M3B Corpus: Multi-Modal Meeting Behavior Corpus for Group Meeting Assessment

Yusuke Soneda
soneda.yusuke.su2@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

Yuki Matsuda
yukimat@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

Yutaka Arakawa
arakawa@ait.kyushu-u.ac.jp
Kyushu University
Fukuoka, Japan

Keiichi Yasumoto
yasumoto@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

ABSTRACT

This paper is the first trial to create a corpus on human-to-human multi-modal communication among multiple persons in group discussions. Our corpus includes not only video conversations but also the head movement and eye gaze. In addition, it includes detailed labels about the behaviors appeared in the discussion. Since we focused on the micro-behavior, we classified the general behavior into more detailed behaviors based on those meaning. For example, we have four types of smile: response, agree, interesting, sympathy. Because it takes much effort to create such corpus having multiple sensor data and detailed labels, it seems that no one has created it. In this work, we first attempted to create a corpus called "*M3B Corpus* (Multi-Modal Meeting Behavior Corpus)," which includes 320 minutes discussion among 21 Japanese students in total by developing the recording system that can handle multiple sensors and 360-degree camera simultaneously and synchronously. In this paper, we introduce our developed recording system and report the detail of *M3B Corpus*.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**; *Computer supported cooperative work.*

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3345588

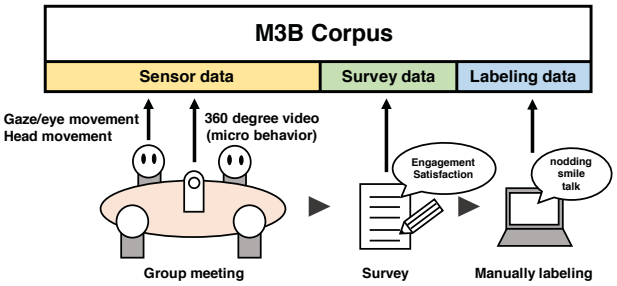

Figure 1: M3B Corpus Constitution

KEYWORDS

corpus, multi-modal communication, multi-sensors fusion, non-verbal communication

ACM Reference Format:
Yusuke Soneda, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. 2019. M3B Corpus: Multi-Modal Meeting Behavior Corpus for Group Meeting Assessment. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, London, UK, 10 pages. https://doi.org/10.1145/3341162.3345588

1 INTRODUCTION

A good communication skill is getting more and more important in social communities such as schools and companies [4]. Communication can generally be classified into verbal communication and non-verbal communication. The former means communication based on language. The latter means communication except for the language, which is usually based on gesture, behavior, and gaze. Non-verbal communication is quite important because it can give a big impression against the other in the social community. However, at the same time, incorrect communication might give the other person an unpleasant impression. Some studies

have been conducted on how non-verbal communication such as posture, gestures, and facial expressions affect social relationships. Ekman's research is famous for the influence of non-verbal communication reflected in the head such as nodding [10], and Peter. E. Bull's research is famous for the influence of the impression on the gesture reflected in the body such as posture and gesture [3].

To understand and recognize the non-verbal communication appeared in a group discussion, it is important to measure posture, gestures, and facial expressions by multiple sensors. However, it is hard to handle those sensors simultaneously and record the collected data synchronously. Therefore, no corpus including multiple sensor data during a group discussion have been released. To cope with this problem, we first developed a system that operates all sensors in one-stop and records multiple sensor data with no deviation in the timestamp.

In our system, we use three sensors: 360-degree camera, mobile eye trackers, and inertial measurement units (IMUs). A 360-degree camera records the facial expressions and gestures from the center of the table. All the participants wear the mobile eye tracker to measure the gaze. IMUs are added on the eye tracker for measuring the head movement. All those sensors are operated remotely and synchronously from another PC.

After 5 minutes of discussion on certain topics, we conducted a survey on the satisfaction level, engagement to participate in the discussion and so on. After finishing the survey operation, we ask the participants to assign the labels by using labeling tool.

