IDIAP RESEARCH REPORT

THE VERNISSAGE CORPUS: A MULTIMODAL HUMAN-ROBOT-INTERACTION DATASET

Dinesh Babu Jayagopi Samira Sheikhi[a]
David Klotz Johannes Wienke Jean-Marc Odobez
Sebastian Wrede Vasil Khalidov
Laurent Son Nguyen[b] Britta Wrede
Daniel Gatica-Perez

Idiap-RR-33-2012

DECEMBER 2012

[a]Idiap Research Institute
[b]Idiap Research Institiute, EPFL

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

The Vernissage Corpus: A Multimodal Human-Robot-Interaction Dataset

***Abstract*—We introduce a new multimodal interaction dataset with extensive annotations in a conversational Human-Robot-Interaction (HRI) scenario. It has been recorded and annotated to benchmark many relevant perceptual tasks, towards enabling a robot to converse with multiple humans, such as speaker localization, key word spotting, speech recognition in audio domain; tracking, pose estimation, nodding, visual focus of attention estimation in visual domain; and an audio-visual task such as addressee detection. Some of the above mentioned tasks could benefit from information sensed from several modalities and recorded states of the robot. As compared to recordings done with a static camera, this corpus involves the head-movement of a humanoid robot (due to gaze change, nodding), making it challenging for tracking. Also, the significant background noise present in a real HRI setting makes tasks in the auditory domain more challenging. From the interaction point of view, our scenario, where the robot explains paintings in a room and then quizzes the participants, allows to analyze the quality of the interaction and the behavior of the human interaction partners.**

I. Introduction

Automatic audio-visual perception of people, that includes tasks such as tracking their location, when they speak, gesture, what they look, whom they address is a relevant problem in many applications ranging from surveillance, smart rooms, telecommunication systems, human behavior understanding, human-robot and human-computer interactions. The challenge in HRI is in providing humanoid robots with the audio-visual perception capabilities to interact with multiple human partners [1], [11]. Towards this goal, new methods in processing unimodal and multimodal data need to be developed and existing methods have to be adapted and redesigned to meet the challenges that accompany recording with the audio-visual sensors on a humanoid robot. In order to be a realistic interaction partner, a humanoid robot needs to perform appropriate actions, for example nodding, or gaze changes which affects the sensing process. Also, the sensing, computing, and communication capabilities on the robot are limited and constrain each other. In contrast, recordings of human-human interactions (HHIs) typically involve stationary cameras or microphone arrays well-placed in the environment where the interaction takes place ([9], [7], [8], [25], [27]), or wearable sensors without the possibility of visual processing [16].

According to our knowledge, the existing HRI datasets in robot perception mainly focus on visual object recognition and navigation (e.g. [32]). Among those corpora that have focused on audio-visual perception tasks [21], [3], [2] in a conversational scenario, none of them have all the advantages of our dataset: an interesting scenario, more than one interaction partner, a commercially available robot (with consumer sensors rather than high-end sensors), extensive annotations, and planned public availability.

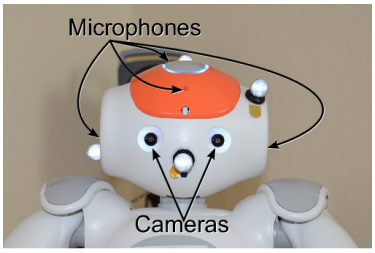

Figure 1. The humanoid robot NAO and its primary sensors used for the recordings. VICON markers (silver balls) for motion capturing are visible.

Our dataset (called 'The Vernissage[1] corpus') has 13 sessions of a humanoid robot, NAO[2] (cf. Fig. 1), interacting with two persons. The robot serves as an art guide, explaining the paintings to the participants and then quizzing them in art and culture. The scenario involves a stationary robot, exhibiting significant gesticulations to facilitate the interaction, for example, turning its head, nodding, and moving the hand towards objects of interest. A wizard-of-oz was used to manage the dialog as well as the robot's gaze and nodding. The behavior of the human partners was unconstrained. Each interaction lasted around 11 minutes. Fig. 2 gives an overview of the corpus.

The dataset comprises of a synchronized multimodal corpus, with multiple auditory, visual, and robotic system information channels. The recording method is inspired by the SInA method proposed in [19] which focuses on synchronizing internal logging data with external manually annotated data in order to analyze specific issues of HRI. NAO video data is mainly mono at VGA resolution and audio data comes from four microphones. In order to have ground-truth information for all the audio-visual processing tasks, 3 close-field external cameras, a motion capturing system and close-talk microphones on the human interaction partners were also deployed.

