A Biological Inspired Cognitive Model of Multi-sensory Joint Attention in Human Robot Collaborative Tasks*

O. Eldardeer, G. Sandini, F. Rea

***Abstract* - In collaborative tasks, human partners and robots need to coordinate in shared environment. One of the fundamental prerequisites for effective collaborations is sharing an attentional focus on the same perceptual events. Joint attention has been widely studied through the investigation of its psychological, cognitive and social defining elements. Also, in the field of human robot interaction, the joint attention has been extensively exploited and defined as fundamental prerequisite for proficient collaborations. We investigate how to enable joint attention between human and robot partner, implementing new attention system from biologically inspired approch, and refined to obtain attentive performance (response time and salient stimulus localization) similar to human partner. We assess the attentive system with the humanoid robot iCub when involved in a joint task with the human partners in order to compare the attentive performance of both when stimulated by the same stimuli from different modalties. The result shows similar performance when both are visually stimulated and significative difference in performance when both are auditory stimulated.**

***Keywords* - Joint attention, cognitive models, computational neuroscience, human-robot interaction, multisensory integration**

I. Introduction

Joint attention is defined as the shared attentional focus on the same perceptual event by multiple individuals that coexist in the same environment. This complex phenomenon has been studied investigating the physiological, cognitive and social aspects of the process but most of the investigations on joint attention process have been carried in rigidly controlled settings and only rarely the research focused on realistic and unstructured environments. In context where the joint attention has been studied involving both artificial agent and human participants the joint attention mechanism has defined through an assessment of exclusively the human performance or the artificial agent performance. Rarely the assessment of the performance regarded the combined performance (including mutual influence) of all the participants involved in the task. This approach limits the thorough assessment of the attentive systems comprising both the human and the artificial agents that operate in joint collaborative tasks. Our contribution aims to improve existing multisensorial systems for attentional redeployment (PROVISION and auditory attention for humanoid robots) to specifically address joint attention in collaborative tasks where the presence of the human agent has significative impacts. The improvement of the multisensory attention system is finalized to provide an attention system that is comparable with the human attention performance in terms of reaction time and accuracy in the localization. The underpinning concept is that more human-like artificial attention systems will make the attention focus of both the parts mutually understandable and consequently the interaction becomes more natural and efficient. At the same time, we aim to strictly measure in realistic scenario the performance (reaction time and localization accuracy) of the entire collaborative group comprising mutually influences between the human participants and the humanoid robot iCub.

II. Background and Related Work

Joint attention phenomenon has been researched in cognitive science [1] [2]. Recent researches shade a light on the neurocircuitry entailed in the mutual focus of two human individuals on a common attentional target. Joint attention has been deeply studied in social studies [3] [4] indicating what are the social processes at the basis of shared attentional focus redeployment. Further, it has been demonstrated that the process of attending jointly the same stimuli has been already observed in infants around the first 18 months of age [5] [6]

The role of joint attention during interaction of human and artificial agents has been investigated through the study of attention timing of gaze patterns [7]

Differently from virtual agents, in Human Robot Interaction studies, the attentional skill of the physically present artificial agent coordinates with the attentional skills of human partner. Therefore, It is important to propose studies of attention-timing that take into consideration the role of robotic agent. In the recent year, different computational models of attention for artificial agents have been developed [8] [9] to respond to visual [10] and auditory stimuli [11] However, only few attention systems have been designed and valuated to specifically address the context of collaboration between the human and the physically present robot partner [12]. In particular, aspects such the attention-timing and the focus redeployment strategies, are e addressed in computational modeling of the attention system but however these neglect the impact of mutual presence in the human robot interaction. Such mutual influence is of fundamental importance especially when assessing joint attention and yet this dimension is often not considered in the joint attention studies. Our contribution has dual objective: first, to improve the existing multisensory systems for attentional redeployment in order to address joint attention in collaborative tasks, and second, to demonstrate the performance of our solution by evaluating the whole collaborative group comprising both the human and the robot participants.

We started by developing the existing PROVISION system: saliency-based selective attention model which

[*] The research has been supported by the FLAG-ERA project Robocom++.

