The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Calibration of a Multimodal Head-Mounted Device for Ecological Assessment of Social Orienting behavior in Children

G. Schiavone, D. Campolo, F. Keller, E. Guglielmelli

Abstract—**In this work a multimodal head-mounted device for the assessment of social orienting behavior in children between 12 and 24 months is presented. The device is specifically designed to be used in poorly structured and uncontrolled environments such as day-care centers. Accordingly, a calibration procedure is described which fully exploits the multimodal approach and which is particularly suitable for an ecological assessment.**

I. Introduction

Human sensory-motor system develops during the first 22 months of age and represents a lifelong neurological foundation for the basic information processing of the brain, which supports social and cognitive development [16], [17]. Losses in the perception, processing, integration and interpretation of sensory information will automatically create serious functional problems and alterations in the development of higher cognitive and social skills [18]. In the domains of sensory-motor system a crucial role is played by the sensory-integration system. Over the past years, increasing research effort has been devoted to the study of multisensory interactions and their role for attention, perception, memory and behavior [19]. Dysfunctions of Sensory Integration (DSI) are Central Nervous System disorders characterized by imbalance among the primary sensations of sight, hearing, touch, taste, or smell, but also vestibular and proprioceptive senses [21]. Poor sensory integration has long been addressed as a cause of motor and social problems in developmental disorders such as Autism Spectrum Disorders (ASD) [20]. The most common characteristics of autism are absence or insufficiency of smiling, laughing, eye contact, limited communication abilities, insistence of routines and sameness, repetitive play, difficulties in pretending play, challenges in interacting with pairs. Failure in orienting towards occurring social stimuli (i.e. facial expressions, speech, gesture) represents one of the earliest and most basic social impairments in autism and may contribute to the later-emerging social and communicative impairments [22]. Early social exchanges require rapid shifting of attention between different stimuli. Because of motivational mechanisms or because social stimuli are complex, variable, unpredictable, children with autism may have difficulty integrating, processing and representing such stimuli, and therefore their attention is not naturally drawn to such stimuli. Impairments in social orienting can alter the developmental pathway of young children by depriving them of appropriate social stimulation [24]. In order to increase the probability of success of reeducation and/or rehabilitation therapies of children with ASD it is crucial to make a diagnosis at a very early age. Technologies now available for early diagnosis provide accurate measurements, they require well-controlled and highly structured environments (i.e., laboratories), artificial environments that can intimidate children and lead to diagnostic artifacts.

In this paper we present the Audio-Visuo-Vestibular Cap (AVVC) which is one of the diagnostic device developed within the TACT (Thought in Action) research Project, financed by the European Union's NEST/Adventure Program [23]. The AVVC is designed to assess sensory integration in social orienting behavior in very young children, from 12 to 24 months of age. It is a multimodal device which allows to monitor at the same time gaze, hearing and head orientation. Multimodality and semiautomatic data analysis are the main technological solutions proposed with the AVVC for the assessment of children behaviors in ecological environments.

This paper is structured as follows: in Section II the state of the art of current technologies for behavioral studies is presented; in Section III the functional and technical specifications for the AVVC device, the experimental scenario, and the system of processing are described in details; the calibration procedure for the AVVC is explained in Section IV, preliminary data on experimental sessions in a day care centre are shown in Section V, Section VI reports the conclusion.

II. State of the art

Current technologies used for sensory motor integration (in particular vision, auditory and motor systems) investigation are more often unimodal, meaning that only one feature is constantly monitored. Stereophotogrammetric systems, such as Motion Capture System from *Vicon*, based on optical devices and markers attached on the body are among the most popular devices used for gait and posture tracking. These systems have a high accuracy and they are useful to capture fast motion data for analysis. However, they have the drawback of being expensive and sensitive to environments. They also need a lot of data processing work. Hearing loss and lack of response to auditory stimuli are monitored with ABR (Auditory Brain Responses) audiometry [14] and OAE (otoacoustic emissions) tests [15],

This work was supported by a grant from the European Union, FP6-NEST/ Adventure program, no. contract 015636