Apart from the richness of the sensor data, the robot system data includes the robot's 2D location of its body, joint angles, wizard commands, internal events for speech and gesture production, usage of CPU, memory, and battery. The corpus is annotated with speech utterances, speech transcription, 2D head-location, nodding, visual focus of attention, and addressees. The motion capturing system gives the 3D location of the participants and their head-pose. In this paper, we present

[1] *vernissage*: French for the opening of an art exhibition
[2] http://www.aldebaran-robotics.com

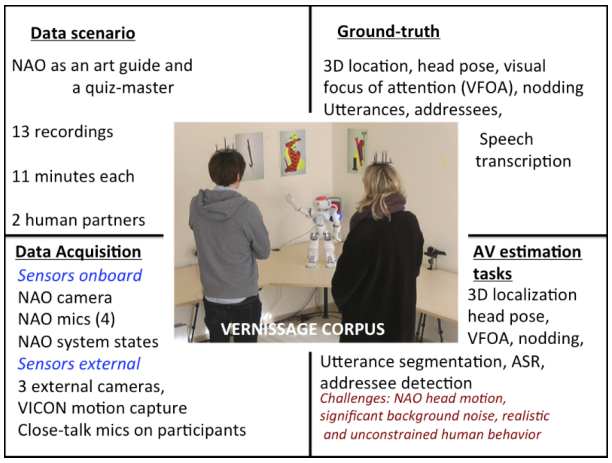

Figure 2. Overview of the Vernissage corpus: scenario, various modalities, annotations, and possible audio-visual perception tasks.

some statistics of the annotations, characterizing the nonverbal cues of human participants to understand the corpus better.

This corpus is a relevant contribution for the multimodal perception community. While datasets exist for benchmarking a single perceptual task (e.g. Rich Transcription Evaluations, [28], [23]), our dataset allows benchmarking several perceptual tasks on a single dataset. According to our knowledge, it is the first publicly available multimodal HRI dataset with extensive annotation of verbal and nonverbal cues relevant to multiparty interactions. As compared to the existing HHI datasets [7], [9], [8], [25], [20], [27], the moving robot camera and natural acoustic background provide significant challenges to audio-visual perception tasks that are a key factor for multiparty interactions. Compared to existing HRI datasets [32], [12], [21], [3], [2], this corpus focuses on multiparty conversational issues. Also, NAO is a commercially available and state-of-the art humanoid robot, and hence facilitates repeatability and comparison of experiments. We believe that the scenario, where NAO acts as an art guide and a quiz master, is both an interesting and reasonably controlled application scenario of humanoid robots.

Section 2 presents related works. Section 3 introduces the scenario of our data collection. Section 4 introduces the data acquisition setup. Section 5 presents the annotations available for the Vernissage corpus. Section 6 discusses the perception tasks that could be attempted on this corpus. Section 7 provides conclusions.

II. Related work

Three kinds of corpora relate to our corpus. First, those corpora focussing on a single perceptual task. Second, those corpora that have studied HHI with annotations for multimodal cues. Third, those HRI corpora that have used a moving camera and microphones, albeit with a different recording scenario.

Focussing on individual perceptual tasks several benchmarking datasets have been collected and used. As this list is quite extensive, we point to a few of them as samples for comparison. Rich Transcription Evaluations [3] have been the traditional benchmark dataset for speaker diarization and ASR studies. Keshet et al have proposed a database [4] specifically for key-word spotting on clean read speech, non-clean read speech and spontaneous speech [15]. Two benchmarking resources for speaker localization are the av16-3corpus [5] [18] and the CHIL corpus [23]. To benchmark face tracking and head-pose estimation methods, the CLEAR evaluation dataset is a good source [28]. Multiperson head-tracking could be evaluated with the BoBoT dataset [6] or the TA2 dataset [7] which includes 2D bounding boxes for people.

Several multimodal, annotated corpora for studying multiparty human interactions exist. They differ in the scenarios and the annotations available. Notably, the ISL corpus [7] collected real and scripted meetings on scenarios such as project planning, military exercises, games, chatting and discussion. The aim of the ISL corpus is to distinguish between different kinds of meetings by characterizing speaking styles. The VACE meeting corpus [9] has been recorded using real-world scenarios (war games and military exercises). The annotations include conversation transcripts, dominant speaker, floor control data, and gesture. The annotations include word tokens, turns, question/non-question, disfluency. For the AMI corpus [8], the meeting participants have assigned roles and work together to design a remote control. Annotations for conversation transcript, addressees, gaze direction, adjacency pairs (question-answer, statement-agreement), dominance ranking (inter-ranking), hand and head gestures exist for different subsets of this corpus. MSC (Mission survival corpus) and the ELEA (Emergent Leadership) corpus have groups solving the winter survival task. While MSC [25], [20] has annotations for task area and socio-emotional functional roles, some aspects of personality, group cohesion, individual and group performance, ELEA [27] has annotations for personality, interpersonal perception (e.g. dominance, leadership, liking, competence), and performance.

Regarding multimodal and annotated HRI datasets, [32] was collected with omnidirectional cameras, laser scanner, odometer and sonars. It was collected in three different home environments, with the purpose of facilitating goal-directed navigation, including localization, path planning, path following, and object categorization. [12] is a multimodal corpus with annotations for gestures and conversational acts in a home-tour scenario. [21] is an interesting dataset with 22 natural HHIs, 22 unnatural HHIs, and 22 human-robot interactions. The scenario is dyadic and involved an instructor and a listener. In what the authors call the natural HHIs, the listener tries to be engaged with the instructor. In the unnatural HHIs, the listener tries to be distracted and uses abnormal nonverbal interaction protocols. The multimodal data collected include physiological data (skin conductance, blood

