actions that go directly to the actuators overriding any bending actions in the action integrator. Note that only reflexes and motor plans can generate actions in EICA
- *Motor Plans*: Reactive processes that along with *attentionality* have an *intentionality* attribute that determine the relative priority of their generated actions.
- *Executers*: Active components that can execute actions directly on the robot.

Along with those customizable action components, EICA has a single process called the *Action Integrator* that combines the actions generated by active motor plans and reflexes to generate the final behavior of the robot. The key features of L_0EICA that makes it appropriate for implementing natural listening behavior are:

1) Sound intention modeling: EICA is based on the interactive model of intention and intention communication proposed by the authors in [4]. This model of intention is based on sound theoretical foundations from neuroscience and experimental psychology [4] that facilitates implementing natural interactive behavior and facilitates debugging and incremental adding of behaviors.
2) Flexible action selection\integration: EICA utilizes a two-levels action integration mechanism that allows it to generate behaviors ranging from Behavior Level Selection to Action Level Combination based on the attributes given to the intentions and actions which are under the control of the developer.
3) Flexible relation between deliberation and reaction: L_0EICA imposes no restrictions on the relation between deliberation and reaction so the robot design can be built in a pure reactive mode and then deliberation can be added as the task complexity supported increases. This is useful for implementing natural listening as the final behavior will not depend only on the nonverbal reactive behavior that is the focus of this paper but also on deliberative behavior based on natural language and context processing.

III. Reactive Gaze Control

Natural listening is a complex process that requires both reactive and deliberative processes. In the current paper only the nonverbal reactive behavior is considered assuming that the robot has no access to the meaning of the explanation or the contents of the environment in which it is immersed.

As a proof of concept a simulation study of the listening behavior of the robot using a minimalist approach was employed. The robot simulated in this study is a Robovie II humanoid robot with 11 degrees of freedom (3 for the head and 4 for every arm). The goal of this simulation was to check the applicability of the EICA architecture in this domain, and to realize a behavior pattern that can be objectively compared with the human-human known behavior in close encounters. As test behaviors mutual gaze, and gaze toward instructor were chosen because of the availability of objective human-human data [1],[9], their relative simplicity, and the important role in natural interactions [14]. Mutual attention behavior of the simulated robot was also studied although human-human

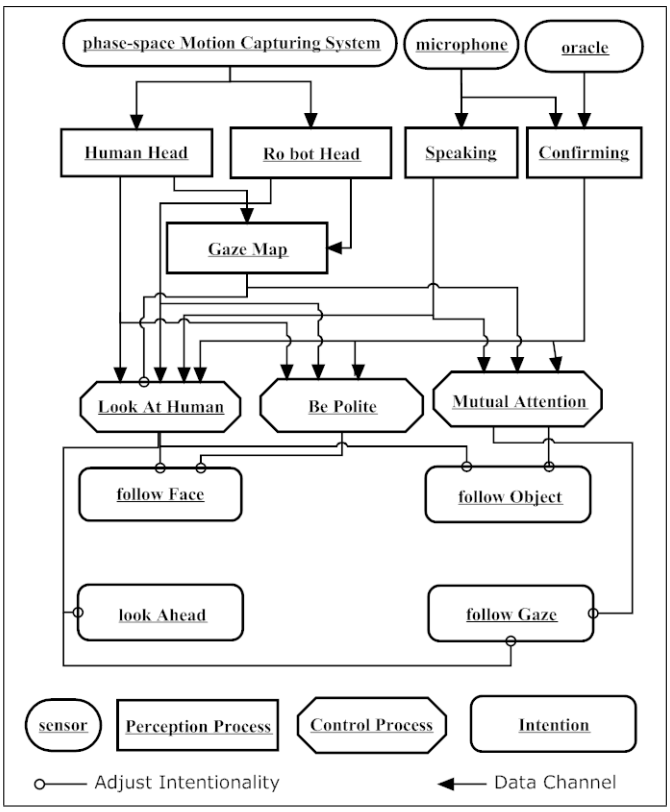

Fig. 2. EICA Processes and motor plans for the Reactive Gaze Controller

behavior data was not available due to the important role of mutual attention in natural listening.

As a minimal design, only the head of the virtual robot was controlled during the simulation. This decision was based on the hypothesis accepted by many researchers in the nonverbal human interaction community that gaze direction is one of the most important nonverbal behaviors involved in realizing mutual attention [1]. The processes and motor plans involved in implementing natural listening behavior in this simulation are shown in Fig. 2.