G.Schiavone,F.Keller, E.Guglielmelli are with the Campus Bio-Medico University in Rome, 00128 Italy. D.Campolo is with the School of Mechanical&Aerospace Engineering Nanyang Technological University, 639798 Singapore. Corresponding author: g.schiavone@ unicampus.it

both the methodologies are implemented in specialized clinical centre and hospitals and therefore they are not directly available to continuously monitor the development of hearing functions. Several technologies, then, exists for recording eye movements. Magnetic scleral search coil [12] is the standard research technique providing the highest spatial and temporal resolution, and it can also detect torsional components, but it is limited to a clinical setting due to discomfort, limited recording time and risk of corneal abrasion or lead breakage. The requirements to stay in the centre of the magnetic field precludes the use of search coils during many natural activities. The standard clinical method for recording of eye movements is the electro-oculography, which allows also measurements with closed eyes but it suffers from low resolution, drift, noise, poor vertical measurements, and motion and EMG artefacts that limit its use during locomotion and other natural activities [1]. The most popular methods are IR oculography and video oculography. The first one uses infrared (IR) lighting to illuminate the pupil and then extract the eye orientation by triangulation of the IR spotlight reflections or others geometrical properties [13]. The major drawbacks of this methodology are the limited linear range, the complicated and time-consuming installation and calibration procedures and poor mechanical stability of the transducer with respect to the eye. Video oculography is an image-based method. There exists external video trackers systems, such as, the *Tobii* eye tracker, which are not wearable and thus, constrain considerably the experimental setup while being very sensitive to head movements. Other examples are described in details in [2-7]. These devices, however, are suited for adults, being too heavy and large for a child. A new head-mounted camera, the Wearcam [8], recently developed by the LASA of Ecole Politechnique de Lausanne (within the TACT project too), is specifically designed for children aged between 6 months and 18 months, to be used in a free-play environment. It films the frontal field of view of the child and a small mirror protruding from the bottom part of the camera reflects the eye portion of the wearer's face [9]-[10]. Compared to the state of the art the performance of the Wearcam seems very low; however it is suitable for data coming from unconstrained environments where the amount of movements or the quality of the image cannot be kept under check. Despite the presented technologies, the AVVC device offers a multimodal approach to systematic diagnosis of early social attention impairments. During experimental sessions, based on social interaction in ecological environments, it can collect several information about the child's behaviors. It works like an *artificial audio-visuo-vestibular* system that can detect sound sources close to the child, child's head orientation, his eyes movements and his facial expressions.

III. THE AVVC DEVICE

A. Functional and technical specifications

The AVVC device is designed in agreement with the following three principles: 1) non-obtrusivity; 2) minimally structured and ecological operating environments; 3) multimodality. The first one assures suitability for continuous monitoring without being distressful or obtrusive for children; this sets technical constraints for the design of the device, such as small in size, lightweight, and portability. The second one points out the field of application of the device, that is in unstructured home-like situations, which differs from laboratories and clinical centers environments. The third one stresses the demand for a complete and integrated analysis of the child behaviors from multiple point of views (i.e. different sensory features) at the same time. Other constraints on the design of the AVVC are associated to the processing system. Current tools ([32-33]) for coding videos and screening of recorded data result tedious and time consuming. Huge amount of data can arise from the collection of multimodal information. The aim of the AVVC is to provide a device able to acquire multimodal data and process them with automatic or at least semiautomatic modality. Therefore, simple and robust algorithms are required both for low level processing of the signals provided by each sensor and for data integration processing.

B. Experimental Scenario

The experimental scenario, that stimulate the child to perform clinically relevant sensory-motor tasks (i.e. orienting behaviors and attention tasks) has been designed, within TACT project, by Gunilla Stenberg of Uppsala University. It takes place in a room of a daycare center. The child is sitting on a chair, and two caregivers are sitting next to him/her, each one at the opposite side of a desk. In a first phase each caregiver, alternately, explains to the child what he is going to do. In a second phase each caregiver puts an object (three colored blocks, positioned one at a time) on the table in front of the child but far enough from the child avoiding he can reach it. The aim of the protocol is to investigate shifts in attention of the child from the object to the caregiver and viceversa, in presence or absence of speech. The child is expected to perform the following actions: i) look at the person who is speaking; ii) look at the person who is acting; iii) look at the objects on the table. The AVVC device, mounted on the child's head has to be able to localize the active sound source, that is the voice of the speaking caregiver, to detect the head orientation of the child and his gaze catching the 'child's point of view'.