[3] http://www.itl.nist.gov/iad/mig/tests/rt/
[4] http://ttic.uchicago.edu/~jkeshet/Keyword_Spotting.html
[5] http://www.idiap.ch/dataset/av16-3
[6] http://www.eecs.qmul.ac.uk/~andrea/PHD-MT.html
[7] http://www.idiap.ch/dataset/ta2

volume pulse, respiration, audio from two microphones, four network cameras, and motion tracking data (using 22 markers on every participant). In [3], a human-centered audio-visual dataset was collected to study the effects of head-movement on existing audio-visual perception methods. A helmet with a stereoscopic camera, binaural microphones, and head-tracking markers was worn by a human or a dummy mannequin to collect the dataset. The scenario is scripted with multiple short sub-scenarios to test different audio-visual perception tasks such as speaker localization. In [2] a multimodal dataset was recorded using a robot head, equipped with two cameras and four microphones. The scenario includes performing certain recognizable actions (e.g. walking, drinking), gestures (e.g. waving, pointing), and certain types of interactions where the robot head interacts with people. Notably, these were short interactions with a fixed-script for both the robot and the humans. Finally, a relevant work that used the NAO robot and shares many data acquisition methods as ours, but used in a soccor-playing scenario is described in [22].

As compared to the existing corpora, the Vernissage corpus has an interesting conversational scenario and comes with much richer annotations and ground-truth from the external sensors and robot internal states. This allows researchers in multimodal perception community to investigate interaction behavior cues at a low level (such as 'who is speaking', 'who is looking at whom', 'nodding') as well as at a higher level (such as 'who is being addressed', turn-taking or conversational behavior).

III. Data scenario

In order to capture a dataset that can be used for HRI analysis as well as to test various audio-visual perception techniques, we decided to choose the Vernissage scenario, where the robot serves as a conversational partner in a reasonably realistic application setting: as an art guide and a quiz master. This scenario offers sufficient flexibility as well as control over the human-robot interaction. More explicitly, it allows for a continuous change in difficulty w.r.t.

- mixed initiative (can range from robot initiative, i.e. just monologues to really mixed)
- speech input (users are not expected to speak, or simple yes/no suffice, up to really sophisticated free questions or answers are possible)
- number of people / complexity of groups
- text to be told by robot
- adaptation capabilities of robot (from none to user- to group-specific adaptation)

This scenario was inspired by a recent work that has studied and documented human interaction experiences with NAO as an art guide in a German art museum [26]. In this scenario, as the robot is stationary, the complexity involved in adapting and extending existing perception methods is reasonable, but still challenging.

In the Vernissage corpus, the scenario unfolded exactly as follows:

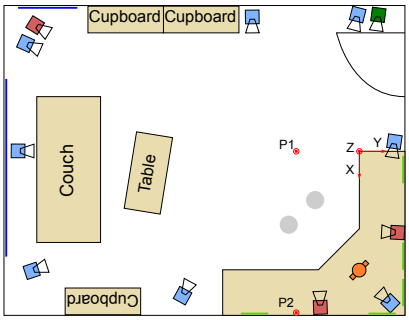

Figure 3. Qualitative overview of the recording room. Orange: NAO, cameras: blue – VICON, red – HD, green – wizard feedback, green lines: paintings, blue lines: windows, red: VICON coordinate system, P1 and P2: position for head pose calibration, gray circles: typical positions of participants

- The visitors arrived in pairs and were greeted by the robot when they entered within a normal interacting distance (Fig. 3 illustrates the top-view of the recording room). After this greeting, the robot offered some explanations about the paintings present in the Vernissage.
- When the visitors agreed to this, the robot started explaining three different groups of paintings using speech and matching gestures. These explanations included pauses intended to elicit comments by the visitors and also gave them the chance to tell the robot if they wanted to hear another explanation at specific points.
- When the explanations were finished, the robot asked the visitors if they were interested in participating in a quiz. After they agreed to this, NAO introduced itself and asked each participant to give their name and to introduce themselves.
- The robot then explained the general quiz rules which included that the visitors should discuss among themselves before giving the answers. The robot then proceeded to ask several questions about the paintings and more general topics and also judged the answers given by the participants.
- After the quiz was finished, the robot asked each visitor to decide on a favorite painting and afterwards told the participants to discuss and choose one common favorite and also to propose a new fitting name for that painting.

The participants spoke in English and they were mostly non-native speakers recruited in a university environment.

Wizard-of-Oz. To govern the behavior of the robot in a repeatable fashion, we used a "Wizard-of-Oz" (WOz) approach [10]. This means that the robot was not acting autonomously, but instead was controlled by human operators. For our recordings, we mostly used two operators (or "wizards"), which worked in a separate room hidden from the participant's view. One operator controlled the utterances and associated gestures of the robot by choosing them from a predefined set of buttons. Limiting the set of possible robot utterances like this was meant to reduces the gap towards an autonomous system with a real dialog engine. The second operator controlled the viewing direction of the robot by choosing points in the live streamed camera image, causing the robot to turn its head

in that direction. In addition to these specialized interfaces, both operators also had access to the sound coming from the microphones the participants wore and the live image of an external camera providing an overview of the interaction. To facilitate later analysis, the button clicks from the wizard interface were also logged as part of the corpus.

