Human Activities Data Collection and Labeling using a Think-aloud Protocol in a Table Setting Scenario

Celeste Mason[1], Moritz Meier, Florian Ahrens, Thorsten Fehr, Manfred Herrmann, Felix Putze, Tanja Schultz

***Abstract*— We describe our efforts in developing a Biosignals Acquisition Space and Environment (BASE) to acquire a large database of human everyday activities along with a procedure to automatically structure and label these high-dimensional data into a valuable resource for research in cognitive robotics. The final dataset is planned to consist of synchronously recorded biosignals from about 100 participants performing everyday activities while describing their task through use of think-aloud protocols. Biosignals encompass multimodal sensor streams of near and far speech & audio, video, marker-based motion tracking, eye-tracking, as well as muscle and brain readings of humans performing everyday activities. This paper provides details of our pilot recordings carried out in the well established and scalable "table setting scenario." Besides presenting initial insights, the paper describes concurrent and retrospective think-aloud protocols and compares their usefulness toward automatic data segmentation and structuring.**

I. INTRODUCTION

The acquisition of a large dataset of human everyday activities in the Biosignals Acquisition Space and Environment (BASE) is an integral part of the EASE collaborative research center (http://ease-crc.org), that aims to enable robots to master the execution of everyday activities. The database is meant to facilitate science and engineering of humans performing and structuring everyday activities. In particular, the goal is to develop narrative-enabled episodic memories (NEEMs) [1], i.e. data structures derived from recorded observations, experiences, and activities to enhance the capabilities of robotic agents.

For this purpose we record sets of biosignals—time-aligned multimodal streams of signals emitted by the human body (see Figure 1)—and apply machine learning techniques to find manifold representations of everyday activities of lower dimensionality than the original data.

By this, we envision creation of a hierarchical temporal structure that is decomposed into general units containing semantic descriptors that are related to the recorded activities.

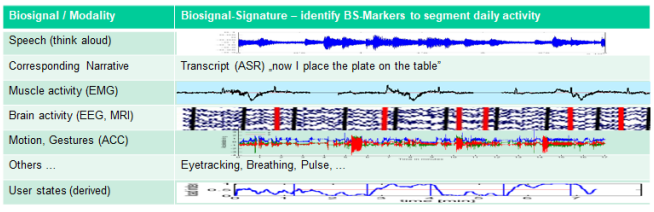

Fig. 1. NEEMs represented as time-aligned manifolds of biosignals

Ultimately, human activity models should be person-independent and cover a wide range of everyday activities. To achieve this goal and make best use of deep learning approaches, large volumes of multimodal sensor data from many subjects are advantageous. For the same reason, automated data processing in terms of segmentation, annotation, and embedding into a semantic representation is favored.

A. The EASE Biosignals Acquisition Space and Environment

EASE focuses on peoples' favorite robot application: taking over household chores. The EASE human activity data collection leverages the Biosignals Lab at the Cognitive Systems Lab at University Bremen that has been extended by an artificial kitchen setup, where participants can carry out kitchen activities in real and virtual settings.

Fig. 2. The Biosignals Acquisition Space and Environment (EASE-BASE)

The Biosignals Acquisition Space and Environment (BASE) consists of an interaction space (5x4m) which allows us to blend real with virtual reality interactions (see figure 2). The sensors, devices, and equipment available include Hololenses, stationary and head-mounted cameras, near- and far-field microphones for speech and audio event recording, a marker-based 9-camera OptiTrack motion capture system, wireless motion tracking based on PLUX inertial sensors, electrodermal activity (EDA) sensors, mobile eye-tracking with Pupil Labs headsets, muscle activity acquisition with stationary 256-channel and mobile 4-channel electromyography (EMG) devices, and brain activity recording based on a BrainAmp 64-channel electroencephalography (EEG) and mobile EEGs based on OpenBCI and g.Tec's g.Nautilus. In addition, an MRI scanner is available from the neuropsychology and behavioral neurobiology lab (see section V).

[1]celeste.mason@uni-bremen.de

B. The EASE Table Setting Dataset (EASE-TSD)

The EASE Human Activity Data (EASE-TSD) will be iteratively built up to, ultimately, cover a wide range of human activities. If ethics approval is granted, the data will be made publicly available. In the first iteration, we focus on the well established Table Setting Scenario, which allows us to leverage prior experience and compare against benchmarks available from several studies, such as the *EPIC-KITCHENS dataset* [2], the *TUM Kitchen Data Set* [3] and the *50 Salads dataset* [4].

