Natural head movement for HRI with a muscular-skeletal head and neck robot

Steve Barker[1], Hooshang Izadi[1], Nigel T. Crook[1,2], Khaled Hayatleh[1], Matthias Rolf[2], Philip Hughes[1] and Neil Fellows[1]

***Abstract*— This paper presents a study of the movements of a humanoid head-and-neck robot called Eddie. Eddie has a musculo-skeletal structure similar to that found in human necks enabling it to perform head movements that are comparable with human head movements. This study compares the movements of Eddie with those of a more conventional robotic neck structure and with those of a human head. Results show that Eddie's movements are perceived as significantly more natural and by trend more lifelike than the conventional head's. No differences were found with respect to the impression of human-likeness, consciousness, and elegance.**

I. INTRODUCTION

Humans are highly adept at detecting genuine human bodily movement and can often reliably differentiate this from artificially generated movements such as those made by computer animated characters or robots. The perception of bodily movements and gestures is particularly important in social settings where they can often convey useful information about the intentions and feelings of others [1]. Human head gestures in particular are known to be important in one-to-one social communication [2].

An increasing number of robots are being developed for human-robot interaction in social settings. However, their ability to mimic with some degree of accuracy the movements and gestures of a human may turn out to be a limiting factor of how well these robots become integrated into these their social settings.

In order to investigate how robots might mimic human head movement, we have designed and implemented a humanoid head and neck robot called Eddie (Figure 1). Eddie has been designed with a skeletal neck that approximates the human cervical spine and is actuated by 8 Pneumatic Artificial Muscles structurally arranged to mimic the 8 primary muscles responsible for creating the movements and gestures of the human head. A full description of the robot is given in [3].

The study presented in this paper investigates the extent to which the head-and-neck mechanism used in Eddie is capable of generating human-like head movements. In particular we are interested in how the movements of this robotic head compares with the movements of (i) a more conventional robot head-and-neck design, and (ii) a real human head. More conventional head-and-neck designs typically consist of three independent (often servo-motor) actuators in the neck that are arranged such that their axes of rotation are orthogonal enabling rotational movements in the yaw, pitch and roll directions. By virtue of the cervical spine, the human head and neck, however, is capable of a much greater range of movements and gestures that conventional head and neck robots are incapable of reproducing. We believe that because of its biomimetic design Eddie is able to achieve some of these 'non-rotational' movements, and that these movements should compare well to the movements of a human head when contrasted to the conventional head and neck robot design.

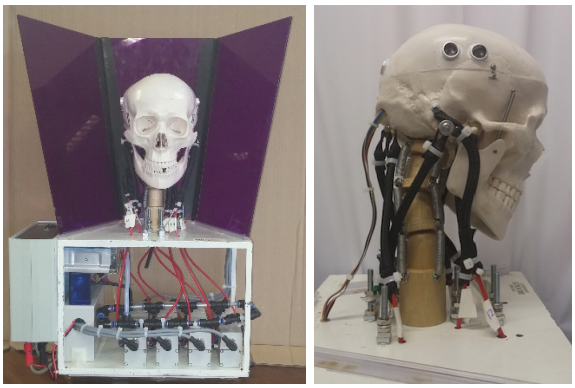

Fig. 1: The humanoid neck "Eddie". (left) Frontal view of complete apparatus. (right) Lateral view of musculo-skeletal neck and spine construction.

II. RELATED WORK

In terms of emotional acceptability, robot behaviour is a major consideration [4], in fact even more important than appearance [5] which is a problem for researchers in robotics, as appearance is an easier problem to address than behaviour. An attempt at making the physical appearance emotionally acceptable can be seen in PEARL (Personal Robotic Assistant for the Elderly) project [6] which is an attempt to develop a mobile personal service robot for the elderly with chronic disorders, that is, problems with no prospect (currently) of a cure, that require some kind of interaction with a helper. In these behaviours, motion is an important component [7], as intentions and drives can be established from motion. A particularly important consideration is head gestures as human communication is multimodal [8], and therefore need to be considered when designing robots to interact with humans [9]. Attempts have been made at producing accurate head gestures (to the non-technical observer)

[1]Department of Mechanical Engineering and Mathematical Sciences, Oxford Brookes University, Oxford, UK

[2]Department of Computing and Communication Technologies, Oxford Brookes University, Oxford, UK

[10] with varying degrees of success, due to their inability to replicate non-rotational lateral translations. Examples include Infanoid [11], Kismet [12], iCat [13], Kobian [14] and iCub [15].