IV. Data acquisition

Recording Hardware. The target platform for the recordings was a NAO robot with a modified head containing improved cameras in a stereo setup and an ATOM processor (cf. Fig. 1). With respect to the processor and the camera type this head equals the recent "V4.0" version. We used only a single camera with VGA resolution due to technical restrictions. This setup is replicable with every "V4.0" NAO.

In addition to the robot system, supplemental devices were recorded to provide ground truth data or facilitate the annotation process. Each participant had to wear a close talk microphone for high-quality recordings of produced utterances. Furthermore, 3 HD cameras were used to obtain external perspectives for annotation. In order to get ground truth information about the locations and head orientations of each participant, we used a VICON motion capturing system[8]. This system is based on reflecting artificial markers. It produces 6D measurements at high frame rates (in our case 100 Hz). We used rigid bodies [6] to obtain good tracking results without needing a training phase for each participant. Comparable to the microphones, the rigid bodies were attached to hard plastic head bands, hence keeping he face free from artifacts. Please refer to Fig. 3 for the arrangement of recording devices.

The aforementioned setup was supported by a distributed computer system connected via gigabit Ethernet links.

Acquisition Software Setup. We decided to use the RSB middleware [30] as the primary recording tool as it provides a hierarchical bus architecture in combination with record-replay tools. This makes it easily possible to a) record the whole system communication including all sensory modalities, b) replay the recorded data in the system without needing modification to processing components, and c) acquire a synchronized dataset without post-processing, as transfered data is accurately timestamped. As a consequence, all communication of the robot system, including the "Wizard-of-Oz" commands, is contained in the corpus. To achieve consistent timestamps across the distributed system NTP[9] was employed.

We reused this functionality also for the recording of the close talk microphones. Due to interface restrictions, the HD cameras as well as the VICON motion capturing system could not be recorded in the RSB middleware and hence required a manual synchronization in a post-processing phase, which is explained subsequently.

Please refer to Table I to get an overview of the concrete data recorded using the aforementioned solutions.

[8]http://www.VICON.com

[9]http://www.ntp.org

Type	Specification
NAO video	Monocular / stereo uncompressed frames, VGA, variable frame rate (∼15 fps mean), YUV422 color mode.
NAO audio	4 channels, 48000 Hz, 16 bit signed.
NAO odometry	est. 2D location of robot body
NAO proprioception	Joint angles, stiffness, last command value, temperature
NAO system	CPU, memory, battery, modules
NAO system and control	Wizard commands, internal events for speech and gesture production
close talk mics	4 channels, 44100 Hz, 24 bit signed
VICON	6D pose for people and NAO, 100 Hz
External HD Cameras	3 perspectives, 1920 × 1080 pixels, 25 Hz. 5.1 channel sound, 48000 Hz.

Table I
RECORDED DATA. ITALIC: RECORDED USING RSB.

Synchronization and Calibration. To generate synchronized video and audio files from the external HD cameras with respect to the RSB-recorded data a synchronization method based on determining the cross-correlation peak between the sound channels of the different video cameras and the audio recorded from NAO was realized using the Praat [5] tool. Since the latter stream was recorded in RSB with timestamps, this allowed us to obtain timestamps w.r.t. the RSB-recorded data for the external videos. Conversion of the recorded data to formats compatible with Praat was performed using a combination of RSB-based tools and plugins developed for GStreamer[10]. The rear HD camera was at an approximate distance of 5 m from the main scene. Therefore, the possible synchronization accuracy is limited by the speed of sound for this distance: $5m/(343.2m/s) = 0.0146s$.

To synchronize the external VICON recordings with the RSB data we used a clapperboard equipped with VICON markers. The closing clapperboard was automatically extracted using a sliding-window approach on the distance of the upper and lower parts and related to the sound event in the microphones, which have timestamps through RSB.

To get calibrated data, we presented a checkerboard pattern to all cameras. Moreover, we presented the calibration device of the VICON system (consisting of several reflecting markers) to all three HD cameras while tracking it in the VICON system. Hence, the HD cameras can be registered in the VICON coordinate frame (not validated).

As each participant was wearing the hard plastic head band with the rigid body in a different configuration we included a calibration phase after each run. In this phase each participant stood at a fixed position and looked straight at a predefined spot (P1 and P2 in Fig. 3) so that a neutral head pose was recorded. We again presented the clapperboard to the VICON system and used the aforementioned detection method for automatic extraction. Additionally, the positions of P1 and P2, paintings, and the robot were recorded using VICON for being able to relate head poses to the scenery.

Annotation Tool Use. For timeline-based annotation we used ELAN [31]. The previously synchronized data was exported

[10]http://gstreamer.freedesktop.org

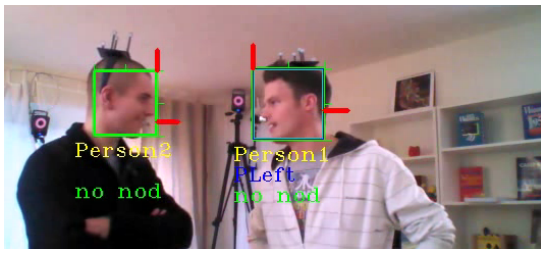

Figure 4. Visualization of annotations: head-pose (pan above and tilt in right of the head bounding box), VFOA (in yellow), addressee in blue (displayed when speaking), and nodding in green. When speaking, bounding box color is partially blue. Videos available as additional material to the paper.

to ELAN projects using automated solutions. These projects contain audio and video tiers. Furthermore, motion capturing and proprioception data can be visualized in ELAN using the "TimeSeries" view.