In preparation for extensive analysis, we are collecting data from recordings of participants performing table setting trials while verbally describing their interactions in the BASE. To uncover activity and task structure, the participants' activity-related biosignals are recorded with the above described sensors under various conditions—e.g. table settings for different numbers of people, different meals, and different degrees of formalism. The data acquisition encompasses multi-channel time series with varied sampling rates, experimental conditions (temperature, etc), automatic markers (trial start/stop), experimental participant demographics, pre- and post-trial feedback questionnaires, and background information, such as familiarity with the experimental task activity. For technical details regarding the recording process, sensor data insight, and analysis methods, see [5].

Initially, semantic annotations on different levels of granularity and abstraction will be manually created post-experiment. Ultimately, we aim to automatically extract annotations from concurrent and retrospective think-aloud descriptions. For this purpose, the participants' verbal descriptions of objects, actions, and thoughts during the activity will be processed using automatic speech recognition (ASR) and other fully automatic annotation methods. The objects and events will be labeled according to an ontology specifically designed for this context and used to incrementally build up the NEEMs for robotic planning algorithms.

C. Related Work

Several studies of everyday human activities have focused on kitchen activities in the attempt to model behaviours. In the framework of CRC 588 on Humanoid Robots, Gehrig et al. [6] focused on intention, activity, and motion recognition. They produced a dataset with video of ten participants performing ten repetitions of seven unique tasks in a kitchen setting, resulting in 700 image sequences. From this, they proceeded to manually annotate the data with 60 motion primitives, and inferred objects from related activities. The TUM Kitchen Dataset [3] focused on activities in a natural kitchen setting using video, full-body motion capture, RFID tag readings and magnetic sensor readings from objects and the environment. With this data, they produced manual motion tracker labels and automatic segmentation of the observed motions into semantic classes. The 50 Salads dataset [4] was produced in a non-natural kitchen setting using a third person, top-down point of view camera, where participants prepared a meal according to a recipe. Sensors used included RGB-D cameras, and accelerometer-mounted objects. The EPIC Kitchens dataset [2] consists of first person point of view video from a head mounted camera of 32 participants performing kitchen activities in their homes. No time limitations were set and participants use a variety of languages, with the instructions to limit articles and other non-critical words. From this data, they produced annotations using 125 verb classes and 352 noun classes. While several datasets of humans performing kitchen activities exist, only EPIC and EASE-TSD use verbal annotation, and only the TSDset uses participants' natural speech. Furthermore, a large dataset covering a manifold of biosignals along with a think-aloud protocol for annotation has, to our knowledge, not yet been created.

II. EXPERIMENTAL SETTING PREPARATION

The setup of the BASE has focused on providing the optimal vantage point to observe the fine details of human interactions in the space while performing everyday activities. The interaction space for the EASE-TSD collection is constructed in a motion capture truss structure with fabric panel walls as depicted in Figure 3. While this is not a natural kitchen setting, it provides a reasonably close approximation to a kitchen and dining space, as described below. Stationary, room-wide sensors include nine IR cameras used for optical motion capture, seven cameras positioned at side views, top-down views for the whole room and each surface, and a scene microphone. Worn sensors include a 16-channel mobile EEG cap, a mobile eye-tracker, a head-mounted microphone for speech, EMG sensors on the arms, EDA on the palm, and accelerometers on both the EEG cap and the participant's back. A suit mounted with passive IR reflective markers is also worn, for use with the motion capture system. The sensors along with their configurations, and methods of synchronized recording are detailed in [5].

Fig. 3. Three views of table setting activity

Within the kitchen space, surfaces are positioned to increase likelihood of diverse interactions. The objects used to recreate a kitchen setting were carefully selected to be familiar to most participants, to provide a manageable vocabulary/ontology and the potential to formulate expectations about task outcome and possible approaches, to offer enough variability to study interesting research questions, to maintain sufficient realism of experience, and to facilitate planned computational processes, such as object recognition for optical tracking. Objects were coated to provide non-reflective surfaces or reduce transparency, and synthetic food items were produced to encourage realistic handling and maintain consistency between trials.

Initially, movable objects such as cutlery, dishes, and food are arranged on stationary objects (e.g. counter, dining table),

in the same manner for each trial. Some were shared between meals (e.g. bread, butter, water, tray), some were intended for lunch (e.g. soup bowls, ladle, forks, salad cutlery), and some for breakfast (e.g. jam, milk, cereal, coffee, mugs, small plates and spoons). Chairs and non-rigid objects such as napkins may be included in later iterations of the experiment.