C. Hardware and software design

For reasons, above specified, the cap is equipped with three different kinds of sensors (Fig. 1):

- a magneto/inertial sensor, sensitive to the orientation of the head in the space, which works like a vestibular system;
- a webcam, called eye-cam, able to detect both slow eye movements and facial expressions of the child;
- a pair of omni-directional microphones to binaurally (stereo) record sounds occurring around the child.

The sensors can be easily moved from one cap to another which better fits the cranial circumference of the child

(estimated from 35 cm to 49 cm for children from 6 to 24 months). The cap is kept fixed on the head of the child using adjustable elastic bands.

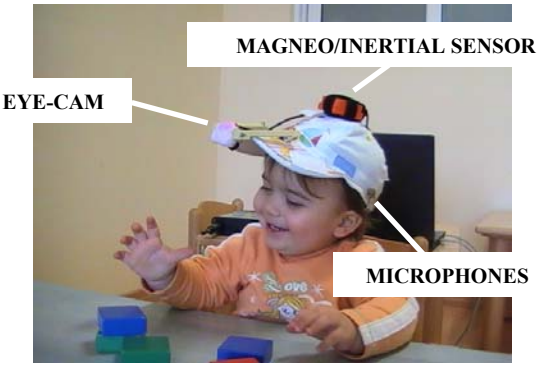

Fig. 1. The AVVC worn by a 18 months-old child.

Inertial/magnetic technology [25] for motion capture has been chosen for several reasons: it is sourceless, it relies solely upon gravitational and geomagnetic fields that are ubiquitously present on Earth and does not require additional field sources; it is available in compact packages, limited in dimension and weight; moreover such systems are easy to calibrate and low cost respect to other motion tracking systems. The main drawback of this technology is the sensitivity to external magnetic fields (i.e. mobile phones, power stations, etc.), however, it is plausible to assume that in the environments in which the AVVC is designed to be applied (daycare centers) electro/magnetic interferences are limited. The head tracker (MTx, Xsens Technologies B.V.), mounted on the top of the cap with velcro bands, transduces head orientation (azimuth, elevation, roll) in 3 dimension at a frequency of 100 Hz, with a dynamic range of all angles in 3D and angular resolution of 0.01° RMS.

Unexpensiveness, light weight, dimensions, and quality of modern day webcams allowing them to be used for eye-tracking have driven to choose them for our purpose. The eye-cam is composed of a ¼" CMOS sensor, with a resolution of 640x480 and a frame rate of 30 frames per seconds. It is positioned on the peak of the cap through an ad hoc light rubber support, designed to hold the camera at a distance of about 10 cm from the face of the child, interfering as less as possible with his field of view, and enabling manual setting of the orientation of the eye-cam. The eye-cam sensor points to the face of the child and it has been provided with a mini-objective of 2.5 mm focal length (Model RE-025S) in order to cover a field of view (the diagonal FOV is 84° for an average of 57° and 71° vertically and horizontally respectively), corresponding a face dimension of 12x15 cm.

Localization of the speaking voice is achieved by processing the signals from two microphones. The two microphones (MKE 2-ew Gold, Lavalier) are fixed to the cap with clips each one in correspondence of the two ears in order to simulate binaural hearing. The audio signal acquisition frequency is set to 44100 Hz, so as to cover an interval of frequencies from 0 to the Nyquist frequency 22050 Hz, in which is contained the frequency range of human auditory perception (20-20000 Hz). Quantization is set to 16-bit per sample, which provides a dynamic interval of 96 dB, close to human ear dynamic range.

All the sensors on the AVVC send data to a PC via USB, which means that the AVVC is still a wired device. The cables are connected together in correspondence of the back of the cap, before being plugged into the PC. The drawback of the wires is that the AVVC cannot be still used in free-play situations, in which the child can move freely, but he has to be sit on a chair during experimental sessions. On the other hand, the advantage of having a wired device is that it is lightweight and it does not require external and cumbersome batteries to energize the sensors. The AVVC device is, therefore, adequate to monitor one-to-one interaction during typical diagnosis tests [11] and in particular is suitable for the presented experimental scenario.