O. Eldardeer (corresponding author: omar (dot) eldardeer (at) iit (dot) it) is with affiliated to DIBRIS, Università di Genova, Italy and to RBCS, Istituto Italiano di Tecnologia - Erzelli, Genova, Italy.

F. Rea, and G. Sandini are affiliated to RBCS, Istituto Italiano di Technologia - Erzelli, Genova, Italy. Emails: {francesco (dot) rea, giulio (dot) sandini (dot) it}.

implemented for iCub robot [13]. In its bottom-up implementation the attention system decomposes the visual scene to a list of feature maps. the features are linearly combined with a specific weight for each of them. The PROVISION system has been recently integrated with auditory attention system based on the Bayesian sound localization model developed for humanoid robotic platform [14]. As well as PROVISION, the auditory attentive system is a biologically inspired model that uses only binaural sensing (two microphones in the head of the humanoid robot iCub) in order to calculate a Bayesian allocentric probability map for the location of the audio.

III. Methodology

A. Model of multisensory attention

Our first contribution is the integration between the audio localization system and the Provision attention system. The is done by extracting an egocentric map from the allocentric output of the audio localization system. Using this extracted map we transferred the Bayesian values of the angles that are in the field of vision to a cartesian saliency map. The cartesian saliency map is taken as an additional input feature to the spatial linear combination of topographic saliency map. The objective is to allow the robot to have an attention mechanism based on the audio source in combination with visual attention. Additionally, to reinforce the selectivity of the targets by relying on both visual and auditory features Instead of visual features only.

The second contribution aims at investigating how biologically plausible multisensory attention system should redesign attentional timing and focus redeployment strategies to promote joint attention between human and robotic platform. As found in other attention models [15] [16], we moved from cyclical selection of attentive focus, typical of attentive systems implemented on robots, to favorite the temporal asynchronous attending at salient changes in the sensorial landscapes. This allows for resembling of asynchronous attentional redeployment of human partner which in turn facilitate joint attention. We propose here a new way to identify the uniqueness trying to avoid continues selection and rather activate actions then the stimuli is clearly unique. Our goal here is to compare this approach with the human behaviour.

This feature is implemented through the acyclic extraction of a saliency *"hot point"*. A saliency *"hot point"* is a spatial location in the scene that is extremely salient when compared to the rest of the scene. It is not only the most salient point in the scene, but it also has a high salience comparing to the whole scene. This spatial location is what is identified as worth attending by the attention of humans interactants when a stimulus is presented Therefore, we compare the maximum salient point with the whole distribution of attentive responses across the entire combined saliency map. If the value of the maximum point exceeded the triple of the standard in comparison to the mean value with a certain threshold then it is considered as a hot point. It means that this point is extreamly salient compared relatively to the full scene. The threshold here represents a tuning parameter for the sensitivity of the system. From this point of the paper we will refer to (mexValue – meanValue – 3 σ) as " gamma Value".

$$(maxValue\ -\ meanValue\ -3\,\sigma)\begin{cases} hot\ point & \text{if,} > threshold \\ not\ a\ hot\ point & \text{if,} \leqslant threshold \end{cases}$$
$$\sigma:\ standard\ diviation\ of\ the\ combined\ seliency\ image$$

And finally, the third contribution in this work is the projection of the retinotopic response into allocentric spatial representation for motor control. The motor control systems is expecting the allocentric world-based 3D location from the location of the target in the scene to properly respond and initiate motor commands [17]. We provide this by using the prior knowledge about the contextualization of the environment. It is reasonable to assume the context of an interaction on the shared working space, Therefore, we defined the attentive plane as the geometrical plane where the speakers and bulb lie on. Knowing the plane and using the single image the 3D projected location of a point in the giving image is expected from the origin of the robot. Mathematically, this is done by computing the 3D projection (x,y,z) of the pixel location (u,v) in the image on the plane defined by the equation "aX + bY + cZ + D + 0". The attentive plane can be represented as the pre-defined region which includes the selected stimuli that require a response in a specific context. It is the area of the cooperative task between the human and the robot. This bit of information is a shared prior knowledge between the robot and the human. Finally, the point will satisfy the response execution if the expected 3D point is in the pre-defined cooperative area of the task. The following equation shows the simple form of the projection 3D image projection where f is the focal length of the camera.