The analysis of the chosen behavior requirements showed the need of three processes. Two processes to generate an approach-escape mechanism controlling looking toward the human operator which is inspired by the Approach-avoidance mechanism suggested in [1] in managing spatial distance in natural human-human situations. Those processes were named Look-At-Human, and Be-Polite. A third process was needed to control the realization of the mutual attention behavior. This process was called Mutual-Attention. A fourth setup process that generates all the other processes and motor plans was used for technical reasons but did not affect the results.

Four reactive motor plans were designed that encapsulate the possible interactive actions that the robot can generate, namely, looking around, following the human face, following the salient object in the environment, and looking at the same place the human is looking at. Each of which is implemented as a simple state machine

The algorithms of the three processes controlling the simulated robot's behavior are shown in Algorithms 1,2,3 and can be explained as follows:

Algorithm 1 Look-At-Human

procedure LAH ▷ attracts the robot to the human
  loop
    $ang \leftarrow \angle(H_{head}, R_{head})$
    if $w_{r_n} = \max_i(w_{r_i}) < \lambda_w$ AND $t_{human-speaking} > \tau_s$ **then**
      if $confirming$ **then**
        $I_{ff} \leftarrow I_{ff} + \varepsilon_{ff}\mu_c\|ang\|$
      else
        $I_{ff} \leftarrow I_{ff} + \varepsilon ff$
      end if
      $I_{fo} = I_{fo} - \varepsilon_{fo}\|ang\|$
      $I_{fg} = I_{fg} - \varepsilon_{fg}\|ang\|$
      $I_{la} = I_{la} - \varepsilon_{la}\|ang\|$
    end if
  end loop
end procedure

Algorithm 2 Be Polite

procedure BP ▷ repulses the robot from the human
  loop
    $ang \leftarrow \angle(H_{head}, R_{head})$
    if $t_{looking-at-human} > \tau_p$ **then**
      if $confirming$ **then**
        $I_{ff} \leftarrow I_{ff} - \breve{\varepsilon}_{ff}\|ang\|$
      else
        $I_{ff} \leftarrow I_{ff} - \breve{\varepsilon}_{ff}\mu_c\|ang\|$
      end if
    end if
  end loop
end procedure

Algorithm 3 Mutual Attention

procedure MI▷ attracts the robot to the direction of gaze
  loop
    if $\arg\max_i(w_i) = \arg\max_i(w_{r_i})$ AND $t_{human-speaking} > \tau_m a$ **then**
      if $confirming$ **then**
        $I_{fg} \leftarrow I_{fg} + \widehat{\varepsilon}_{fg}$
      else
        $I_{fo} \leftarrow I_{fo} - \widehat{\varepsilon}_{fo}$
      end if
    end if
  end loop
end procedure

1) *Look-At-Human*: This process is responsible of generating an attractive virtual force that pulls the robot's head

direction to the location of the human face. This process first checks the Gaze-Map's current modes and if their weights are less than a specific threshold for more than 10 seconds, and if the human is speaking for more than 4 seconds , it increases the intentionality of the *followFace* motor plan and decreases the intentionality of the other three reactive motor plans based on the difference in angle between the line of sight of the human and the robot and the *Confirming* condition (if the human is confirming the robot should look at him more as was noted in [14]).

2) *Be-Polite*: This process works against the *Look-At-Human* process by decreasing the intentionality of the *followFace* motor plan in reverse proportion to the angle between the line of sight of the human and the robot depending on the period the human is speaking.
3) *Mutual-Attention*: This process increases the intentionality of the *followObject* or the intentionality of the *followGaze*. The rate of intentionality increase is determined based on the confirmation mode.