V. Data Annotation

Two types of annotations or ground truth (GT) will be made available with the dataset: first, the GT data automatically generated by the recording process, which comprises 3D head-location, head-pose, and NAO system data. Second the manual annotations which includes several important cues to study the HRI process and analyze the verbal and nonverbal behavior patterns (Fig. 4). In the following, we motivate the annotation of the selected cues, and then describe and comment the labeling procedure, some statistics about the labels, and the reliability across coders for nods, VFOA, and addressee annotations.

A. Utterance

Motivation and Label Set. In order to evaluate speech and speaker turn detection and enable addressee labeling and speech transcription, we decided to label the silence segments and speech segments (the utterances) of the audio channel. An utterance is the basic speech unit and following the literature on addressee detection, we defined it as 'a speech turn followed by silence more than 0.5 seconds' (e.g.[29]). We decided to also include a 'Laughter' label to differentiate actual speech turns and laughter, so our three labels were. were *Speech*, *Silence*, *Laughter*.

Annotations and Statistics. As the task of manual segmentation and then assigning a label is quite cumbersome, we used a semi-automatic approach. We started with an automatic method (speech activity detection by cross-talk suppression) to obtain the speech/silence segmentation. Then an annotator revisited and adjusted the segmentation and labels. This process was carried out using ELAN. Each recording has an average of 60 utterances. The average duration of an utterance being 1.3 seconds.

B. Speech Transcription

In addition to the segmentation of the human utterances into the broad categories of *Speech*, *Silence* and *Laughter*, we have also transcribed what exactly was said by the participants during the interaction. Besides the evaluation of algorithms

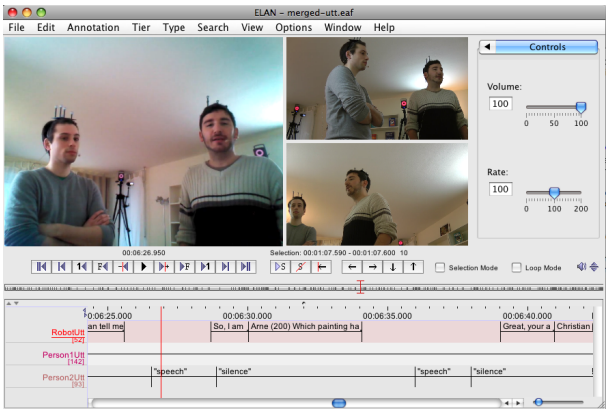

Figure 5. Screenshot of the ELAN annotation tool [31] showing participant utterance segmentation & robot utterances.

for ASR tasks, keyword spotting or dialogue control, this also enables understanding how the content of the spoken interaction affects other behavioral analysis tasks like addressee detection. Although the participants were instructed to talk only in English, they sometimes switched between English and German (their native language) when talking among each other. The annotation marks marks those segments where the utterances are in German.

For the text that was said by the robot, we generated annotation tiers in ELAN with transcriptions automatically by exploiting the communication between the WOz interface and the text-to-speech and gesture system which was also recorded including timing information as part of the dataset. Fig. 5 shows an example of the human utterance segmentation describe earlier in combination with this transcription of the robot's utterances.

C. Head Location

Motivation. We annotated the image location of people in the recording. This allows to evaluate tracking algorithms under two following challenging conditions:

- Given our scenario, there are several elements of attraction besides NAO other people, paintings), resulting in an interesting variety of head poses that makes tracking challenging.
- By contrast to standard static camera systems which are used when using Avatars in a Human Computer Interaction (HCI) scenario, in natural conversations NAO moves his head (for nodding, or for head deictic gestures), which generate large motions of people in the video recordings. Also, the people disappear and re-appear frequently. The median number of head motions (including both gaze-shift and nodding) was 76. This generates challenging situations for trackers who need to quickly re-acquire a person track but cannot necessarily rely on frontal head pose detectors if, for example, people are currently looking at a painting and are seen from profile.

These issues have not been dealt with much in the literature, and we believe that a public dataset addressing them would

	Mean	SD	Min.	Max.
Obvious	8.15	6.92	0	22
Subtle	8.54	3.38	3	15
Total	16.69	8.64	3	34

Table II
STATISTICS OF THE NUMBER OF HEAD NODS PER PARTICIPANT FOR 13 SEQUENCES OF THE VERNISSAGE SCENARIO, ANNOTATED BY THE PRIMARY ANNOTATOR. HEAD NODS ANNOTATIONS WERE SEPARATED INTO TWO CLASSES, *obvious* AND *subtle* NODS.

be valuable for the research community.