$$\begin{pmatrix} u \\ v \end{pmatrix} = f/z \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$
$$u:\ the\ vertical\ component\ of\ the\ vector\ in\ the\ image$$
$$v:\ the\ horizontal\ component\ of\ the\ vector\ in\ the\ image$$
$$z:\ the\ pre-defined\ depth\ of\ the\ plane$$
$$x,y,z\ :\ the\ 3D\ location\ in\ space$$
$$f:\ focal\ length\ of\ the\ camera$$

B. Software Infrastructure

The system is developed using yarp [18] in C++ and python. In the design of the software infrastructure we opted for modularity and for multiple connections between modules. The Fig. 1 shows the structure of the system implementation. The first stage after acquiring the visual and audio data from the sensors (camera and microphones) comprises some specific modules:

Egocentric audio cropper: in this module, we generate the cartesian saliency map from the allocentric audio Bayesian map. The allocentric audio Bayesian map can be represented as an array of 360 probabilistic values corresponding to the 360 degrees centralized across the axis of the robot. Based on the current head azimuth angle and the gaze angle we cut only the relative angles that represent the field of vision (FOV). If the maximum probability of the sound source location is among the field of vision exceeded a certain threshold a strap is creating in that location and then extended to a 2D cartesian image. This Cartesian image is then sent as a cartesian feature map to the provision attention model. In the paper we present the maximum probability as the confidence value of the cartesian image of the audio.

Attention Manager: This module is a controller for the whole attention system. It is also a middle module between attention and any other applications or modules. It is responsible for the computational process of the hot point as well as sending/ receiving commands with external systems. It receive the combined cartesian saliency image after the linear combination and also communicate with the action execution part with the required information. This module has the full control to suspension and resume the attention process as well as manipulating the parameters of the provision system as well as the audio integration stage. This module represents the gate for any external modules need to communicate with the attention system.
Attention Action Linker is the module that estimates the 3D location of the point of interest in the scene and checks if it is inside or outside the attentive plane. It gets the hot point from the attention manager module. It is also communicating with the attention manager with the state of the action. This module limits the action execution to be done under certain pre-defined conditions. The conditions are location constrains in the 3D world. We build this module as a decision-making layer to decide when and how the action will be executed. Additionally, it sends the state of the action execution to attention manager so that the manager is able to suspend the attention system and the gaze during the action execution and resume them after finishing the action.

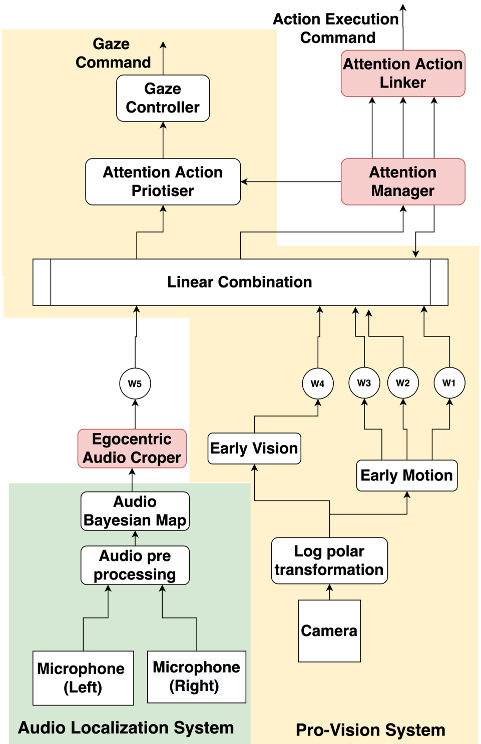

Figure 1 : software infrastructure integrating auditory and visual attention systems in order to provide attentive motor commands. The yellow part is the existing provision attention system and the greed part is the audio localaization system. The modules with the red color are the new modules which we implemented.

C. *Experiment*

The main goal of this experiment is to address the performance assessment of the joint attention system measuring the performance of the robot and human subject in response to two different sensorial stimulutions. The first one is based exclusively by auditory stimuli only and the second one combines audio-visual stimuli.