The created corpus is named "*M3B Corpus* (Multi-Modal Meeting Behavior Corpus)." As shown in Figure 1, our corpus consists of three data: multi-modal sensor data, survey data, and labels. *M3B Corpus* is available online at http://data.ubi-lab.com/M3B. The voice during the discussion may include personal information, so *M3B Corpus* does not include the voice data.

2 RELATED WORK

Hori et al. [13] developed a system that analyzes face orientation and facial expression in real time by combining multiple omnidirectional cameras and directional microphones. However, it is necessary to learn the shape of the room in advance for speech analysis. Also, it requires an expensive resource for analysis. So, it is hard to set up the same system in various meeting rooms.

Onishi et al. [1] developed a system to identify the nodding motion, the turn of the head, and talking person by using a 9-axis acceleration sensor attached on the head. Head acceleration data has been reported to be practical data in determining face movement. However, it is necessary to correct the timestamp of the acquired data manually for synchronizing multiple persons' data. Philipp et al. [12] pointed out the importance of combining the timestamp of each device, especially when creating corpus for activity recognition. For example, we use multiple sensors to collect sensor data with multiple people, the number of sensors increases with the number of people. When three sensors are attached to one person and four persons participate in the experiment, a total of 12 sensors exist in the system. If the participants turned on these sensors manually and measurement is started, deviations occur in the timestamp of each sensor data. Therefore, we have to match the timestamp of each sensor at the time of the corpus generation, and the task becomes more difficult if the number of sensors increases. To prevent deviation of each timestamp of each sensor, it is necessary to have a system that can operate one-stop all sensors at the time of data collection.

Most of the existing corpus published for communication analysis is not for human-to-human, but for human-to-robot communication [5, 8]. The multi-modal corpus about human-to-human has not been confirmed.

Inoue et al. [7] checked the acceleration data of the head at the time of nodding and the impression of the nodding. However, their data is not that of communication in the real environment, because the nodding is prompted.

Our system does not require expensive resources and can collect the data on group discussions by three kinds of sensors synchronously, so we can easily create a multi-modal corpus without timestamp deviations. This makes it possible to compare the labeling information not only with one sensor but with multiple sensors in analyzing micro-behavior. *M3B Corpus* consists of sensor data acquired during this group discussion, survey data on discussions, and labels for micro-behavior.

3 DATA COLLECTION OF M3B CORPUS

We developed a system for acquiring data during group discussion and created a corpus. In this section, we describe the outline of the experiment and the method of creating the corpus.

Assumed Environment

We asked the participants to discuss certain topics for five minutes. The number of participants per discussion is four. The topic is chosen from the topics that will not be affected by the knowledge possessed in advance and whose answer can be narrowed down to two choices (e.g., If you can use a time machine, would you go to the past or future ?). It aims to avoid the situation where the superiority of the discussion occurs depending on the presence or absence of prior knowledge and the situation where the discussion diverges due to not deciding the content of the discussion.

The study received ethics approval (approval no :2018-I-28) after review by the research ethics committee at Nara Institute of Science and Technology.

Data Collecting System

We use the following sensors to get multi-modal data appeared in group discussions: RICOH THETA V [11], Pupil Labs Mobile Eye Tracking Headset [9] (hereinafter, this is called "Pupil") and LPMS-B2 [6]. Figure 2 shows the architecture of the developed system. All the sensors are connected to the control PC via WiFi, LAN, and Bluetooth. We developed the program that can handle all the sensors simultaneously.

Figure 3 shows the layout of a desk, chairs, participants, and a camera, where each participant wears a Pupil and LPMS-B2 on their head. Laptop PCs (using Microsoft Surface Pro) required for running a Pupil are set on the desk. It is also used for labeling work.

Features of M3B Corpus

Sensor data. THETA V is a 360-degree camera developed by RICOH that captures the motion of the participant as a video. The frame rate at the time of video shooting is 29.97 fps, and the number of pixels is 3840 × 1920 pixel (4K). Pupil is a wearable gaze system developed by Pupil Labs and is used to know where the participant was looking at. Figure 5 shows an image of Pupil. LPMS-B2 is a 9-axis acceleration sensor developed by LP-Research. LPMS-B2 can control the frequency to acquire data and was set to 100 Hz in this experiment. Also, it is possible to obtain data on the environment such as air pressure and temperature, but we focused on 3-axis acceleration and 3-axis angular acceleration data that are considered to be important in recognizing micro-behavior. Figure 4 shows the elements of the image and output of LPMS-B2. By attaching LPMS-B2 to Pupil as shown in Figure 6, we obtained head motion data.