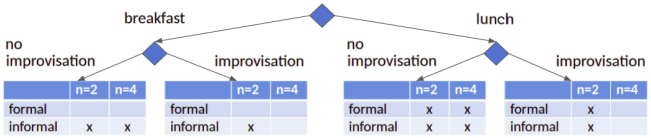

Fig. 4. Breakdown of scenario variations

The sequence of trials and variation of conditions were carefully devised to allow the observation of a wide array of behaviors. Variation in table setting approaches that may emerge include differences between breakfast and lunch configurations, formality based on diners attending the meal, variations in collection and dispersal for groups of smaller or larger numbers, and changes in behaviour due to unexpected events (e.g. missing items) in the improvisation trial variant. The breakdown of trial variants is shown in Figure 4.

The context was selected to optimize the potential behaviour variations of subjects in terms of arrangements, strategies, speed, precision, degree and type of movement, reactions, etc. For example, breakfast may require more objects that might be shared between diners. Provision of a tray would allow participants the option to move many objects at once, potentially improving efficiency. Inclusion of formal meal setting variations may result in differences in the speed and attention to detail displayed during the setting activity. Omission of a required number of utensils may reveal how people react to and cope with unforeseen situations in the improvisation variants.

III. Experimental Procedure

An experimental session adheres to a pre-defined sequence, shown in Figure 5. Participants receive information about the study, give consent, go through sensor setup and calibration, receive instructions, perform a series of randomly selected trials, and are then debriefed. Additionally, pre-trial and post-trial questionnaires are administered.

A. Experiment Instructions and Consent

General instructions for all trials are as follows:

- You and your [companions:formal/informal] stay in a vacation apartment. It is your first day.
- The others are out shopping, you agreed to set the table.
- You want to set this table for [a meal: breakfast/lunch] for [some number of (2 or 4)] people.
- You want to put all the necessary objects from the counter to the table and have everything ready to eat as your friends will return shortly.
- Please think aloud about what you are doing as you perform the activity - [timing: concurrent / retrospective]
- No time constraints (perform at a natural pace)
- No detailed context knowledge

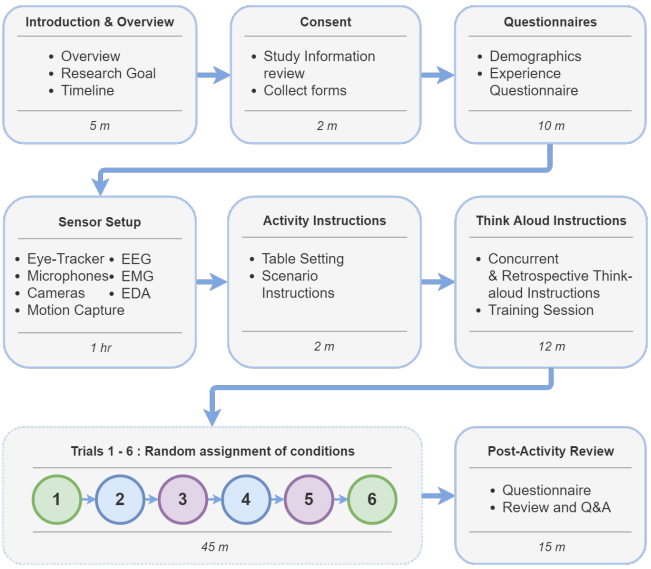

Fig. 5. Overview of the flow of an experiment session

B. Suit up and Sensor Calibration

An optimal sensor configuration procedure minimizes the impact felt by the participant wearing numerous sensors and a full body motion capture suit. The eye-tracker, EEG, and motion capture systems must be individually and expediently calibrated prior to each experimental session.

IV. Think-aloud Protocol Pilot Study

Determining the best means of gaining activity information based on the thoughts and perceptions of participants performing the activities requires investigation of the interrelated influences on the participant's ability to perform the activity while verbalizing in sufficient detail [7][8].

A. Concurrent vs Retrospective Protocols

The two primary methods being considered are concurrent think-aloud protocols and retrospective reports of activities while participants observe their behavior through recorded video, as shown in Figure 6. Based on discussions in the literature [9][10][11][12], variations in such methods have been investigated to attempt to tune the protocol to the needs of this study. While think-aloud methods in HCI studies may allow experimenters to interact to various degrees with participants in order to elicit more frequent or detailed information, the need to maintain as natural as possible context when performing table setting motivates the decision to provide instruction only during intermissions between trials.