We used the humanoid robot iCub [19] in our study. It is equipped with different sensors including two cameras (eyes) and two microphones (ears) which we used in our experiment to perceive the environment.

In the experiment, the subject is sitting on a chair facing the robot. In between the subject and the robot, there is a black table with the stimuli distribution on it. The audio stimuli are produced using four identical black Bluetooth speakers distributed in a horizontal line with fifteen centimeters apart. The audio stimuli are 240 Hz sin wave. The distance between the speakers' line and the robot is 65 centimeters while for the human is 45 centimeters. The first speaker on the left side of the subject is coupled with a colored bulb which represents the visual stimuli. In this experiment, we used a fixed color (blue).

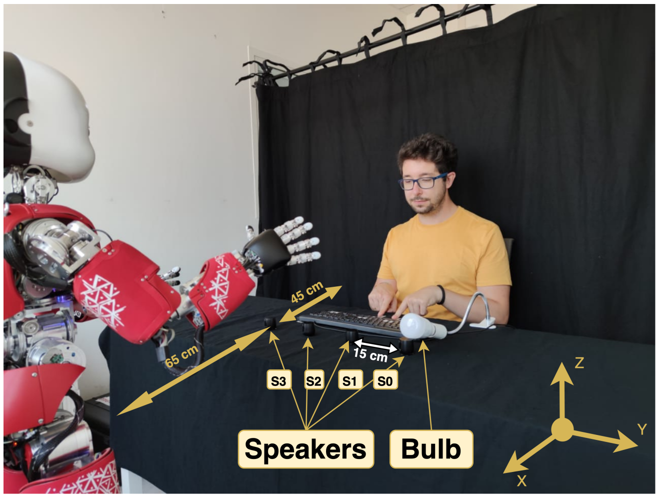

Figure 2 Experimental setup where both the humanoid robot and the human participants are stimulated with visual-auditory stimuli provided on the table. Four stimuli are presented (S0, S1,S2, and S3) the fisrt stimuli has both visual and auditory sources and the fisrt stimuli has both visual and auditory sources and the rest are only auditory. The keyboard allows the participant to indicate the source of the stimulation and to measure the reaction time. The axis shows the directions of the axis and its origin is at the torso of the robot. The distances between the stimuli are shown also the distance between the stimuli and both the human and the robot.

The experiment consists of 32 rounds for each subject. In each round, a random number is selected between one and four corresponding to the location of the stimuli. Then the stimuli are activated exclusively for the selected speaker. If the round is running for the first speaker then the bulb will turn on in a synchronic way with the speaker. The turn-on duration is {10} seconds. The subject is requested to react as fast as possible by pressing the two buttons corresponding to the activated stimuli using both of his/her index fingers and return then back to the initial position on the edge of the table. The buttons in the keyboard are selected to be approximately equal distance from the initial position. The time between the round is 10 seconds so the total time of the round is 20 seconds. The location pattern is randomized but with the same sequence for all subjects as the following sequence: (S1, S0, S1, S3, S2, S2, S0, S3, S1, S3, S2, S0, S3, S2, S3, S1, S0, S2, S3, S0, S1, S0, S0, S2, S1, S3, S1, S2, S0, S1, S3, S0).

D. *Measurements*

The main goal is comparing the joint-attention performance of the human and the artificial agents in the same cooperative task. Therefore, and for the human side, we recorded the reaction of the human which keys are pressed. For the robot we recorded the profile of the audio as well as the full attention system and the action execution commands. The profile of the audio consists of the value of the maximum confidence as well as its egocentric location. From the full system we recorded the analysis of the combined scene to have the gamma value and its expected location in the 3D world.

IV. Results and Analysis

In order to compare the performance of the robot and related with the performance of the human in the joint attention task, we analyzed the overall performance their temporal response after the stimulus is presented. The stimulus presented in characterized as auditory only when the auditory stimulus is reproduced by one of the speakers and characterized as visual-auditory when the stimulus is produced by the light of bulb and the sound from the speaker. We considered the time window of 20sec representing the cyclic stimulus production. In the cycle, the first 10sec the stimulus is provided in the first 10sec whereas in the second 10sec no other stimulus is produced in the scene.