Five perception processes were needed to implement the aforementioned control processes and motor plans:

1) *Human-Head*, which continuously updates a list containing the position and direction of the human head during the last 30 seconds sampled 50 times per second.
2) *Robot-Head*, which continuously updates a list containing the position and direction of the robot head during the last 30 seconds sampled 50 times per second.
3) *Gaze-Map*, which continuously updates a representation of the distribution of the human gaze both in the spatial and temporal directions. The spatial distribution is stored as a mixture-of-Gaussians like structure where the mean μ_i represents the location of an important object and the variance σ_i is a measure of the size of that object. The weight w_i represents the importance of the place according to the gaze of the human. Those weights have a self decay parameter β to focus the representation on the currently important parts of the space. The temporal evolution of the gaze direction is captured in the weights w_{r_i}. Algorithm 4 shows the exact algorithm of this process.
4) *Speaking*, uses the power of the sound signal to detect the existence of human speech. The current implementation simply assumes there is a human speech whenever the sound signal is not zero. This was acceptable in the simulation but with real world data a more complex algorithm that utilizes Fourier analysis will be used.
5) *Confirming*, specifies whether or not the human is making a confirming action. Currently an oracle is used to do this operation although the algorithm proposed in [14] can be used.

The evaluation data was collected as follows:

1) Six different explanation scenarios were collected in which a person is explaining the procedure of operating a hypothetical machine that involves pressing three different buttons, rotating a knob, and noticing results in an LCD screen in front of a Robovie II robot while pretending that the robot is listening to the explanation. The data was collected using the PhaseSpace Motion Digitizer system [12] by utilizing 28 LED markers attached to various parts of the environment and the explainer as follows:
   - 8 markers attached around the head of the explainer
   - 4 markers attached around every wrest of the explainer
   - 2 markers attached to both sides of the explainer's palm
   - 1 marker attached to the index finger
   - 1 marker in the location of every button (total 3 markers) and one marker in the location of the knob, another in the location of the LCD and sixth marker between the LCD and the knob locations.

   The data was logged 460 times per second.
2) The logged data were used as the input to the robot simulator and the behavior of the robot's head was analyzed.
3) For every scenario 20 new synthetic scenarios were generated by utilizing 20 different levels of noise. The error level is defined as the percentage of the mean value of the noise term to the mean of the raw signal. The behavior of the simulator was analyzed for every one of the resulting 120 scenarios and compared to the original performance.

It should be noted that in the current setup of the experiment the speaker's behavior cannot be assumed completely normal because the robot is not actually executing natural listening behavior which breaks the interaction loop. This is the reason that in the current proof of concept experiment only the means of the time of execution of various interaction behaviors are compared to the known human-human means. It was assumed that the effect of the inaccuracies resulted from not using a human in the listener position will have smaller effect on the means compared to its effects on the details of the behavior. In the full scale experiment the listening robot will be actually *listening* to the instructor using the software presented in this paper and the interaction loop will be closed which will make it possible to compare the details of the robot's behavior to the human-human case.

Fig. 3 shows the evolution of intentionality of the aforementioned four basic reactive motor plans under the control of the three control processes used to implement natural interaction in one case.

In the beginning the robot was scanning the environment for salient features that require attention. The interaction with the human started when the human directed his gaze to the robot for a few seconds. The *Look-At-Human* process increased the intentionality of the *followFace* motor plan while decreasing the intentionality of the *LookAround* motor plan which initialized the eye contact that started the interaction. After a while (25 seconds in average) the *Be-Polite* process takes over reducing the intentionality of the *followFace* which along with

Algorithm 4 Gaze-Map Process

procedure GAZE-MAP ▷ Builds a gaze map in the spatial and temporal dimensions

$G = \phi, i = 0$

loop

$g(i) \leftarrow$ point of current humans gaze focus

if $\left\| g(i) - \sum_{j=0}^{i-1} (j+1)\, g(j) / \sum_{j=0}^{i-1} (j+1) \right\| < \varepsilon \text{for} \nu \text{steps}$ **then**

$H_g \leftarrow \frac{1}{i+1} \sum_{j=0}^{i} g(j)$, $\rho(H_g) \leftarrow \sum_{k=0}^{m-1} \frac{w_k}{(2\pi)^{1.5}\sigma_k} \exp\left(-\frac{1}{2}\sigma_k^2 (H_g - \overrightarrow{\mu_k})^T I_3 (H_g - \overrightarrow{\mu_k})\right)$

if $\rho(H_g) < \delta_0 \sum w_k$ **then**

$G \leftarrow G \cap \left\{ w_0, \mathrm{H}_g, \sigma_0, \max_i (w_{r_i}) + 0.5, now \right\}$

else

$n \leftarrow \arg\max_k (\|\mu_k - H_g\|)$

$w_n \leftarrow w_n + \eta(\|\mu_k - H_g\|)$, $w_{r_n} \leftarrow \max_i (w_{r_i}) + 0.5$, $t_n \leftarrow now$