The video recorded using THETA V is converted to point cloud data by applying OpenFace [14]. OpenFace changes face image to point cloud data and calculate the face angle and gaze angle. Figure 7 shows how OpenFace is applied to the face image and the elements used for the analysis in this paper. Also, Figure 8 shows the image when applying OpenFace to the video recorded by THETA V. By converting videos into point cloud data, there are merits that we can protect the private data and the combination with labels can be simplified.

Survey data. At the end of each discussion, participants answered a survey in Table 1. As for A1 and A2, about the theme of the discussion, participants answer 1 in the case of the former opinion, 5 in the case of the latter opinion. For example, if the participant answered 1 for the theme "If you can use a time machine, would you go to the past or future ?", it means that he/she wants to go to the past. As for B1 to B8, participants answer closer 1 in the case of a negative opinion, closer to 5 in the case of a positive opinion.

Labels. Each participant used a tool called "ELAN" [2] to label their own micro-behavior. ELAN is OSS software that is mainly used to label time series data such as video. Table 2 shows the contents of labelling. In this experiment, we decided the contents to be labeled focusing on the head movements that are simple and easy to check. Based on Peter's book [3], we defined the contents of labeling data that consider micro-behavior means. Although it has been reported that the posture and movement that appear in the body affect communication, the movement is complicated by people, and in this experiment, labeling work was not performed for body action.

4 SUMMARY OF M3B CORPUS

M3B Corpus was created through 16 discussions between the period from May 7 to June 5, 2019. A unit of discussion was set to five minutes, and four people participated in the discussion. In total, 80 minutes of discussion data, 320 minutes of sensor data and labels were collected. The participants consisted of students in the laboratory, they are 17 males and 4 females aged 22-24. Their native language is Japanese. Of these, 10 people participated in the experiment once, and 11 people participated in the experiment twice. It took a total of 130 hours to create the corpus. Most of the time is used for labeling work.

Corpus statistics

Figure 9, 10, 11 show the graphs of the survey results. From the graphs of Figure 9 and Figure 10, we confirmed the tendency of the opinion to become neutral through the discussion. In Figure 11 survey ID B8 "I have an attention for the 360-degree camera" point is as low as 1.375, most participants did not pay attention to the 360-degree camera. Placing a 360-degree camera in the center of the table during an actual discussion, it is expected not to cause a distraction.

Table 3 shows the statistics of time taken for each micro-behavior per person during a five-minute discussion and Figure 12 shows the statistical graph. The time units of tables and graphs in this paper are all "second". Figure 13 shows a visualization of how per-person micro-behavior is distributed during the discussion as an example. Next, Table 4 shows the statistics about the duration per micro-behavior and Figure 14 shows the statistical graph. For printing reasons, smile is abbreviated to "s", nodding is abbreviated to "n", and talk is abbreviated to "t" as a prefix.

From Figure 12, it is confirmed that there are some participants with few nodding, smile or talking. This case occurred