A. *Overall performance*

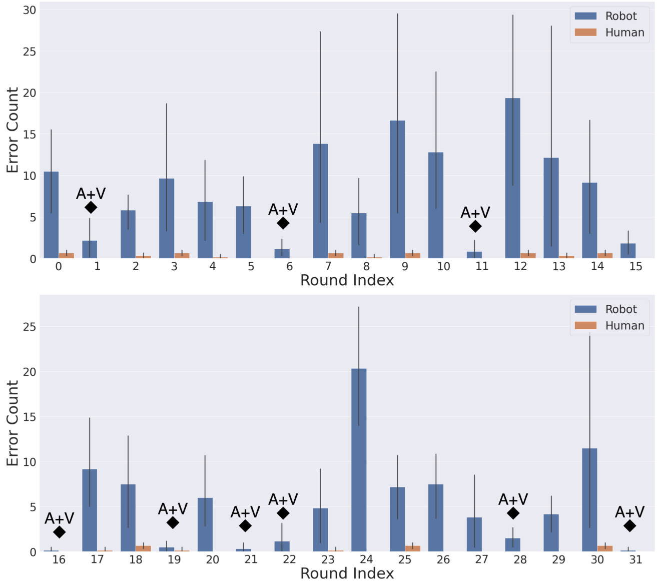

Figure 3 : The Error Count for both the robot and the human participants. The robot in blue and the human is in orange. Starting from the first round to the last round, the figure shows the error count in each round across the six experiments

The overall performance of the humanoid robot is compared with the response of human participants. In particular we indicate the number N_{wrong} as the *number of wrong attended location* $L_{attended}$ computed according the following formula: $N_{wrong} = N_{wrong} + 1$, if $|L_{target} - L_{attended}| > \theta$.

The threshold $\theta = 0.10$[m] since the distance between two consecutive stimulus location is 0.15[m]. For the robot it is computed knowing the fixation point commanded by the attention system and for the human participant it counts the number of wrong stimulus selection at the keyboard. The number of wrong attended location is comparable only for trials where visual-auditory stimulus is presented whereas there is a significative difference between the performance of the robot and the human participant when the stimulus is exclusively auditory. It worth noting that especially for auditory stimulation the task resulted difficult for the human participants since the auditory localization of pure tone without head movement provided within a azimuthal angle range is perceptually challenging.

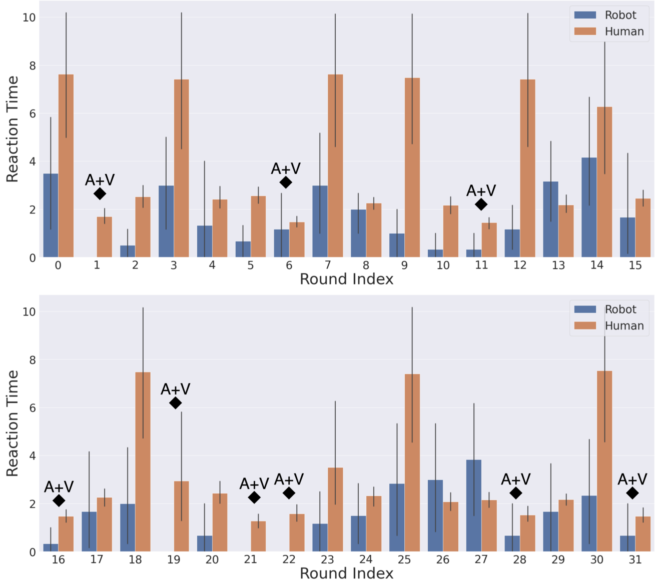

Figure 4 :Reaction time in seconds for both the robot and the human participants. The robot in blue and the human is in orange. Starting from the first round to the last round, the figure shows the reaction time in each round across the six experiments

Comparing instead the performance of the robot and the human participants in average, it is worth noting how the robot has wider variability in response time RT_{robot} but the response time is not significantly different with respect to human participant`s response time $RT_{human}$. This is also true if we discard $RT_{human} > 5$[s] as human subject`s mistakes. In fact if both the keys failed to be synchronously pressed the system does not record the human`s response. The robot instead shows immediate response indicating quick detection of the presence of a new stimulus but it is also not as accurate as human participant in localizing the stimulus source in space.