if $\|\mu_n - H_g\| < d_0$ **then** $\sigma_n \leftarrow \sigma_n + \eta_\sigma(\|\mu_k - H_g\|)$

else $\mu_n = \mu_n + \eta_\mu(\|\mu_k - H_g\|)$

end if

end if

else

$i \leftarrow 0$

end if

for all tuples in G **do**

$w_i \leftarrow w_i(1 - \beta)$

end for

end loop

end procedure

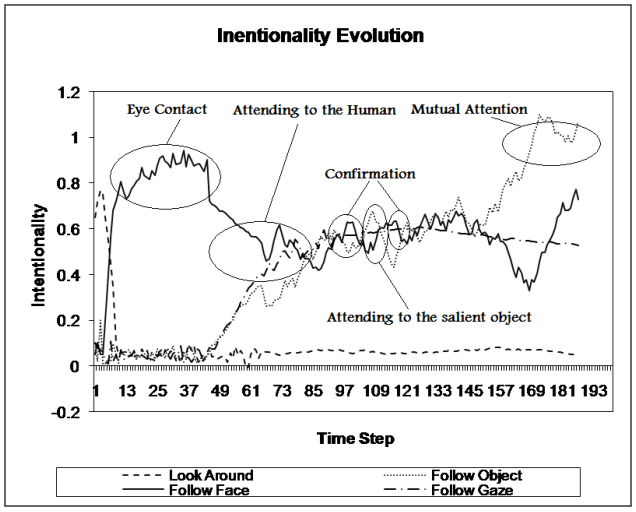

Fig. 3. The Evolution of Intentionality of the four Basic Motor Plans Implementing Natural Listening Behavior

the increased intentionality of *followGaze* and *followObject* results in breaking the eye contact. As the interaction goes the robot will tend to look to the human for around 78.87% of the time while attending to the shared objects of interest around 53.12% if the time.

IV. Results and Discussion

TABLE I
COMPARISON BETWEEN THE SIMULATED AND NATURAL BEHAVIOR

Item	Statistic	Simulation	H-H value
Mutual Gaze	Mean	31.5%	30%
	Std.Dev.	1.94%	–
Gaze Toward Instructor	Mean	77.87%	75%
	Std.Dev.	3.04%	–
Mutual Attention	Mean	53.12%	unknown
	Std.Dev.	4.66%	–

To analyze the applicability of EICA to the natural listening behavior an objective evaluation criteria was selected. For the three behaviors chosen the mean and standard deviation of the time spent doing each of them were calculated and compared with the known values in natural human-human interactions when available.

Some of the results of numerical simulations of the listening behavior of the robot are given in Table I. The table shows the average value obtained from the simulated robot in comparison to the known values measured in human-human interaction situations as stated in [1]. As the table shows the behavior of the robot is similar to the known average behavior in the human-human case for both mutual gaze and gaze toward instructor behaviors and the standard deviation in both cases is less than 7% of the mean value which predicts robust operation in real world situations. These results suggest that the proposed approach is at least applicable to implement natural listening behavior. It should be noted that the naturalness of the behavior does not only depend on the averages specified but on the detailed movements of the head and eye during the interaction which will be measured in the final experiment with the Robovie II robot. This simulation results, nevertheless, can be considered a first step proof of applicability given that the reported simulation statistics were not hard coded in any of the processes controlling the robot.

Fig. 5 shows the effect of increasing the error level from zero to 100% of the input on the percentage of time mutual gaze, gaze toward instructor, and mutual attention behaviors were recognized in the simulation. As expected the amount of time spent on these interactive behaviors decreases with increased error level although this decrease is not linear but can be well approximated with a quadratic function.

To study this phenomenon further, an error term is defined as: *Error in behavior X* is the difference in the mean time spent in doing behavior X in the simulation and the reported human-human mean relative to the reported human-human value.