Annotation. Exploiting the VICON 3D location data was not possible since it did not localize the 2D head bounding box in the image captured by NAO as NAO's camera is constantly changing its orientation and position. We thus resorted to simple manual annotation of the visibility, ID, and position (bounding box) of each person. Annotation was done at 1 frame per second. Interpolation was automatically generated, and manually revised (i.e. intermediate frames were annotated) whenever these interpolations, displayed on the image, deviated too much from the true head location, to have sufficient accuracy at important transition points.

D. Head Nods

Motivation and Label Set. In human face-to-face communication, nonverbal behavior is a major mode of communication as it provides information in parallel to the spoken language [17]. Providing a humanoid robot with the capacity to properly detect these nonverbal cues is an essential step towards the goal of natural interactions with humans, and a first step for the inference of higher-level social constructs such as the engagement or the interest of the participants. Head movements and in particular head nods are one important cue involved in this communication channel. Head nods are defined as vertical up-and-down movements of the head rhythmically raised and lowered and fulfill various functions: signaling 'yes', displaying interest, enhancing communicative attention by occurring in synchrony with the other's speech, or anticipating an attempt to capture the floor (i.e. signaling a turn claim) [13]. In order to benchmark a head nod detection method, annotations were made on the Vernissage data. Depending on the amplitude and duration of the up-and-down oscillatory movements, head nods can be difficult to code; two classes of head nods were therefore defined: *obvious* and *subtle*.

Annotations and Statistics. The annotation of head nods were completed on the full dataset by one person who noted the on-set an off-set time of an event, and qualitatively decided the nod class based on nod amplitude and duration. Table II displays the statistics of the head nod annotations, completed on 13 sequences of the Vernissage scenario. The average duration of a head nod is approximately 1.09 seconds. In this dataset, we observe significant variability of nodding behavior across participants: some never nodded while others produced a high number of head nods. In total, 111 *obvious* and 106 *subtle* nods were annotated. Although these numbers are relatively small, they are sufficient to benchmark a head nod detection method in a realistic HRI context (illumination changes, camera motion, non-frontal nods).

	Obvious	Subtle	Not Annotated
Obvious	75	15	6
Subtle	16	21	41

Table III
CONFUSION MATRIX OF ANNOTATED HEAD NODS FOR 8 VIDEOS.

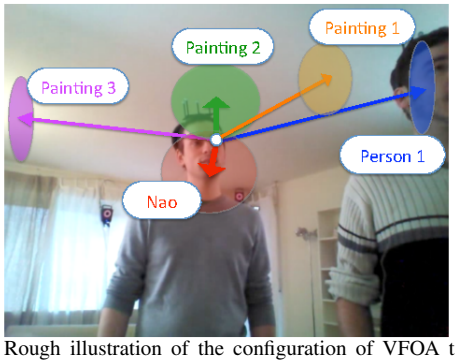

Figure 6. Rough illustration of the configuration of VFOA targets in the scene.

In order to evaluate the reliability of the annotations, eight videos were annotated by a secondary annotator. Interrater agreement was studied using a confusion matrix constructed as follows: for each nod annotated by one annotator (*subtle* or *obvious*) we evaluated whether there was a match with the second annotator, and in the positive case, what was the assigned label. Table III displays the resulting confusion matrix. It shows that head nods are difficult to annotate. As expected, most part of the confusion comes from the distinction between obvious vs subtle, and subtle vs not annotated. However, only 8% of the nods labeled as obvious are not annotated by the second annotator; therefore, nods labeled as obvious can be trusted.

E. Visual Focus of Attention (VFOA)

Motivations. As with nods, gaze is another important nonverbal communication cue which helps to keep the smoothness of HHI. It has functions such as establishing contact (through mutual gaze), display of attention towards the speaker, and floor control. Besides communication, gaze plays also a role of displaying the current attention of a person. For instance, in our scenario, when NAO describes a painting, one could expect participants to look at this painting, hence showing that they have followed the explanations given by the robot. Recognizing the VFOA is thus fundamental for a robot to understand how the interaction is going, monitor the engagement of participants, and react appropriately. It is however a difficult task that involves gaze tracking (or head pose as a proxy given the available head image resolution) as well as context tracking (conversation state like speakers, probability of end of turns, topic of conversation) to identify and timely reduce the number of relevant potential gaze targets as shown in [4]. The latter is essential for removing the potential ambiguities arising from only taking into account the plain geometrical gaze direction.

Labels and Annotations. Given the scenario, 5 main VFOA

Label	NAO	OP	Pai1	Pai2	Pai3	OT	NV	DK
FraFreq	0.43	0.11	0.06	0.14	0.06	0.09	0.06	0.05
EvFreq	0.25	0.10	0.06	0.14	0.06	0.17	0.04	0.19
AvgeDur	78.9	54.1	44.3	47.2	48.4	25.7	67.6	11.9

Table IV
VFOA FREQUENCY IN PERCENTAGE OF FRAMES (FRAFREQ), EVENTS (EVFREQ), AND AVERAGE EVENT DURATION (AVGEDUR) IN NUMBER OF FRAMES (30 FPS).