The low level processing of the data is made by processing separately the signals from the three sensors. The head rotation angle in the horizontal plane (head azimuth, ψ_h) is extracted by performing some computations on the rotation matrix provided by the magneto\inertial sensor. The sensor calculates the orientation between the sensor coordinate system, **H**, in agreement with the head, and a fixed reference coordinate system, **O**. The fixed reference coordinate system used is defined as a right handed Cartesian coordinate system with:

- X positive when pointing in the direction of the nose of the wearer.
- Y according to right handed co-ordinates.
- Z positive when pointing up.

The output provided by the sensor is in form of rotation matrix, $\mathbf{R_{OH}}$ (1), interpreted as the unit-vector components of the sensor coordinate system **H** expressed in **O**. The columns of the matrix $\mathbf{R_{OH}}$ are the unit vectors of **H**.

$$R_{OH} = [\, X_H \quad Y_H \quad Z_H \,] = \begin{bmatrix} x_H x_O & y_H x_O & z_H x_O \\ x_H y_O & y_H y_O & z_H y_O \\ x_H z_O & y_H z_O & z_H z_O \end{bmatrix} \qquad (1)$$

The azimuth of the head lies on the xy-plane of the fixed reference co-ordinate system and it is defined as the angle between the unit vector $\mathbf{X_O}$ of the fixed reference coordinate system and the vector **P**, projection of the $\mathbf{X_H}$ vector on the xy-plane of the fixed reference co-ordinate system. The vector **P** is the column vector defined as:

$$P = \begin{bmatrix} x_H x_O \\ x_H y_O \\ 0 \end{bmatrix} = \begin{bmatrix} cos(\theta)cos(\psi) \\ cos(\theta)sin(\psi) \\ 0 \end{bmatrix} \qquad (2)$$

The first two elements of the vector **P** are the first components of the unit vector $\mathbf{X_H}$ of the rotation matrix, where θ is the pitch, describing rotations around the axis $\mathbf{Y_O}$, and ψ is the yaw, describing rotations around the axis $\mathbf{Z_O}$. The azimuth of the head corresponds to the yaw and it is obtained by the trigonometric function *arctan2(y, x)*, where x and y are real arguments and not both equal to zero.

$$\psi = arctan\,2(y,x) = 2\,arctan\left(\frac{y}{\sqrt{x^2+y^2}+x}\right) \qquad (3)$$

By substituting the arguments x and y with the first two components of the vector **P** , the head azimuth is estimated in the range (-π, π]:

$$\psi = 2\arctan\left(\frac{\sin(\psi)}{1+\cos(\psi)}\right) \quad (4)$$

The signals provided by the two microphones are used to compute sound localization in the horizontal plane. Assuming that the distance between each observation point and the sound source is different, the sound waves produced by the source will arrive at the observation points at different time (time difference of arrival, TDOA) due to the finite speed of the sound. Most of the state-of-the-art sound localization systems rely on TDOA estimaton [26]. In order to determine the delay in the signal captured by the two microphones, a coherence measure has to be defined. The most common coherence measure is a simple cross-correlation [28] between the signals perceived by the two microphones. Considering windowed frames of N samples with 50% overlap, the cross-correlation function for the a single frame is expressed by:

$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n]\, x_j[n-\tau] \quad (5)$$

where $x_i[n]$ and $x_j[n]$ are the signals received by microphone *i* and microphone *j*, $\hat{o}$ is the correlation lag in samples (equally distributed from -1 and 1 ms, which means in samples -44< π <44). In order to reduce the computational load from a complexity of $O(N^2)$ to a complexity of $O(N\log_2 N)$ the inverse Fourier transform of the cross-spectrum is computed:

$$R_{ij}(\tau) \approx \sum_{k=0}^{N-1} X_i[k]\, X_j^*[k]\, e^{j2k\pi\tau/N} \quad (6)$$