The think-aloud protocols and activities that have been examined in previous studies do not correspond directly to our context. Consequently, differences in the level of detail and accuracy of recollections when comparing concurrent vs. retrospective protocols cannot be considered directly admissible in the contexts we examine. Most studies relying on concurrent methodologies do not strictly enforce concurrency, instead opting to allow switching between focused

physical activity and verbalization. Due to the physical nature of the table setting activity and instructions to continuously speak while performing, a main focus of this pilot study is to examine which method serves this context best. Because participants will not be reminded during the course of a trial, it becomes necessary to provide clear, concise instructions and allow participants to perform think-aloud procedures during a tangentially related activity as a training exercise.

B. Selection Criteria

Given the many competing factors to weigh, the ultimate goal must be defined so that the advantages of these approaches address the needs of the experiment. In this case, we must require a think-aloud protocol, or combination of different variants, that allow us to best fulfill the following criteria: (1) Challenges in identifying the optimal think-aloud procedure relate to the need to balance quality of data collected with the unintended effects the process of following the protocol may have on other aspects of the experiment. (2) The need to maintain a natural behavioural setting motivates the preliminary instruction and training session, encouraging a specific level of detailed verbalization without priming subjects for specific behavior, increasing cognitive load, or otherwise interfering with natural behaviours significantly during the trials. (3) The influence of speech production on other sensor recordings must be minimized–in particular, related artifacts in EEG and eye tracking data. (4) This must be balanced with the goal of producing semantic descriptions at a specified level of detail and abstraction that provides accurate information about priorities of the participant and basic scene understanding. (5) We must account for time limitations of participants ability to maintain focus and engagement during repetitive tasks, and endure recording with full coverage on-body sensors for approximately two hour sessions. (6) We must determine whether think-aloud procedures performed during or after a trial provide the best value versus time, considering effect on participants performance and variations in individuals' memories. To determine the best methods for data collection in this context, we performed pilot trials with ten participants comparing concurrent and retrospective procedures during a subset of six trials selected from the previously detailed experimental trial variations. The same basic instructions were given for each method:

- Describe what you do (at each moment) as you interact with the environment
- Talk constantly, so long as it does not interfere with the ability to perform the activity
- Treat each trial independently, repeat words as needed

C. Preliminary findings

Pilot study participants are German or English speaking volunteers from associated labs. Seven participants spoke German and three spoke English during the trials. Three participants were female. All participants were right handed and between the ages of 21 and 41.

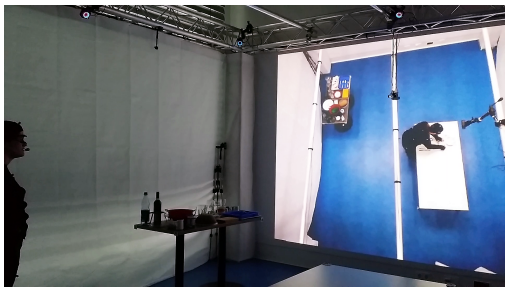

Fig. 6. Participant performing retrospective report using third person video

TABLE I
THINK-ALOUD TRIAL VARIANT SPEECH DURATIONS

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**AVG**
C1	80	131	137	72	167	288	125	249	206	69	**152**
C2	79	92	130	91	182	236	82	363	127	118	**150**
C3	94	121	134	100	198	235	140	264	151	102	**154**

Participants performed six table setting activity trial variations, defined as a representative subset of variations from the planned study, with the addition of a think-aloud protocol type variation—concurrent, retrospective, or both.

Challenges encountered during development of this experimental setup included synchronization of the multimodal sensor recordings, balancing the value of sensor recording methods against interference with one another, and overcoming the physical limitations of such a dense electronic recording setup, while trying to maintain a natural experimental setting. Extensive automation was developed and refined to provide a smooth experimental session, while minimizing the potential for human error. The eye-tracker recording configuration and calibration has gone through continuous refinement in order to overcome bandwidth, heat, and participant point-of-view limitations. The wearable sensors required a customized harness incorporating individually tailored attachment of the EEG, eye-tracker, voice-microphone, and Plux sensors to increase comfort and proper functioning for extended experimental session periods.

From review of initial recordings, participants provided slightly more vocal descriptions of the activity when reviewing recorded video than when speaking during the activity, with fewer (longer) pauses. Descriptions included low (infrequent) and high (frequent) level behaviors and plans, observations, and reasoning. Table I provides the average duration of speech (in seconds) during think aloud trial recording for each participant (S1-S10) for each condition (C1-C3), with average durations. Condition C1 is "concurrent," C2 is "retrospective," and C3 is the average of C1 and C2 during the "both" condition trials. Based on comparison of average trial durations, it may be assumed that the concurrent protocol does not significantly slow the performance of activities.