Label	NAO	OP	Pai1	Pai2	Pai3
NAO	21221	22	15	1502	46
OP	6	3812	132	2	33
Pai1	36	1	4617	110	48
Pai2	894	5	29	5177	47
Pai3	22	415	0	44	2576

Table V
CONFUSION MATRIX - PRIMARY VS SECONDARY FOR VFOA ANNOTATION

targets have been identified and considered as labels. They are: NAO, OP (the other participant), and the three paintings Pai1, Pai2, and Pai3. In addition, we defined a label OT (others) to denote a person looking at any other place in the room, and a DK (don't know) label when there is too much ambiguity between several VFOA targets and making a decision is not possible. Fig. 6 illustrates the approximate configuration of different targets in the scenario.
One person annotated an entire video sequence. Several annotators were used. The annotator performed the labeling using an interface displaying the video acquired from NAO (i.e. taking the robot perspective). Annotation was done with a precision of 150 ms on the average. Table IV provides the statistics from eight of the recordings. As can be seen, as a consequence of the scenario, looking at NAO is clearly dominating, esp. in terms of durations and is characterized by long gazes (average duration of 4.5 s). Note that the occurrence frequencies are not distributed evenly during the sessions: in the first part (introduction to the paintings), looking at paintings obviously happen more often; during the quiz part, interacting with and looking at the other person is more frequent.

In order to check the reliability of our annotations, we carried out secondary annotation on 2 minutes of data for 15 randomly chosen people among the total 26 participants. Table V contains the confusion matrix for our two annotation sets with the five main labels of interest. As seen from the table, the confusion between NAO and Painting 2 was high as the painting was right above NAO as seen in Fig. 2. Apart from this, the annotations are very reliable.

F. Addressee

Motivations. Addressee is the person or group of people to whom 'a speech utterance is intended to'. From NAO's point of view, knowing the addressee of a person's utterance is important in multiparty interactions. This information is useful for the robot to decide automatically if it 'has to' or 'should not' or 'can respond'. Though gaze information about 'who the current speaker is looking at' carries valuable information about addressee, previous research has shown that this cue is not always sufficient. Other contextual cues from the dialog state and spoken words in utterances have been shown to improve the detection accuracies [14], [29], [24].

Label	NAO	OPerson	Group	NoLabel
NAO	238	3	0	0
OPerson	11	242	0	0
Group	12	3	40	0
NoLabel	0	0	0	67

Table VI
CONFUSION MATRIX - PRIMARY VS SECONDARY FOR ADDRESSEE ANNOTATION

Labels and Annotations. Given the scenario, we are interested in labeling the addressee of the utterances from the two human participants. We assigned the following labels: {NAO, PRight, PLeft, Group, NO LABEL}. PRight and PLeft are the persons to the right and left of NAO. Group label corresponds to the situation where one participant addresses jointly NAO and the other participant. We assign NO LABEL if the current utterance has no addressee or if it is a speech act like 'Laughter'. The labeling of each utterance was done by one annotator having full access to the audio-visual recording. The GROUP label mainly occurred during the self-introduction phase. 13 interactions were used to compute the statistics.

Inorder to test the reliability of the annotations, a secondary coder performed the annotation for 4 out of 13 interactions (i.e. 30% overlap). The results show that Cohen's Kappa, the interannotator agreement, was 0.93, meaning they are infact quite reliable.

VI. DATA USE CASES

The Vernissage corpus allows to benchmark several perception methods using single or multiple modalities. NAO system data and other available annotations could serve as contextual information. This helps to understand what each of the modalities contributes and how the available annotations and dialog state information can act as additional context to make better estimations. Below, we illustrate the research tasks that could benefit from our dataset.

Some of the perceptual tasks could make use of inputs from different sources in single modalities for comparison. For example, audio tasks such as ASR and utterance estimation could use close-talk microphones or NAO's microphones. When performing automatic speaker localization using NAO's microphone data, we can benchmark the performance of the processing methods using the utterance annotation. The loss in performance when using the close-talk microphones and different NAO microphones tells us how challenging the task is. Studying this, better speech enhancement techniques could be devised to improve the signal-to-noise ratio. Visual tasks such as head-pose estimation and head-tracking could use video from NAO's camera or other external cameras (close to a indoor surveillance-type scenario). In certain cases, the effect of errors in a low-level cue such as speaker localization

on a high-level cue such as addressee estimation could be studied.

Audio-visual tasks can also be attempted. E.g., audio-visual people tracking, that could exploit both audio data (i.e. speaking information) and video data (i.e. head motion information). Addressee estimation requires both utterance information as well as gaze information. Both nodding estimation and prediction can make use of audio-visual cues such as speaking cues of others and head-movement of self.

Apart from the audio-visual modalities, NAO system data, other automatically estimated cues or their ground-truth could also serve as an important context for some tasks such as VFOA and addressee estimation. For example, VFOA estimation could be improved with the dialog state of NAO (e.g., what painting he is talking about). Wizard commands and their timing could be used to study how to automate gaze-shift and nodding. For estimating a higher-level cue such as 'who is being addressed', lower-level cues such as VFOA, and dialog context (from both NAO as well as the participants) could be relevant. It would be interesting to study how useful each of these cues is (using the ground-truth), what the degradation when using automatically extracted cues is, and also study what the best way of fusing the information is.