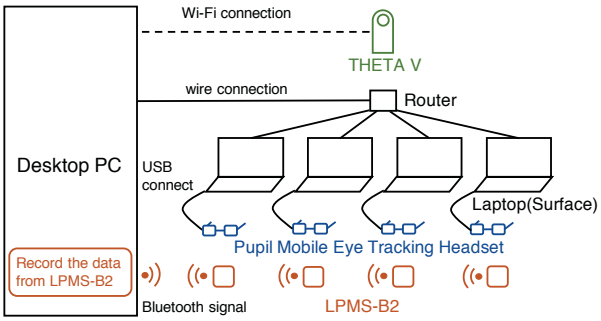

Figure 2: Architecture of developed system

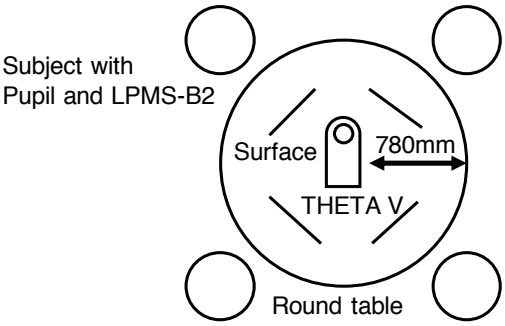

Figure 3: Layout of a desk, sensors, and participants

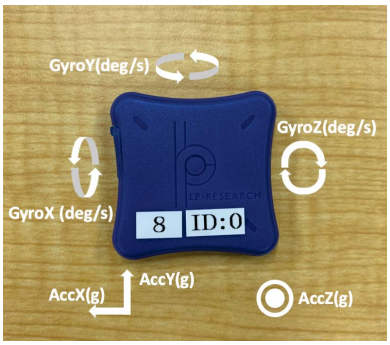

Figure 4: LPMS-B2 output element

Figure 5: The role of each Pupil camera

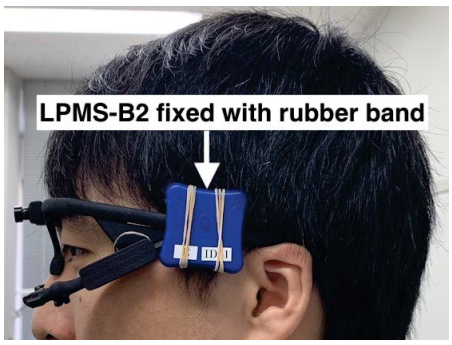

Figure 6: Pupil with LPMS-B2

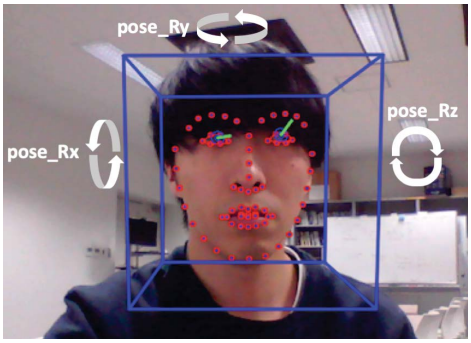

Figure 7: OpenFace and output element

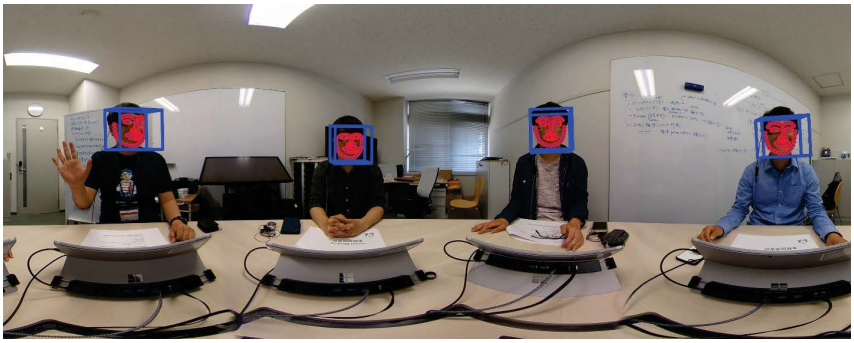

Figure 8: Apply OpenFace to group discussion recorded by THETA V

when a particular person continued to express his/her opinion and some participant could not get the opportunity to speak. We also obtained some opinion in the interview after the experiment that "It was difficult to explain my opinion to the seniors in the laboratory". When conducting group discussions, it is conceivable that the relationship between participants affects the satisfaction level of the discussion and the ease of speaking.

Referring to Figure 14, there was no tendency for the duration to change because of the meaning of nodding and smile. On the other hand, focusing on the talk labeling, it was confirmed that "description" took longer time than the other categories. The statistics for each micro-behavior shown in the Table 4 are essential to determine the size of the window function when applying the sliding window algorithm for activity recognition in the future.

Example of data in M3B Corpus

In this paper, we analyzed sensor data and labels of nodding. We checked how the output of OpenFace and LPMS-B2 correspond to the result of labeling.

