The MULTISIMO Multimodal Corpus of Collaborative Interactions

Maria Koutsombogera
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
koutsomm@scss.tcd.ie

Carl Vogel
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland

ABSTRACT

This paper describes a recently created multimodal corpus that has been designed to address multiparty interaction modelling, specifically collaborative aspects in task-based group interactions. A set of human-human interactions was collected with HD cameras, microphones and a Kinect sensor. The scenario involves 2 participants playing a game instructed and guided by a facilitator. Additionally to the recordings, survey material was collected, including personality tests of the participants and experience assessment questionnaires. The corpus will be exploited for modelling behavioral aspects in collaborative group interaction by taking into account the speakers' multimodal signals and psychological variables.

CCS CONCEPTS

• **Human-centered computing** → **Laboratory experiments**; **Collaborative interaction**;

KEYWORDS

Multimodal corpus; group interaction; collaboration; personality

ACM Reference Format:
Maria Koutsombogera and Carl Vogel. 2017. The MULTISIMO Multimodal Corpus of Collaborative Interactions. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3136755.3151018

1 INTRODUCTION

The use or development of multimodal corpora has become a common requirement in multimodal behavior modelling. Corpora are a knowledge base providing rich information about the structure of the interaction but also about the communicative intent and the affective state of the speakers [1], [4]. The MULTISIMO corpus was developed to enable the understanding of human social behavior in multiparty interaction, the structural representation of the interaction flow, and the modelling of the social communication mechanisms to be integrated into intelligent collaborative systems exhibiting natural behavior. In the following sections we describe the experimental design and setup that led to the data acquisition as well as the resulting data collection.

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3151018

2 EXPERIMENTAL DESIGN AND SETUP

We aimed to implement a scenario whose nature is expected to trigger the collaboration between group participants. We thus designed sessions, in which a group of 3 persons, i.e. 2 players and 1 facilitator, was sitting around a table and was engaged in a collaborative discussion about a game-like task. The task was that the players collaborate with each other to find the 3 most popular answers to each of 3 questions (based on survey questions posed to a sample of 100 people), and to order their answers in terms of popularity. Participants were guided by the facilitator who instructed them throughout the session.

MULTISIMO investigates, among others, the effect of participants' personality traits on the task success and aspects of collaboration in participants' behavior. To have an objective measure of the participants' personality traits, all participants completed a personality test prior to the recordings. We employed the Big Five Inventory (BFI) [2], a self-report inventory designed to measure the Big Five dimensions. Furthermore, after each session the participants completed an experience assessment survey.

The data acquisition setup includes three frontal HD cameras, one 360 camera, three head-mounted microphones, one omnidirectional microphone and one Kinect 2 sensor. Two of the frontal HD cameras (1920x1080 px, 29.97 fps) are placed opposite each of the two players (cf. Figure 2). The third frontal camera (1920x1080 px, 25 fps) is placed opposite the facilitator and captures the whole scene (cf. Figure 1); its zoomed angle is used to isolate the facilitator's front view. The 360 camera (3840x2160 px, 29.97 fps) was placed in the middle of the table to capture the whole scene from a low angle (cf. Figure 3). The head-mounted microphones enable the recording of individual audio signals (SR 44.1 kHz), while the omnidirectional microphone is used as a backup audio source (SR 44.1 kHz). Finally, the Kinect 2 was placed in a way that it would perform the skeletal tracking of all participants (cf. Figure 4).

The recordings took place at Trinity College Dublin (TCD). The experiment, including the recordings and the surveys lasted about 40 minutes per participant. All participants were rewarded with a 10 euro voucher. In total, 49 participants were recruited; 3 of them shared the role of the facilitator throughout the 23 sessions and 46 of them were grouped into pairs of players. The facilitators were trained before the actual recordings via pilot experiments. All sessions were carried out in English.

The experimental design was supervised by the TCD ethics committee to ensure good ethical practice. Our goal is to make this data

Figure 1: Sample of a recorded session depicting the experimental setting.

Figure 2: Samples of close-up angles.

Figure 3: Sample of 360 low angle.

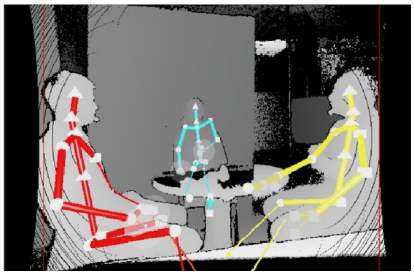

Figure 4: 3D view of Kinect performing skeletal tracking.

collection available to the research community, and at the same time protect the participants' rights to privacy and confidentiality. Therefore, we followed the informed consent practice so that participants would decide whether they'd wish to disclose their personal information to third parties [3].

3 CORPUS DESCRIPTION

The data collection consists of 23 sessions and its duration is approximately 4 hours. The average session duration is 10 minutes (min=6, max=16). Grouping was performed randomly. In most of the groups participants are unacquainted with each other. The average age of the participants is 30 years old (min=19, max=44). Gender is balanced in terms of participants (F=25, M=24). However, the gender distribution varies depending on the pairing of the players. For example, there are groups where both of the players are female, or groups with male players, and groups with both genders. The participants span eighteen nationalities and one third of them are native English speakers. Gender, language and familiarity are important factors to be co-examined when correlations with the multimodal behavior will be performed. Furthermore, scores of personality tests and experience assessment surveys are information that will lead to significant findings when associated with collaboration and task success aspects. The corpus is currently being processed and annotated at various levels, including speech transcription, automatic feature extraction and manual coding of low and high-level signals.

We aim to make the corpus publicly available, i.e. the sessions for which the participants have given their consent to be public. The dataset will include audio, video (in high and low resolution formats) and Kinect files, accompanied by documentation and annotation files. Sample data from the corpus are available for viewing at https://www.scss.tcd.ie/clg/MULTISIMO/.

The demo on site will showcase the data collection, the annotation scheme used for the manual coding of various features as well as automatically extracted visual and audio features. It will also provide additional details about the experimental setup, the data analysis and the investigation of the multimodal behavior of the participants. Finally, we will discuss tasks (implemented or in progress) that exploit the corpus information and representations, and specifically, tasks related to measuring collaboration and task success based on visual and temporal (audio-based) features.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 701621 (MULTISIMO). We would also like to thank Akira Hayakawa, Fasih Haider and Nick Campbell for their significant contribution to the recordings implementation.

REFERENCES

[1] Anna Esposito, Antonietta M. Esposito, and Carl Vogel. 2015. Needs and Challenges in Human Computer Interaction for Processing Social Emotional Information. *Pattern Recognition Letters* 66 (2015), 41–51. https://doi.org/10.1016/j.patrec.2015.02.013

[2] Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research* 3 (2008), 114–158.

[3] Maria Koutsombogera and Carl Vogel. 2017. Ethical Responsibilities of Researchers and Participants in the Development of Multimodal Interaction Corpora. In *8th IEEE Conference on Cognitive Infocommunications, Debrecen, Hungary, (September 11-14, 2017)*. IEEE, 277–282.

[4] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Emer Gilmartin, Zakia Hammal, et al. 2015. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation* 7, 4 (2015), 397–413.