B. *Temporal analysis*

For the temporal analysis of the auditory attention behaviour of the humanoid robot iCub the value δ indicates the confidence that salient auditory stimulus is present in the scene. In average the Figure 5 shows how the value δ changes during the 20secs of trial cycle and in particular how the confidence increases for the first 10secs and decreases for the last 10 secs of the trial. All the value of δ that exceed the threshold TH_{δ} indicate the presence of salient auditory stimulus and the localization process of the auditory stimulus is initiated.

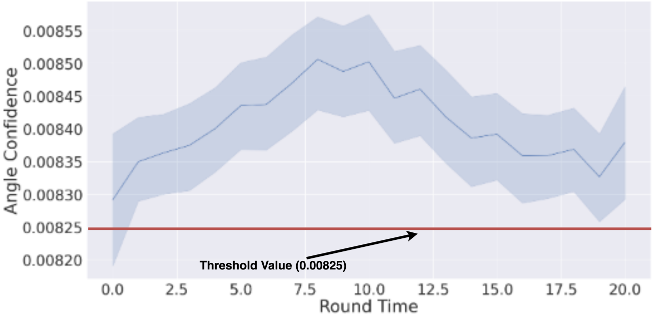

Figure 5 :Confidence value δ profile

The value of TH_{δ} is fundamental parameter that impacts on the number of errors since a lower TH_{δ} will activate localization even when the Bayesian network is not confident enough on the stimulus location but also impact on the reaction time since higher level will prevent the system for attending the auditory stimulus. In the correct implementation we opted for TH_{δ} = 0.0085 which has impacted on fast reaction time but also in the high number of errors.

For the temporal analysis of the visual attentive behaviour of the humanoid robot iCub involved in joint attention task we focus on the 20 seconds after the stimulus on-set observing in particular how the attentive system responds in presence of the stimulation (first 10 seconds) and in absence of the stimulation (second 10 seconds).

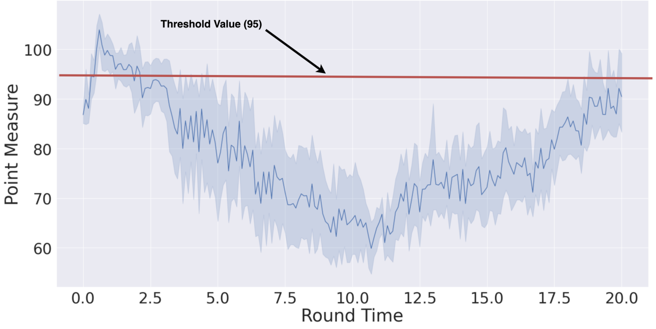

Figure 6 : average gamma value (±st.error) temporal profile for all the trials involving visual-auditory stimulation. The figure shows a swift increase when the stimuli is presented

In figure 6 we report the average of the γ progression in the 20 seconds time window for all the subjects and all the trials. In particular, the γ value, which is the value that indicates the uniqueness of the most salient stimulus, increases for the first 10seconds and decreases in absence of the salient stimulus. When the γ value exceeds the threshold TH_{γ}, the multimodal attention system detects the presence of unique salient stimulus significantly different from the rest of the scene and it proceeds for the spatial localization of the target stimulus. It is worth observing the swift increment of confidence in average of γ value (<500ms in average) that explains the faster response rate RT_{robot} with respect to the RT_{human}. In absence of the stimulus (after 10 secs) the γ increase because other stimuli in the scene become salient with respect to the rest of the scene. The process is gradual but the γ value exceeds the TH_{γ} in average after 18 secs generating new responses of saliency and consequently new target localization attempts. This is where mostly of the wrong attended locations are generated thus explaining the greater number of errors in stimulus localization.