Analysis of the effect of noise on the behavior of the robot showed an advantage of using EICA in terms of the robustness in the resulting behavior. In both mutual gaze and gaze toward instructor the difference in the mean between

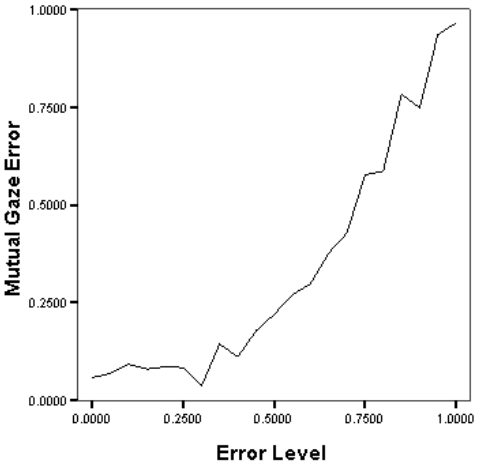

Fig. 6. Effect on the error level on the difference between mutual gaze mean and the human-human case

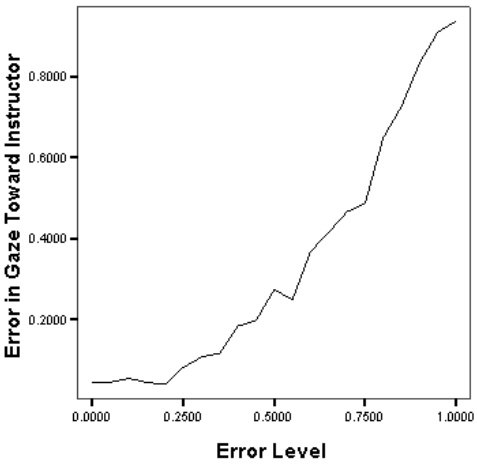

Fig. 7. Effect on the error level on the difference between mean time of gazing toward the instructor and the human-human case

the simulated behavior and the natural human-human reported values is growing much slower than the error in the input to the system. Regression Analysis revealed that in both cases the effect on the mean time spent doing the studied behavior grows with the square of the inverse SNR (Signal to Noise ratio) as Fig. 8 shows for the mutual gaze behavior. As Fig. 6 and Fig. 7 suggests, the current implementation has excellent noise rejection properties as long as the noise is less than 25% of the original signal.

V. Conclusion

This paper reports a simulation study to test the applicability of the EICA architecture to implement simplified natural nonverbal listening behavior in humanoid robots. The results

(A) Mutual Attention

(B) Mutual Gaze

Fig. 4. A synchronized view of the human instructor and the listener robot showing two interesting interactive behaviors. The behavior of the human was recorded using the motion capture data and applied offline to the gaze controller proposed in this paper, the motion commands was then applied to the Robovie to visualize the interactive behavior.

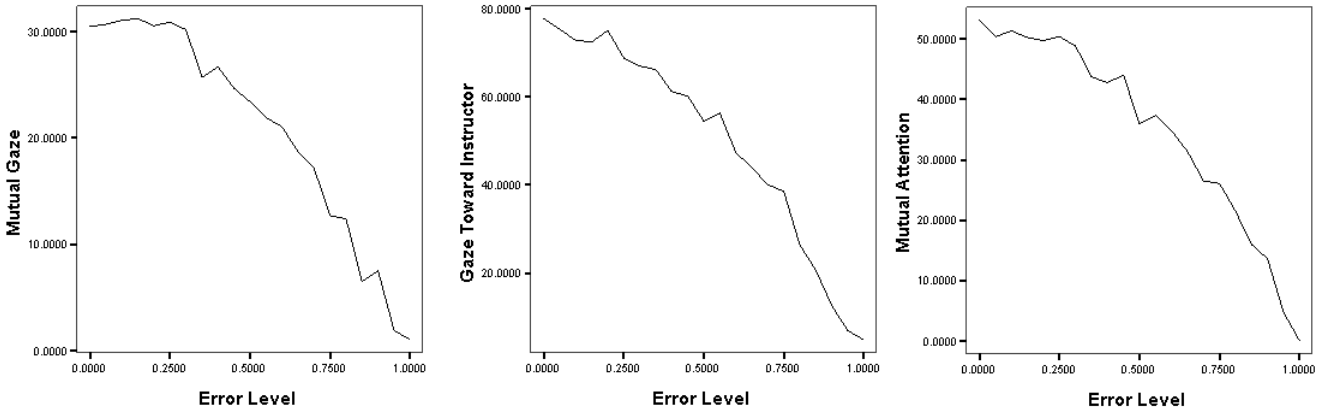

Fig. 5. Effect on the error level on the behavior of the robot

of this study suggest that the interaction between very simple EICA processes can achieve a behavior comparable to the human natural listening behavior. Moreover the study of the effect of noise level on the final robot behavior revealed that the EICA based implementation has good noise rejection properties. In the future a user study with an actual implementation of this system on the Robovie II humanoid robot will be conducted to test the subjective evaluation of the listening behavior of the robot and compare this evaluation with previous work in this area.