III. EXPERIMENTS AND RESULTS

We adopted an experimental approach to comparing the movements generated by the proposed robot head-and-neck design with those generated by a more conventional design and a human head and neck. To this end, a series of three videos were made, with video A capturing the head movements of the Eddie robot, video B capturing the movements of a Robothespian head, which has a conventional neck design, and video C capturing the movements of a human head.

Fig. 2: (left) the mask in daylight, (right) the mask in dark conditions with backlit eyes and mouth.

Fig. 3: (left) the masked human, (center) the masked Robothespian, (right) the masked Eddie.

In order to maintain 'non bias', it was necessary to ensure that the robots and the human could not be identified by appearance. The videos were therefore designed to conceal the visual identity of the human and the robots whilst capturing their movements. This was achieved by creating a mask that could be worn by both the robots and the human. The mask was made of matt black material with the eyes and mouth backlit (Figures 2 and 3). When filmed in dark conditions, the videos captured the movements of the backlit eyes and mouth without revealing the identity of the wearer of the mask, as desired. Each video lasted 55 seconds, capturing random movements of the corresponding head in each case. The videos are attached as video submission.

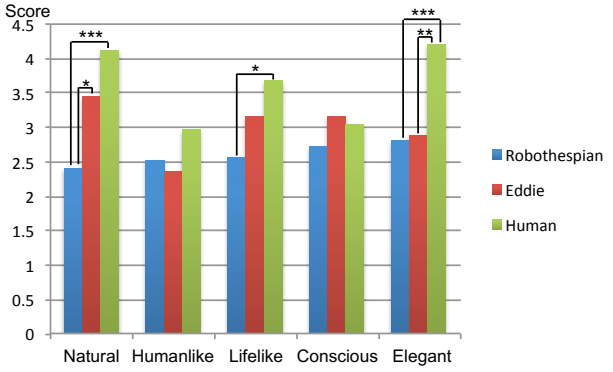

Fig. 4: Overview of results: Eddie clearly outperforms Robothespian on the "natural" score and is statistically similar to the human score on "lifelike" where Robothespian falls behind. Both robots perform poorly on the "elegant" score compared to human movement.

A 'within subjects' experimental design was adopted using a group of 25 participants sampled from a broad demographic. The experimental setup involved a human participant sat at a desk in front of a computer and a questionnaire. The experimental procedure involved asking the participant to watch video A (Eddie) and then asking them to evaluate the movements depicted in the video using a questionnaire (Table I). This procedure was repeated for videos B (Robothespian) and C (human) using the same questionnaire in each case.

The questionnaire used in the study was a subset of the questions in the validated measurement tool developed by Barkneck et al [16] that specifically evaluate movement and human-likeness. The questionnaire uses a five point Likert scale with the extremes Fake/Natural, Machine-like/Humanlike, Artificial/Lifelike, Unconscious/Conscious and Moving rigidly/Moving elegantly (Table I). Each participant, therefore, completed the questionnaire three times, once in response to each movement video. Results are summarized in Figure 4.

TABLE I: The questionnaire

Fake	1	2	3	4	5	Natural
Machine-like	1	2	3	4	5	Humanlike
Artificial	1	2	3	4	5	Lifelike
Unconscious	1	2	3	4	5	Conscious
Moving rigidly	1	2	3	4	5	Moving elegantly

A Repeated Measures analysis of variance (ANOVA)[1] was used to test the null hypothesis that the mean overall rating for the human and robot head movement videos were the same (i.e. that $H_0 : \mu_A = \mu_B = \mu_C$, where μ_X is the mean overall rating for video X, against the alternative that at least

[1]SPSS v. 23

two of the means are different). In all the following cases Mauchley's test assumed sphericity.

The results show that $F(2, 48) = 16.193$ with the p-value < 0.001, indicating a less than 0.1% probability of the means being the same across all videos (see also Appendix, Table XII). The null hypothesis can therefore be rejected and a post-hoc test used to discover where the differences lie in the evaluations of the videos. The post-hoc test was done using pairwise comparisons between the questionnaire results for the 3 videos. The results of the pairwise test are shown in Table II.