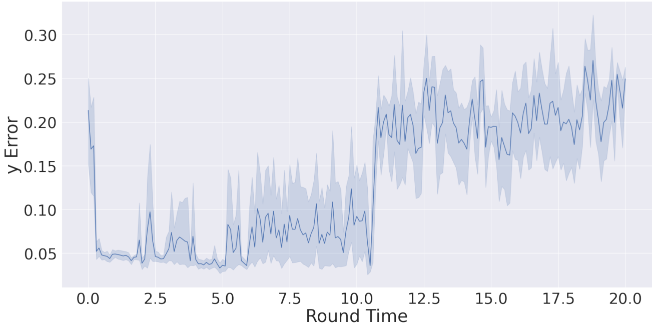

Figure 7 : Localization error for attentive task involving visual-auditory stimulation

Concerning the process that spatially determines the position of the most salient stimulus in the scene, the performance is especially promising for the localization of visual-auditory stimuli. Figure 7 shows the error along y axis (the direction of the speaker deployment and azimuthal angle for the robot) since the x (depth) and z (elevation) axis do not show significative changes. As shown in Figure 7, as soon as the stimulus is set the error $Error_{y}$ drops to 0.05[m] in less than 500ms . Such RT_{robot} observed even before is comparable reaction time with respect the RT_{human}. In average for the first 5secs the error oscillates around the 0.05[m] and only partially adjusts in the second 5secs remain in average below the 0.10[m]. After the stimulus stops (after 10secs) the robot attends other locations not necessarily corresponding to the target of this trial. Such behaviour of the robot in absence of salient stimulus is comparable with the unconstrained behaviour of the human participants that in absence of stimulus attends other salient locations in the scene (not necessarily the target of the trial) and sometimes even the robot partner.

V. Discussion and Future Work

The architecture proposed for joint attention in the context of human robot interaction shows promising behaviour although not comparable with the attentive pattern of human counterpart. It is clear that, whereas the visual attention system swiftly and correctly captures changes in visual field that correctly drive the attentive system, the auditory attention remains a challenge for both the human participant and the robot. In particular, the majority of the subject mentioned the difficulty in localizing the source of sound when only the pure tone is produced. The interesting question for our future study focuses on the test with complex tone sounds or speech which is richer in terms of auditory features for sound recognition and localization. Whereas the human participant can refine the estimation of the stimulus location by head rotations and then solve the auditory problem [20] he robot does not fully leverage on its motor capabilities to disambiguate the uncertain situation. In future work, we plan to endow the humanoid robot iCub with motor control finalized to the refinement of its estimation and to ask the human participants to wear eye and head tracking system that can give insights on the degree of involvement of head rotation and eye fixation process. Further, the auditory attentive system can be automatically adapted to the contextualization of the specific task. It is more efficient to reallocate the limited computational resources of the robot in the direction of assessing exclusively what is happening in the task (e.g.: limiting the auditory beams to the exclusive area in front of the robot, adjusting the frequency bands to the range of interest) and this can make the robot more efficiently react to the auditory stimulation.

Another interesting point is that multisensory information equally supports attentive process of both the human observer and the robot observer. In situations where both the auditory and visual stimuli mutually reinforce, both the human partner and the humanoid robot iCub improve their attending performance. It is also worth to mention that the results provided in this paper are preliminary. Therefore, a richer testbed with multisensory stimulation (auditory and visual) for all the stimulus locations in our future study will give us insights and deeper results on the benefit of multisensory integration across different locations

VI. Conclusion

If robots are going to be used in support to daily activities, it is important to understand how process of joint attentions work in typical human robot interactions. The joint attention is important mediator for efficient collaborations between the interactants however, it is challenging to endow robotic platforms with the same attentive capabilities (reaction time and localization accuracy) of human partners. Our contribution aims at improving existing auditory and visual attention systems with specific mechanisms that promote attention in human-robot collaborations. We demonstrated that the performance of improved system is comparable with the performance of human participants when the multisensory stimulations (auditory and visual) is presented at the same time to both the human participant and the robot. On the other hand the challenge of attending only auditory stimuli is only partially fulfill for the humanoid robot since the robot correctly recognizes the presence of salient new stimulation but shows limitation in the correct localization of the stimulus.