Table 1: Survey questions

Survey ID	The content of survey
A1	Before the discussion, is your opinion the former or the latter ?
A2	After the discussion, is your opinion the former or the latter ?
B1	I was satisfied with the discussion.
B2	I could talk my own opinion.
B3	I heard what people with the same opinion say.
B4	I heard what people with the opposite opinion say.
B5	The discussion was enjoyable.
B6	The group often cut off my speech.
B7	I would like to have discussion with this group in future.
B8	I have an attention for the 360-degree camera.

Table 2: The contents of labeling

Label name	Detailed category	Explanation of category
smile	response	without special meaning generated by the listener
	agree	smile with consent
	interesting	smile occurs when the discussion is interesting
	sympathy	seeking empathy
nodding	response	without special meaning generated by the listener
	agree	nodding with consent
talk	description	explain his/her opinion
	objection	deny the opinion of the other
	agree	agree with the other opinion
	say	give the right to speak to an other person

Table 3: Statistics on per person behavior during 5 min discussion

statistics	s_total	s_response	s_interesting	s_agree	s_sympathy	n_total	n_response	n_agree	t_total	t_description	t_objection	t_agree	t_say
mean(s)	49.30	15.15	22.03	5.67	6.43	31.06	18.40	12.66	72.63	45.07	12.55	8.78	6.21
median(s)	45.67	10.22	15.41	2.56	2.49	27.08	14.11	10.11	67.04	38.54	4.01	7.30	2.63
SD(s)	31.94	16.82	21.82	8.06	10.65	21.41	16.65	10.23	35.31	24.31	19.21	7.94	8.53
max(s)	128.06	91.83	85.82	36.73	51	102.70	99.93	39.58	154.56	122.59	97.51	30.17	33.35

Table 4: Statistics for a duration of single behavior

statistics	s_total	s_response	s_interesting	s_agree	s_sympathy	n_total	n_response	n_agree	t_total	t_description	t_objection	t_agree	t_say
count(time)	906	324	352	112	118	1404	859	545	1143	481	181	315	166
mean(s)	3.25	2.78	3.77	3.23	3.06	1.44	1.39	1.53	4.01	6.07	4.09	1.74	2.28
median(s)	2.75	2.36	3.06	2.96	2.36	1.11	1.07	1.2	1.96	2.98	2.44	1.32	1.73
SD(s)	2.4	1.85	2.79	1.72	2.76	1.19	1.12	1.29	6.14	8.35	5.08	1.29	1.47
max(s)	29.07	12.76	29.08	9.39	22.72	17.37	10.97	17.37	61.25	61.25	39.63	7.99	7.2

For OpenFace, as shown in the Figure 7, we focused on the angle of the face and visualized them as a time series waveform graph. For LPMS-B2, as shown in Figure 4, we focused on three-axis acceleration and three-axis angular acceleration and visualized as a time series waveform graph. In this paper, we analyzed data on two persons. Figure 15 and Figure 16 show the graphs of time series data including the moment when a nodding occurred.

Figure 15 and 16 show graph of time-series behavior data of two different participant. Each Figure 15(a), 16(a) is the output about the face angle of OpenFace, Figure 15(b), 16(b) is the output of the accelerations of LPMS-B2, Figure 15(c),