TABLE II: Pairwise comparisons based on estimated marginal means

(I) Rating	(J) Rating	(I-J) Mean Difference	Std. Error	Sig[b]
Human	Eddie	3.000*	0.733	0.001
	Robothespian	5.000*	0.902	0.000
Eddie	Robothespian	2.000	0.998	0.160

[b] Adjustment for multiple comparisons: Sidak
* The mean difference is significant at the 0.05 level

The post hoc tests, using Sidak adjustment, test the null hypotheses that $H_0 : \mu_i = \mu_j$ for $i, j \quad 1, 2, 3$. The outcome of this test is:
Reject $H_0 : \mu_{Human} = \mu_{Eddie}$, p-value < 0.001.
Fail to reject $H_0 : \mu_{Eddie} = \mu_{Robothespian}$ at 5% as p-value $= 0.16$.

Hence, the combined scores already allow to establish a difference between human and Eddie, i.e. that Eddie does not yet reach human scores. The combined scores do not yet allow to distinguish Eddie and Robothespian, but, as we will see, a closer look into the individual scores shows significant differences.

The validity of Repeated Measures ANOVA is based on the underlying assumptions of normality for the population in each group. The graphical checks and normality tests that were conducted show that this assumption is valid. Moreover, the equal group sizes guarantee that even moderate departures from the underlying assumptions are not problematic. However, to confirm the results even more emphatically, a non-parametric test (Friedman) was used. The test rejects the null hypothesis of equality of medians with $p < 0.001$ (see also Appendix, Table XIII), and the post hoc test based on Studentized Range Test confirms the results of the parametric test above. Having now established a statistically significant difference between the overall means of the 3 subjects, tests were then carried out on the different aspects of the questionnaire.

Given that general differences between the three conditions are established, a Repeated Measures ANOVA was used for each individual of the five scores to test the hypothesis that the mean overall rating for the three videos were the same (i.e. test that $H_0 : \mu_{Eddie} = \mu_{Robothespian} = \mu_{Human}$ against the alternative that at least two means are different. If significant differences are found, a post hoc tests, using Sidak adjustment for multiple comparisons, tests the null hypothesis that the scores are identical for the three experimental

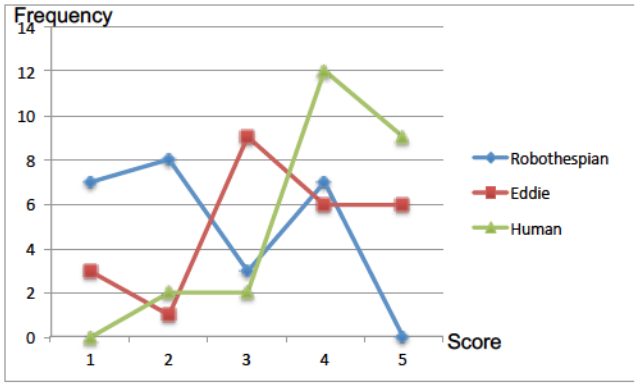

Fig. 5: Population plot Fake vs Natural: Robothespian did not receive any 5. Eddie received mostly ratings between 3 and 5. The human received almost exclusively 4's and 5's.

conditions $H_0 : \mu_i = \mu_j$ for $i, j = 1, 2, 3$. Further, we employ Friedman tests to test the null hypothesis of equality of the medians.

A. *Fake vs Natural*

The Repeated Measures ANOVA results for the fake/natural score show $F(2, 48) = 14.727$, and p-value < 0.001 (see also Appendix, Table XIV). The null hypothesis can therefore be rejected and a post hoc test can be used to find out where the differences lie. The results of this are shown in Tables III and IV.

TABLE III: Marginal Means for Fake vs Natural

F_N	Mean	Std. Error	Median
Robothespian	2.400	0.238	2
Eddie	3.440	0.252	3
Human	4.120	0.176	4

TABLE IV: Pairwise comparisons for Fake vs Natural

(I) Rating	(J) Rating	(I-J) Mean Difference	Std. Error	Sig[b]
Human	Eddie	0.680	0.304	**0.101**
	Robothespian	1.720*	0.297	0.000
Eddie	Robothespian	1.040*	0.353	0.021

[b] Adjustment for multiple comparisons: Sidak
* The mean difference is significant at the 0.05 level

The pair-wise test gives:
Do not reject $H_0 : \mu_{Human} = \mu_{Eddie}$, p-value $= 0.101$.
Reject $H_0 : \mu_{Human} = \mu_{Robothespian}$ and $H_0 : \mu_{Eddie} = \mu_{Robothespian}$.