Acknowledgment

This experiment could never have been done without the help of Lucas Grasse and Dr. Matthew Tata (University of Lethbridge, Canada), supporting us in the design and development of the cognitive models of the auditory attention.

References

[1] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Current directions in psychological science,* vol. 16, no. 5, pp. 269-274, 2007.

[2] L. Schilbach, M. Wilms, S. B. Eickhoff, S. Romanzetti, R. Tepest, G. Bente, N. J. Shah, G. R. Fink and K. Vogeley, "Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry," *Journal of cognitive neuroscience,* vol. 22, no. 12, pp. 2702-2715, 2010.

[3] A. Frischen, A. Bayliss and S. Tipper, "Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences," *Psychological bulletin,* vol. 133, no. 133, pp. 694-724, 2007.

[4] M. Jording, A. Hartz, G. Bente, M. Schulte-Rüther and K. Vogeley, "The “Social Gaze Space”: A Taxonomy for Gaze-Based Communication in Triadic Interactions," *Frontiers in Psychology,* vol. 9, 2018.

[5] C. Moore and P. J. Dunham, Joint attention: Its origins and role in development, Psychology Press, 2014.

[6] P. Mundy, J. Block, C. Delgado, Y. Pomares, A. V. Van Hecke and M. V. Parlade, "Individual differences and the development of joint attention in infanc," *Child developmen,* vol. 78, no. 3, pp. 938-954, 2007.

[7] N. Pfeiffer-Lessmann, T. Pfeiffer and I. Wachsmuth, "An Operational Model of Joint Attention--Timing of the Initiate-Act in Interactions with a Virtual Human," in *Proceedings of KogWis*, 2012.

[8] Y. Nagai, K. Hosoda, A. Morita and M. Asada, "A constructive model for the development of joint attention," *Connection Science,* vol. 15, no. 4, pp. 211-229, 2003.

[9] J. Triesch, C. Teuscher, G. O. Deák and E. Carlson, "Gaze following: Why (not) learn it?," *Developmental science,* vol. 9, no. 2, pp. 125-147, 2006.

[10] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience,* vol. 2, no. 3, pp. 194-203, 2001.

[11] A. Treisman, "The binding problem," *Current opinion in neurobiology,* vol. 6, no. 2, pp. 171-178, 1996.

[12] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction,* vol. 6, no. 1, pp. 25-63, 2017.

[13] F. Rea, G. Sandini and G. Metta, "Motor biases in visual attention for a humanoid robot," in *2014 IEEE-RAS International Conference on Humanoid Robots*, 2014.

[14] A. Kothig, M. Ilievski, L. Grasse, F. Rea and M. Tata, "A Bayesian System for Noise-Robust Binaural Sound Localisation for Humanoid Robots," in *2019 IEEE*

International Symposium on Robotic and Sensors Environments (ROSE), IEEE, 2019, pp. 1-7.

[15] D. Ognibene and G. Baldassare, "Ecological active vision: four bioinspired principles to integrate bottom--up and adaptive top--down attention tested with a simple camera-arm robot," *IEEE transactions on autonomous mental development,* vol. 7, no. 1, pp. 3-25, 2014.

[16] G. Baldassarre, W. Lord, G. Granato and V. G. Santucci, "An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning," *Frontiers in neurorobotics,* vol. 13, p. 45, 2019.

[17] C. L. Colby, "Action-oriented spatial reference frames in cortex," *Neuron,* vol. 20, no. 1, pp. 15-24, 1998.

[18] P. Fitzpatrick, G. Metta and L. Natale, "Towards long-lived robot genes," *Robotics and Autonomous systems,* vol. 56, no. 1, pp. 29-45, 2008.

[19] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardinoe and L. Montesanoe, "The iCub humanoid robot: An open-systems platform for research in cognitive development," *Neural networks,* vol. 23, no. 8-9, pp. 1125-1134, 2010.

[20] A. Butcher, S. W. Govenlock and M. S. Tata, "A lateralized auditory evoked potential elicited when auditory objects are defined by spatial motion," *Hearing research,* vol. 272, pp. 58-68, 2011.