VII. Conclusion

In this paper, we presented a new corpus for benchmarking multimodal perception tasks, with extensive annotations. We systematically presented the scenario of the recording, how the corpus was acquired with all the relevant sub-modules, what ground-truth and annotations are available, the reliabilities, and the possible use cases of the dataset.

This dataset will serve as a good benchmark to evaluate audio-visual perception algorithms in a multiparty conversational context. The scenario makes the corpus valuable to study the interaction behavior for example, interaction quality and turn-taking or gazing patterns, apart from low-level nonverbal cue extraction methodologies. The movement of the participants and the robot head, noise from the fans of the robot, and limited sensing capabilities on the robot make the nonverbal cue extraction challenging.

The corpus along with all the meta-data is planned to be made publicly available end of this year. The released corpus will include annotations for speech utterances, speech transcription, 2D head location, nodding, visual focus of attention, and addressees. 3D head location and pose of the participants will also be available.

References

[1] B. Adams et al. Humanoid robots: a new kind of tool. *Intelligent Systems and Their Applications, IEEE*, 15(4):25–31, 2000.

[2] X. Alameda-Pineda et al. The RAVEL data set. In *ICMI 2011 Workshop on Multimodal Corpora*, Alicante, Spain, Nov 2011.

[3] E. Arnaud et al. The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In *Proc. ICMI*. ACM, 2008.

[4] S. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2008.

[5] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.3.03). http://www.praat.org, 2012. Computer program.

[6] B. Brüning et al. Automatic detection of motion sequences for motion analysis. In *13th Int. Conf. on Multimodality*, Alicante, Spain, 2011.

[7] S. Burger et al. The ISL meeting corpus: the impact of meeting type on speech style. In *Int. Conf. on Spoken Language Processing*, 2002.

[8] J. Carletta et al. The ami meeting corpus: A pre-announcement. In *Workshop on Machine Learning and Multimodal Interaction*, 2005.

[9] L. Chen et al. Vace multimodal meeting corpus. In *Workshop on Machine Learning and Multimodal Interaction*, 2005.

[10] N. Dahlback et al. Wizard of oz studies–why and how. *Knowledge-based systems*, 6(4):258–266, 1993.

[11] T. Fong et al. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[12] A. Green et al. Developing a contextualized multimodal corpus for human-robot interaction. In *Proc. LREC*, 2006.

[13] U. Hadar et al. Head movement during listening turns in conversation. *Nonverbal Behavior*, 9(4):214–228, 1985.

[14] M. Katzenmaier et al. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. of the 6th int. conf. on multimodal interfaces*. ACM, 2004.

[15] J. Keshet, D. Grangier, and S. Bengio. Discriminative keyword spotting. *Speech Communication*, 51(4):317–329, 2009.

[16] T. Kim et al. Meeting mediator: enhancing group collaboration using sociometric feedback. In *Proc. CSCW*. ACM, 2008.

[17] M. L. Knapp and J. A. Hall. *Nonverbal communication in human interaction*. Wadsworth, Cengage Learning, 7 edition, 2009.

[18] G.. Lathoud et al. Av16. 3: an audio-visual corpus for speaker localization and tracking. *Machine Learning for Multimodal Interaction*, pages 182–195, 2005.

[19] M. Lohse et al. Systemic interaction analysis (SInA) in HRI. In *Proc. Human-Robot Interaction (HRI)*, San Diego, CA, USA, 2009.

[20] N. Mana et al. Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In *Workshop on Tagging, mining and retrieval of human related activity information*, 2007.

[21] Y. Mohammad et al. The h3r explanation corpus human-human and base human-robot interaction dataset. In *Proc. ISSNIP*. IEEE, 2008.

[22] T. Niemüller et al. Providing ground-truth data for the nao robot platform. *RoboCup 2010: Robot Soccer World Cup XIV*, pages 133–144, 2011.

[23] M. Omologo et al. Speaker localization in chil lectures. *Machine Learning for Multimodal Interaction*, pages 476–487, 2006.

[24] R. op den Akker and D. Traum. A comparison of addressee detection methods for multiparty conversations. In *Proc. 13th Workshop on the Semantics and Pragmatics of Dialogue*, 2009.

[25] F. Pianesi et al. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41:409–429, 2007.

[26] K. Pitsch et al. Attitude of german museum visitors towards an interactive art guide robot. In *Proc. HRI*. ACM, 2011.

[27] D. Sanchez-Cortes et al. An audio visual corpus for emergent leader analysis. In *Proc. ICMI Workshop on Multimodal Corpora*, 2011.

[28] R. Stiefelhagen et al. The clear 2007 evaluation. *Multimodal Technologies for Perception of Humans*, pages 3–34, 2008.

[29] Y. Takemae et al. Automatic addressee identification based on participants' head orientation and utterances for multiparty conversations. In *Proc. ICME*, 2006.

[30] J. Wienke and S. Wrede. A middleware for collaborative research in experimental robotics. In *Proc. SII2011*, Kyoto, Japan, 2011. IEEE, IEEE.

[31] P. Wittenburg et al. ELAN: a professional framework for multimodality research. In *Proc. LREC*, 2006.

[32] Z. Zivkovic et al. From sensors to human spatial concepts. *Robotics and Autonomous Systems*, 55(5):357–358, 2007.