This shows when rating the fake/natural score, the means ratings for the human and Edie were not statistically different, whereas the mean rating for Robothespian video was statistically different from the other two. Together with the mean scores (see also Figure 4) this shows that Eddie's movement is clearly perceived as more natural than Robothespian's, and close to the human's rating (see population plot in Figure 5).

The Friedman Test for the Fake vs Natural case rejects the null hypothesis of equality of the medians and the post hoc

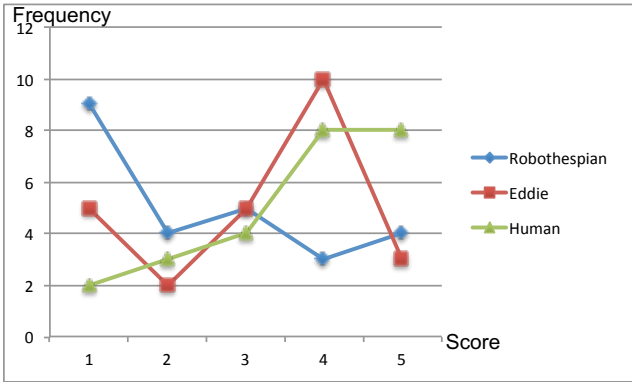

Fig. 6: Population plot Artificial vs Lifelike: Robothespian's scores are clearly tilted to 1 (artificial), while the human scores are tilted to 5 (natural). Eddie is in between with no clearly visible tendency.

test based on the Studentized Range Test confirms the same result (details in Appendix, Table XV).

B. Machine-like vs Humanlike

The mean scores for Machine-like vs Humanlike are shown in Table V (see also Figure 4).

TABLE V: Marginal Means for Machine-like vs Humanlike

M_H	Mean	Std. Error	Median
Robothespian	2.520	0.295	2
Eddie	2.360	0.276	2
Human	2.960	0.300	3

Results in all three conditions are very similar. The Repeated Measures ANOVA does not reveal any significant differences (see Appendix, Table XVI). The test of equality of medians only gives marginal results (see Appendix, Table XVII). Therefore, regarding the machine-like/humanlike question, the evaluation of all three videos exhibited no significant differences.

C. Artificial vs Lifelike

The Repeated Measures ANOVA on Artificial/Lifelike rejects $H_0 : \mu_{Eddie} = \mu_{Robothespian} = \mu_{Human}$ with $p < 0.05$ (details in Appendix, Table XVIII), thereby showing significant differences between the 3 videos.

TABLE VI: Marginal Means for Artificial vs Lifelike

A_L	Mean	Std. Error	Median
Robothespian	2.560	0.300	2
Eddie	3.160	0.269	2
Human	3.680	0.256	4

The p-value of a sharp 5% is too marginal to rule out the difference between the means for video B (Robothespian) and video C (human). While Robothespian and human are different at least be trend, it is interesting to see that Eddie reaches statistically similar performance to the human's score. Visual inspection of the population plot of participants' answers (Figure 6) confirms this: while

TABLE VII: Pairwise comparisons for Artificial vs Lifelike

(I) Rating	(J) Rating	(I-J) Mean Difference	Std. Error	Sig[b]
Human	Eddie	0.520	0.366	0.424
	Robothespian	1.120	0.437	**0.050**
Eddie	Robothespian	0.600	0.436	0.451

[b] Adjustment for multiple comparisons: Sidak
* The mean difference is significant at the 0.05 level

Robothespian's scores are tilted to 1/"artificial", Eddie's answer distribution is visually closer to the human's answers distribution which is tilted towards 5/"lifelike".

However, the non-parametric test fails to reject the equality of medians (details in Appendix, Table XIX). Therefore, overall it would appear that the means for the artificial/lifelike parameter are significantly similar for video B (Robothespian) and video C (human), as well as the medians.

D. Unconscious vs Conscious

The Repeated Measures ANOVA on this measure does not allow to reject $H_0 : \mu_{Eddie} = \mu_{Robothespian} = \mu_{Human}$ (details in Appendix, table XX). Therefore, there is no significant difference between the means. Also the hypothesis of equality of medians can not be rejected (Appendix, Table XXI). Therefore, for the unconscious/conscious question, all 3 videos showed statistically similar results.