where $X_i[k]$ is the discrete Fourier transform of $x_i[n]$and $X_i[k]X_j^*[k]$ is the cross-power spectrum of $x_i[n]$ and $x_j[n]$. The drawback is that (6) is strictly dependent on the statistical properties of the source signal. Since most signals, including voice, are generally low-pass, the correlation between adjacent samples is high and generates cross-correlation peaks that can be very wide. The problem of wide cross-correlation peak can be solved by whitening the spectrum of the signal prior to compute the cross-correlation [29]. The resulting whitened cross-correlation, also commonly referred to as Phase Transform (PHAT) technique, is:

$$R_{ij}^{PHAT}(\tau) \approx \sum_{k=0}^{N-1} \frac{X_i[k]\, X_j^*[k]}{\left|X_i[k]\right|\left|X_j[k]\right|} e^{j2\pi k\tau/N} \quad (7)$$

This approach allows to only take the phase of $X_i[k]$ into account, narrowing the wide maxima caused by the correlation between the received signals; it does not require any knowledge about the spectrum of the microphone dependent noises and shows good performance in low-noise, reverberative environments [30]-[31]. Although there exists more robust and accurate algorithms for sound localization, the PHAT algorithm has been chosen because of its simplicity and lower complexity. The TDOA, ΔT_{12}, for each time frame, between the two microphones, can be found by locating the peak in the cross-correlation function:

$$\Delta T_{12} = \arg\max_{\tau}\left(R_{12}^{PHAT}(\tau)\right) \quad (8)$$

As the TDOA has been obtained, the sound source location in the horizontal plane can be estimated theoretically by the model described in [27]. This model works well when the wavelength of the sound is higher than the dimensions of the head. For higher frequencies, the wavelength of the sound wave is smaller than the dimension of the head and other factors, such as the shadowing of the head, need to be considered. Moreover the child is free to orient his head in the space and the sound sources nearby the child (i.e. caregivers interacting with him) are not fixed. Thus an in loco calibration procedure (described in Session IV) is needed to correlate an estimated TDOA to a specific angle. The processing of the video recordings for the extraction of the pupil coordinates is not presented in this paper.

IV. THE CALIBRATION PROCEDURE

Two calibration procedures are required to calibrate the AVVC device: the first one is the *vestibulo-auditory calibration* which relates the estimated TDOA with the *binaural azimuth* (ψ_b), that is the angular position of the sound source in the horizontal plane; the second one is the *vestibulo-ocular calibration*, which allows to derive an angle of orientation of the eyes in the horizontal plane, the *visual azimuth* (ψ_e), from the pixel coordinates of the pupil. Both the vestibulo-auditory calibration and the vestibulo-ocular calibration are processes of the artificial sensory-motor integration system. In this paper only the *vestibulo-auditory calibration* is described, calibration data results for the *vestibulo-ocular calibration* will be presented separately in a future work.

The experimental set up for mapping of the TDOAs with angular positions of sound sources usually consists of a fixed set of observation points (array of microphones) and a fixed set of sound sources, located in known orientations with respect to the observation points. In an ecological environment, which is poorly structured, the typical experimental set up is difficult to reproduce, especially if the head of the child, where the microphones are mounted, is free to move. The proposed *vestibulo-auditory calibration* is properly designed for those environments. It exploits the free movements of the child's head to determine a correlation between the TDOAs and the sound sources directions. During the calibration procedure (Fig.2) the child is sitting on a chair at the long side of a desk while wearing the AVVC device. A caregiver (C1) is sitting in front of him, on the opposite side of the desk, and exhorts the child to orient the head toward a second caregiver (C2). C2 moves in a semicircle in the frontal space of the child, and captures his attention by holding a toy without speaking. C1 represents a fixed sound source in front of the child, while he/she is orienting the head from left to right and viceversa towards the toy. This procedure is specular to a setting in which the sound source moves in several positions around a

semicircle and the observation points are fixed in the middle of the semicircle.