ACKNOWLEDGMENT

The first author would like to thank the Egyptian Government for supporting his PhD study during which this work took place.

REFERENCES

[1] M. Argyle, *Bodily Communication*. Routledge; New Ed edition, 2001.
[2] T. Kanda, M. Kamasima, M. Imai, T. Ono, D. Sakamoto, H. Ishiguro, and Y. Anzai, "A humanoid robot that pretends to listen to route guidance from a human," *Autonomous Robots*, vol. 22, no. 1, pp. 87–100, 2007.
[3] C. D. Kidds and C. Breazeal, "Effect of a robot on user perceptions," in *IEEE/RSJ Conference on Intelligent Robots and Systems 2004 (IROS 2004)*, vol. 4. IEEE, September 2004, pp. 3559–3564.
[4] Y. F. O. Mohammad and T. Nishida, "A new, hri inspired, view of intention," in *AAAI-07 Workshop on Human Implications of Human-Robot Interactions*, July, pp. 21–27.

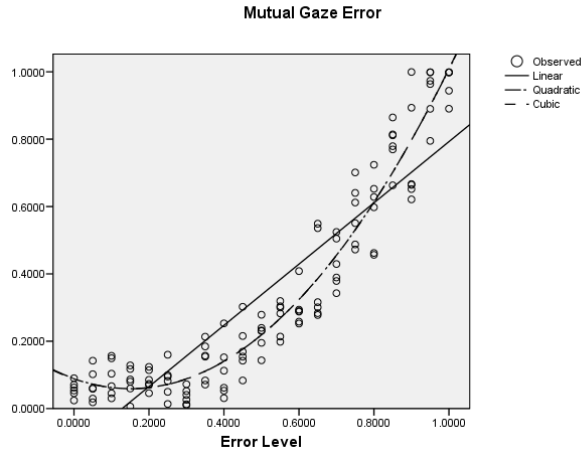

Fig. 8. Linear and Quadratic fit of the effect of noise on the error in the mutual gaze behavior.

[5] ——, "Eica: Combining interactivity with autonomy for social robots," in *International Workshop of Social Intelligence Design 2007 (SID2007)*, July 2007, pp. 227–236.

[6] ——, "Intention thorugh interaction: Towards mutual intention in human-robot interactions," in *IEA/AIE 2007 conference*, June 2007, pp. 114–124.

[7] Y. F. O. Mohammad, T. Ohya, T. Hiramatsu, Y. Sumi, and T. Nishida, "Embodiment of knowledge into the interaction and physical domains using robots," in *International Conference on Control, Automation and Systems*, October 2007, pp. 737–744.

[8] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell, "Towards a model of face-to-face grounding," in *Association for Computational Linguistics*, December 2003.

[9] T. Nishida, K. Terada, T. Takima, M. Hatakeyama, Y. Ogasawara, Y. Sumi, Y. Xu, Y. Mohammad, K. Tarasenko, T. Ohya, , and T. Hiramatsu, "Toward robots as embodied knowledge media," *IEICA Trans. Inf. and Syst.*, vol. E89-D, no. 6, pp. 1768–1780, June 2005.

[10] Y. Ogasawara, M. Okamoto, Y. I. Nakano, and T. Nishida, "Establishing natural communication environment between a human and a listener robot," in *Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction*, 2005, pp. 42–51, aISB.

[11] H. Ogawa and T. Watanabe, "Interrobot: Speech-driven embodied interaction robot," *Advanced Robots*, vol. 15, no. 3, pp. 371–377, 2001.

[12] (2007) The phasespace inc. official web site. [Online]. Available: http://www.phasespace.com/

[13] B. Scasellatti, "Investigating models of social development using a humanoid robot," in *Biorobotics*, B. Webb and T. Consi, Eds. MIT Press, 2000.

[14] T. N. T. Tajima, Y. Xu, "Entrainment based human-agent interaction," in *IEEE Conference on Robotics, Automation, ans Mechatronics*, December 2004.