V. NEUROIMAGING EXPERIMENTS ON EASE-TSD

In addition to providing the environment for the behavioral investigations, the experimental setup also provides a back-

drop for the acquisition of first person table setting videos, which are used for analyzing human brain activity derived from functional Magnetic Resonance Tomography (fMRI), as shown in Figure 7, and high density multi-channel EEG data.

Resembling the actions of the participants in the table setting experiments, the videos encompass a variety of scenarios and depict both confidently finished runs as well as scenarios causing interference resolution or erroneous behavior induced by missing or misplaced objects. The actions in these videos are carried out by an experimenter, who employs articulate and easily traceable movements of the arms and hands as well as smooth pans of the view. After recording, they are manually annotated and their timelines organized into substantially distinct episodes.

Fig. 7. fMRI measurement during video observation

These first person table setting scenarios then serve as standardized tools to be shown in neuroimaging studies at the neuropsychology and behavioral neurobiology lab. Participants of these studies employ motor imagery [13] to actively put themselves into the observed situation, while their brain activity is measured either via EEG or fMRI and later correlated to the established episodes. The use of both EEG- and MRI devices in the viewing scenario will allow us to take full advantage of the higher spatiotemporal signal information and to introduce a combined fMRI constrained source analysis [14], so the results will demonstrate a wide range of brain network activation with a high temporal and spatial resolution. We particularly aim at the analysis of the dimensionality of involved networks during the planning and execution of complex everyday activities and the handling of unexpected situations. Besides giving novel insights for members of the neuroscience community, this will notably serve researchers in the field of robotics, by providing them with a template on how these demanding tasks are solved by the organisms they strive to emulate.

VI. Outlook

We have provided an overview of experimental table setting scenarios used when recording a multi-modal human activity dataset. Future work will entail detailed analysis of the value versus cost of the two think-aloud protocols in terms of description frequency, levels of detail and abstraction, time-alignment, and interference with other recording modalities.

Acknowledgment

The research reported in this paper has been supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 EASE - Everyday Activity Science and Engineering, University of Bremen (http://www.ease-crc.org/). The research was conducted in subproject H03 Descriptive models of human everyday activity.

References

[1] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, and G. Bartels, "Knowrob 2 . 0 a 2nd generation knowledge processing framework for cognition-enabled robotic agents," 2018.

[2] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[3] M. Tenorth, J. Bandouch, and M. Beetz, "The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1089–1096.

[4] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 729–738.

[5] M. Meier, C. Mason, R. Porzel, F. Putze, and T. Schultz, "Synchronized multimodal recording of a table setting dataset," in *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges, Madrid, Spain*, 2018 (submitted).

[6] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen, "Combined intention, activity, and motion recognition for a humanoid household robot," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2011, pp. 4819–4825.

[7] K. A. Ericsson and A. S. Herbert, "Verbal reports as data," vol. 87, no. 3, pp. 215–251, 1980.

[8] T. Boren and J. Ramey, "Thinking aloud: reconciling theory and practice," *IEEE Transactions on Professional Communication*, vol. 43, no. 3, pp. 261–278, Sept 2000.

[9] M. Hertzum, K. D. Hansen, and H. H. Andersen, "Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload?" *Behaviour & Information Technology*, vol. 28, no. 2, pp. 165–181, 2009. [Online]. Available: https://doi.org/10.1080/01449290701773842

[10] T. Zhao and S. McDonald, "Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ser. NordiCHI '10. New York, NY, USA: ACM, 2010, pp. 581–590. [Online]. Available: http://doi.acm.org/10.1145/1868914.1868979

[11] J. C. Welsh, S. A. Dewhurst, and J. L. Perry, "Thinking aloud: An exploration of cognitions in professional snooker," *Psychology of Sport and Exercise*, vol. 36, pp. 197 – 208, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1469029217306763

[12] T. Towne, K. Ericsson, and A. Sumner, "Uncovering mechanisms in video game research: suggestions from the expert-performance approach," *Frontiers in Psychology*, vol. 5, p. 161, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00161

[13] M. Jeannerod, "Mental imagery in the motor context," *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, nov 1995. [Online]. Available: https://doi.org/10.1016/0028-3932(95)00073-C

[14] S. A. Trautmann-Lengsfeld, J. Domínguez-Borràs, C. Escera, M. Herrmann, and T. Fehr, "The perception of dynamic and static facial expressions of happiness and disgust investigated by ERPs and fMRI constrained source analysis," *PLoS ONE*, vol. 8, no. 6, p. e66997, jun 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0066997