TABLE VIII: Marginal Means for Unconscious vs Conscious

U_C	Mean	Std. Error	Median
Robothespian	2.720	0.268	3
Eddie	3.160	0.243	3
Human	3.040	0.241	3

E. Rigidly vs Elegantly

The Repeated Measures ANOVA allows to clearly reject $H_0 : \mu_{Eddie} = \mu_{Robothespian} = \mu_{Human}$ with $p < 0.001$ (see details in Appendix, Table XXII). Also the hypothesis of equality of medians can be rejected (Appendix, Table XXIII).

TABLE IX: Marginal Means for Rigidly vs Elegantly

R_E	Mean	Std. Error	Median
Robothespian	2.800	0.265	3
Eddie	2.880	0.273	3
Human	4.200	0.173	4

TABLE X: Pairwise comparisons for Rigidly vs Elegantly

(I) Rating	(J) Rating	(I-J) Mean Difference	Std. Error	Sig[b]
Human	Eddie	1.320*	0.330	0.002
	Robothespian	1.400*	0.316	0.001
Eddie	Robothespian	0.080	0.346	0.994

[b] Adjustment for multiple comparisons: Sidak
* The mean difference is significant at the 0.05 level

The pair-wise comparison results show a significant difference in means for video C (human) and video A (Eddie) and video C and video B (Robothespian), but not in the means

for video A and video B. Hence, Robothespian and Eddie perform equally poor on this measure compared to human movement.

IV. CONCLUSION

TABLE XI: Summary of the results

	A/C	A/B	B/C
Overall mean	Dissimilar	Similar	Dissimilar
Fake vs Natural	Similar	Dissimilar	Dissimilar
Machine-like vs Humanlike	Similar	Similar	Similar
Artificial vs Lifelike	Similar	Similar	Similar
Unconscious vs Lifelike	Similar	Similar	Similar
Rigidly vs Elegantly	Dissimilar	Similar	Dissimilar

A = video of Eddie robot
B = video of Robothespian robot
C = video of human

The terms similarity and dissimilarity in this table denote failing to reject and rejecting H0 at 5% significance level, respectively.

This study investigated whether Eddie [3] the biomimetic robot head/neck system improves over traditional robot head/neck systems in terms in terms of the perceived characteristics of its motion. In summary (see Figure 4 and Table XI) Eddie's motion is perceived as more natural and lifelike than the conventional robot head, but not as natural and lifelike as human motion, yet. In comparing artificial vs lifelike and unconscious vs conscious all the movement in the videos were seen as similar. This indicates that the conventional and the muscular skeletal robot head both give perceived realistic movement for these Likert human likeness scales. Although there is this improvement, with the muscular skeletal robot, the robot movement videos were still seen to be more similar to each other than to the human movement video. Most markedly when it comes to rigidity vs elegantly. When examining the mean values for each video it can be seen that participants often struggled to judge realistic movement (Tables V, VI, VIII) with values falling close to the mid-scale value of 3. Apart from unconscious vs conscious, Table VIII, the human movement video was always perceived to be more realistic (higher mean values). For fake vs natural and rigidity vs elegantly participants were much more able to identify the human movement video as being realistic (Tables III, IX). Although improvement has been seen with the muscular skeletal robot; work is still required to give convincing human movement, particularly when comparing Rigidity vs Elegantly.

APPENDIX

TABLE XII: Tests of within-subject effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Rating	316.667	2	158.333	**16.193**	**.000**
Error (Rating)	469.333	48	9.778		

TABLE XIII: Friedman Test results

N	25
Chi-Square	20.702
df	2
Asymp. Sig.	0.000
Exact Sig.	0.000
Point Probability	0.000

A. Fake vs Natural

TABLE XIV: Tests of within-subject effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
F_N	37.520	2	18.760	14.727	0.000
Error (F_N)	61.147	48	1.274		

TABLE XV: Friedman Test results for Fake vs Natural

N	25
Chi-Square	16.349
df	2
Asymp. Sig.	0.000
Exact Sig.	0.000
Point Probability	0.000

B. Machine-like vs Humanlike

TABLE XVI: Tests of within-subject effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
M_H	4.827	2	2.413	1.252	0.295
Error (M_H)	92.507	48	1.927		