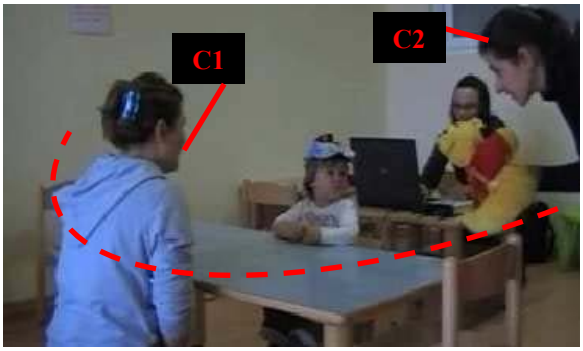

Fig. 2. Vestibulo-auditory calibration procedure in a room of the day-care center: C1, caregiver speaking in front of the child, C2, caregiver moving a toy in a semicircle in front of the child.

The calibration curve (Fig.3) is obtained by correlating the TDOAs estimated by the localization algorithm and the head orientation data provided by the magneto/inertial sensor. The curve shows a linear correlation between TDOAs and angles of rotation.

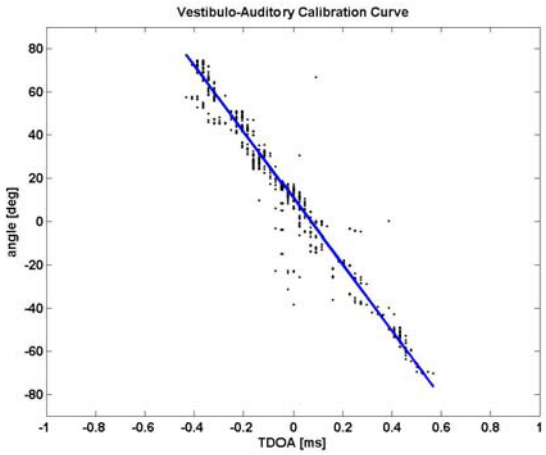

Fig. 3. Vestibulo-auditory calibration curve: black dots, raw data, blue solid line, fitted curve. TDOAs, in ms, are correlated to angular directions, in deg, by a linear relation with 95% confidence bounds (R-square, $R^2 = 0.98$)

The gain (G_b) and the offset (O_b) of the linear fitting are then used to estimate the binaural azimuth, ψ_b, according to the following relation:

$$\psi_b = G_b TDOA + O_b \quad (9)$$

The knowledge of the binaural azimuth is, then, used to determine the absolute direction of the active sound source ψ_{voice} (the speaking voice):

$$\psi_{voice} = \psi_h - \psi_b \quad (10)$$

According to the calibration procedure the sound source is expected to be estimated in a 0° direction, just in front of the child.

V. PRELIMINARY RESULTS

The calibration procedure has been tested on 11 healthy children between 12 and 24 months of age. Here preliminary results from the vestibulo-auditory calibration are presented. The Probability Density Function (PDF) of the sound source direction, ψ_{voice}, estimated for each calibration session, has been fitted using a Gaussian model, with 95% confidence bounds. In Fig. 4 the PDF of the estimated sound source direction for a the calibration session is shown. As expected, the probability to find the sound source, the caregiver' s voice, is maximum at about 0° (i.e. in front of the child).

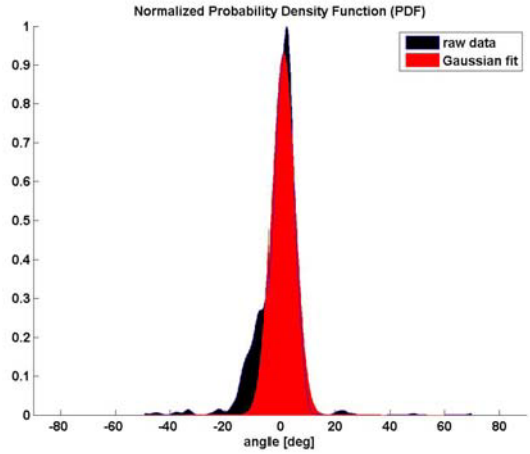

Fig. 4. Probability density function (PDF) of the sound source location estimated during the vestibule-auditory calibration. Black, raw data, red, Gaussian fit (mean = 0.37°; standard deviation = 16°, $R^2 = 0.98$).

The coefficients (mean and standard deviation) of the fitted PDFs, obtained for each subject, together with the R-square coefficients are listed in Table 1. R-square coefficients are close to 1 (see Table 1), thus confirming that the estimated sound source distribution is well fitted by a Gaussian process.

TABLE I

Subjects	Mean [°]	Std [°]	R^2
1	-0.34	17	0.98
2	0.89	24	0.96
3	0.23	21	0.97
4	-0.07	20	0.98
5	0.02	15	0.99
6	0.37	16	0.98
7	-0.01	20	0.99
8	1.09	18	0.98
9	0.01	15	0.99
10	0.09	15	0.98
11	-0.12	16	0.98

Mean, Standard Deviation (Std) and R-square coefficient (R^2) of the estimated PDFs are reported for each subject.

A T-test on the mean values of the PDFs (see Table 1) shows that there is not a statistically significant discrepancy among the subjects (*p-value*=0.16, α=0.05). Also standard deviation values, measure of the width of the bells, are not discrepant (*p-value*=1, α=0.05, the mean of standard deviation values is 18°). These first results confirm that the calibration procedure is consistent and reliable. The standard deviation is an index of the accuracy of the system. Compared to the human ability to localize sound source (1°-

5°), the accuracy of the AVVC device is very low. The dispersion of the PDFs is due to several factors: the performance of the algorithm used for sound localization in a noisy and un-controlled environment, such as a day-care centre, the head and trunk movement of the caregiver speaking in front of the child. Although these considerations, the accuracy of the AVVC is satisfactory for the designed experimental scenario in which the two sound sources are located in angular positions higher than 18° respect to the child.

VI. CONCLUSION

In this work a novel device for multimodal behavior assessment for children from 12 to 24 months of age is presented. The AVVC device is designed as an artificial audio-visuo-vestibular system: it can sense, process and integrate signals coming from different sensory channels. This can be useful for early diagnosis of DSI disorders by providing quantitative data on children orienting behavior in social situations (i.e. if and how the child orients to persons, to objects, to voices) and by providing, as a technological solution to the tedious videos coding, a tool for semiautomatic data analysis, which could help to reduce diagnostic time. While the device is less accurate with respect to current technologies for behavioral analysis, it offers the advantage that it can be used in unstructured and ecological environments. The calibration procedure for the device proved to be simple and it can be configured as a sort of a game play with the child. Also, the statistical analysis performed on the collected data showed the consistency and reliability of the designed procedure. We are currently working both for integrating the vestibulo-ocular calibration to the described procedure and for presenting the data collected during the experimental scenario.

ACKNOWLEDGMENT

The authors would like to thank the personnel of the daycare 'La Primavera del Campus' for helping with the preliminary tests, Gunilla Stenberg for the design of the experimental scenario and Mr. Luca Lonini and Mr. Emiliano Schena for their precious comments about this work.

REFERENCES

[1] A. DiScenna, V. Das, A. Zivotofsky, S. Seidman, R.J. Leigh, Evaluation of a Video Tracking device for measurement of Horizontal and Vertical Eye Rotations During Locomotion, submitted to Neuroscience Meth.

[2] R.S. Allison, M. Eiyenman, B.S.K. Cheung, IEEE transaction on Biomedical Engineering, Volume 43, Issue 11, Nov. 1996, pp. 1073-1082

[3] D. Winfield, Dongheng Li, J. Babcock, D.J. Parkhurst, Towards an open-hardware open-software toolkit for robust low-cost eye tracking in HCI applications, Iowa State University Human Computer Interaction Technical Report ISU-HCI, April 2005

[4] Dongheng Li, J. Babcock, D.J. Parkhurst, Proceedings of the 2006 symposium on Eye tracking research & application, pp. 95-100, San Diego, California, 2006

[5] Dongheng Li, D. Winfield,D.J. Parkhurst, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 3, pp. 79, June 2005

[6] A. Iijima, M. Haida, N. Ishikawa, H. Minamitani, Z. Shinohara, Engineering in Medicine and Biology Society, 2003, Proceedings of the 25th Annual International Conference of IEEE, Volume 4, 17-21, Sept. 2003, pp 3225-3228.