TABLE XVII: Friedman Test results for Machine-like vs Humanlike

N	25
Chi-Square	5.692
df	2
Asymp. Sig.	0.058
Exact Sig.	0.058
Point Probability	0.004

C. Artificial vs Lifelike

TABLE XVIII: Tests of within-subject effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
A_L	15.707	2	7.853	3.661	0.033
Error (A_L)	102.960	48	2.145		

TABLE XIX: Friedman Test results for Artificial vs Lifelike

N	25
Chi-Square	4.207
df	2
Asymp. Sig.	**0.122**
Exact Sig.	**0.123**
Point Probability	0.005

D. Unconscious vs Conscious

TABLE XX: Tests of within-subject effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
U_C	2.587	2	1.293	0.846	0.436
Error (U_C)	73.413	48	1.529		

TABLE XXI: Friedman Test results for Unconscious vs Conscious

N	25
Chi-Square	0.494
df	2
Asymp. Sig.	0.781

E. Rigidly vs Elegantly

TABLE XXII: Tests of within-subject effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
R_E	30.907	2	15.453	11.280	0.000
Error (R_E)	65.760	48	1.370		

TABLE XXIII: Friedman Test results for Rigidly vs Elegantly

N	25
Chi-Square	17.062
df	2
Asymp. Sig.	0.000
Exact Sig.	0.000
Point Probability	0.000

REFERENCES

[1] R. Blake and M. Shiffrar, "Perception of human motion," *Annu. Rev. Psychol.*, vol. 58, pp. 47–73, 2007.

[2] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," *Artificial Intelligence*, vol. 171, no. 8-9, pp. 568–585, 2007.

[3] S. Barker, L. A. Fuente, K. Hayatleh, N. Fellows, J. J. Steil, and N. T. Crook, "Design of a biologically inspired humanoid neck," in *Robotics and Biomimetics (ROBIO), 2015 IEEE International Conference on*. IEEE, 2015, pp. 25–30.

[4] C. ÖZGEN, "Human-like robot head design," Ph.D. dissertation, MIDDLE EAST TECHNICAL UNIVERSITY, 2007.

[5] S. Maddock, J. Edge, and M. Sanchez, "Movement realism in computer facial animation," in *19th British HCI group annual conference, workshop on human-animated characters interaction*, vol. 6, 2005, pp. 1–4.

[6] M. E. Pollack, L. Brown, D. Colbry, C. Orosz, B. Peintner, S. Ramakrishnan, S. Engberg, J. T. Matthews, J. Dunbar-Jacob, C. E. McCarthy *et al.*, "Pearl: A mobile robotic assistant for the elderly," in *AAAI workshop on automation as eldercare*, vol. 2002, 2002, pp. 85–91.

[7] H. Ishiguro, "Interactive humanoids and androids as ideal interfaces for humans," in *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 2006, pp. 2–9.

[8] R. Bischoff and V. Graefe, "Dependable multimodal communication and interaction with robotic assistants," in *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*. IEEE, 2002, pp. 300–305.

[9] Y. Kuno, K. Sadazuka, M. Kawashima, S. Tsuruta, K. Yamazaki, and A. Yamazaki, "Effective head gestures for museum guide robots in interaction with humans," in *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*. IEEE, 2007, pp. 151–156.

[10] I. Lütkebohle, F. Hegel, S. Schulz, M. Hackel, B. Wrede, S. Wachsmuth, and G. Sagerer, "The bielefeld anthropomorphic robot head flobi," in *Robotics and automation (ICRA), 2010 IEEE international conference on*. IEEE, 2010, pp. 3384–3391.

[11] H. Kozima, C. Nakagawa, and H. Yano, "Using robots for the study of human social development," in *AAAI Spring Symposium on Developmental Robotics*, 2005, pp. 111–114.

[12] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003.

[13] A. van Breemen, X. Yan, and B. Meerbeek, "icat: an animated user-interface robot with personality," in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. ACM, 2005, pp. 143–144.

[14] N. Endo, S. Momoki, M. Zecca, M. Saito, Y. Mizoguchi, K. Itoh, and A. Takanishi, "Development of whole-body emotion expression humanoid robot," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 2140–2145.

[15] G. Metta, L. Natale, F. Nori, and G. Sandini, "The icub project: An open source platform for research in embodied cognition," in *Advanced Robotics and its Social Impacts (ARSO), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–26.

[16] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.