[7] R. el Kaliouby, A. Teeters, R.W. Picard, in Proceedings of International Workshop on Wearable and Implantable Body Sensor Network, BSN 2006

[8] L. Picardi, B. Noris, G. Schiavone, F. Keller, C. Von Hofsten, and A. G. Billard, In RO-MAN '07: Poceedings of the 16th International Symposium on Robot and Human Interactive Communication

[9] B. Noris, K. Benmachiche, J. Meynet, J.-P. Thiran and A. Billard, Computer Recognition Systems, 2(2007), 663-670

[10] B. Noris, K. Benmachiche and A. Billard, In Proceedings of the International Conference on Computer Vision Theory and Applications, (2008).

[11] P.A. Filipek *et al*., Journal of Autism and Developmental Disorders, Vol. 29, Issue 6, pp. 439-484, 1999

[12] D. A. Robinson, IEEE Trans. Biomed. Eng., vol. BME-l0, pp. 137-145, 1963

[13] T.N. Cornsweet, H.D. Crane (1973), J Opt Soc Am. 63, 921-8.

[14] A. Davis, J. Bamford, I. Wilson, T . Ramkalawan, M. Forshaw ,S. Wright, Health Technol Assess. 1997;1(10).

[15] D.T. Kemp, J Acoust Soc Am. 1978;64:1386–1391

[16] J. Piaget, The origins of intelligence in children, International Universities Press, NewYork, 1952

[17] J. Ayres, Sensory Integration and the Child, Western Psychological Services, Los Angeles, California, 1970

[18] M. S. Williams, S. Shellenberger, How Does Your Engine Run? A Leader's Guide to The Alert Program for Self-Regulation, TherapyWorks, Inc., Albuquerque, NM, 1996

[19] G. A. Calvert et al., The handbook of multisensory processes, Massachusetts Institute of Technology, Calvert, G. A. and Spence, C. and Stein, B. E., Cambridge MA, 2004

[20] C. Trevarthen and S. Daniel, Brain and Development, n. 27, pp. S25-S34, 2005

[21] C. S Kranowitz, The Out-of-Sync Child: Recognizing and Coping with Sensory Integration Dysfunction, The Berkley Publishing Group, New York, NY, 1998

[22] G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw, Developmental Psychology, 2004, Vol. 40, No. 2, 271–283

[23] D. Campolo, C. Laschi, F. Keller, E. Guglielmelli A Mechatronic Platform for Early Diagnosis of Neurodevelopmental Disorders, on RSJ Advanced Robotics Journal, Vol. 21, No. 10, pp. 1131-1150, 2007 .

[24] P. Mundy & R. Neal (2001), Neural plasticity, joint attention and a transactional social-orienting model of autism. In L. Glidden (Ed.), International review of research in mental reterdation: Vol. 23, Autism (pp. 139-168), New York: Accademic Press

[25] G. Welch, E. Foxlin, Motion Tracking: No Silver Bullet, but a Respectable Arsenal, Motion Tracking Survey, IEEE Computer Graphics and Applications, November/December 2002

[26] P. AArabi, Multi-sense artificial awareness, M.A.Sc. Thesis, Department of Electrical and Computer Engineering, University of Toronto, Ontario, Canada, 1998

[27] M. S. Brandstein and H. Silverman, in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 375-378, Munich, Germany,April 1997

[28] C. H. Knapp and G. Carter, IEEE Trans. Acoustics, Speech and Signal Processing, vol. 24, no. 4, pp. 320-327, 1976

[29] M. Omologo and P. Svaizer, in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. II-273-II-276, 1994

[30] K. D. Donohue, J. Hannemann and H. G. Dietz, Signal Processing, Volume 87, Issue 7, July 2007, Pages 1677-1691

[31] C. Zhang, D. Florencio, Z. Zhang, IEEE Int.. Conf. on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. Volume , Issue , March 31 2008-April 4 2008 Page(s):2565 – 2568

[32] M. Kipp, Anvil-a generic annotation tool for multimodal dialogue. In Proc. Eurospeech., 2001

[33] D. Reidsma, N. Jovanovic, and D. Hofs, Designing annotation tools based on properties annotation problems, In Measuring Behavior 2005, 5th Int. Conf. On Methods and Techniques in Behavioral Research, 30 August-2 September 2005.