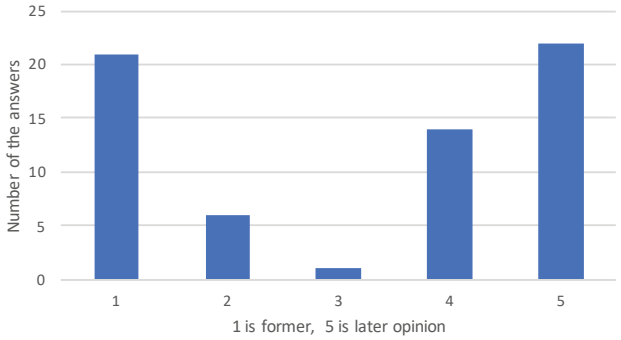

Figure 9: A1: Opinion bias before discussion

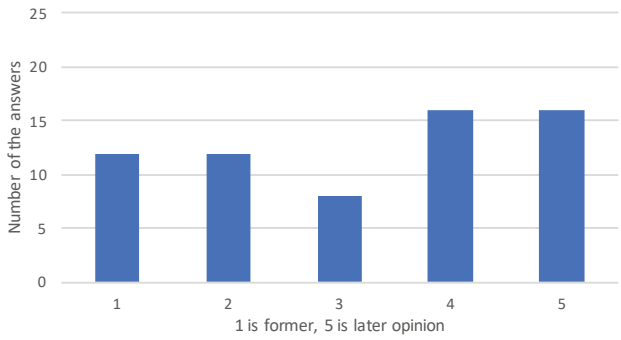

Figure 10: A2: Opinion bias after discussion

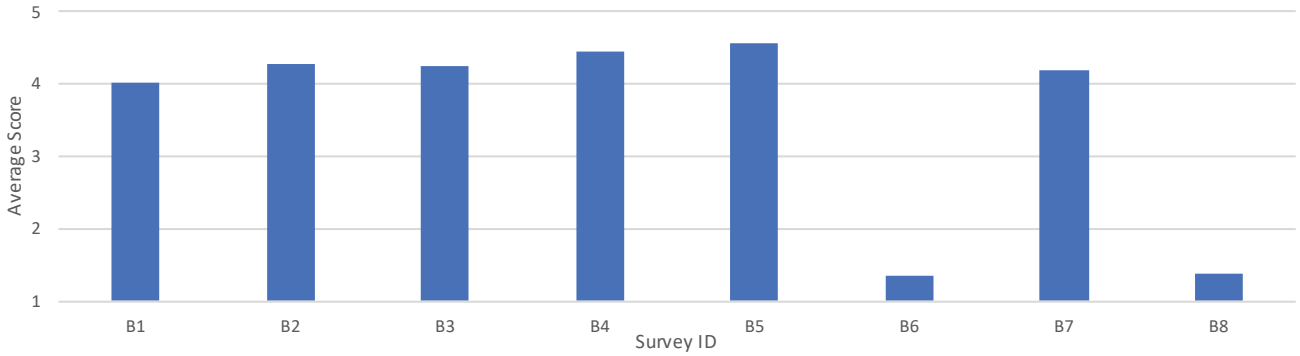

Figure 11: B1-B8: Average scores for each question

16(c) is the output of the angular accelerations of LPMS-B2. Each blue square wave shows the part labeled as nodding. The participant No. 1's nodding generated multiple nodding in succession. We can confirm that this situation is reflected in the waveforms of pose_Rx in OpenFace and AccX, AccY, GyroZ in LPMS-B2 regarding Figure 15. Hence, we can confirm that it is possible to recognize nodding in the actual group discussion from the 360-degree cameras' video, head acceleration and angular acceleration data. On the other hand, from Figure 16, we can confirm that it is difficult to clarify the relationship between the waveform and the labels because the interval of the nodding is very short. As in the two examples listed here, when we confirmed the other participants' data, the type of nodding-behavior is divided into a group of people who nods continuously and who nods momentarily. Although many participants belong to former group, it is necessary to consider how to try to recognize the micro-behavior of the latter group as well.

In labeling, it was difficult for the person to judge whether or not he/she was a nodding. There were individual differences in the size and speed of the nodding, depending on the person.

Missing Data

LPMS-B2 needs to be manually assigned an ID for the reason of the provided software. Because a human error caused an error in the ID assignment of the acceleration sensor, there is a case where the sensor data of the two LMPS-B2 was mixed.

Figure 17 and Figure 18 show a part of the video shot with Pupil. As shown in Figure 17, the green point represents the gaze point of the person wearing Pupil, and in this case, it is understood that the person on the right is in focus. However, as shown in the Figure 18, it was found that Pupil cannot acquire eye movement correctly when user performs action such as smiling, that causes eyes to narrow.

5 M3B CORPUS AVAILABILITY

We make our *M3B Corpus* available for research community only for academic non-commercial purposes. It is available online at http://data.ubi-lab.com/M3B/ for researchers. We delete face image and audio data from video to protect private information. The videos included in *M3B Corpus* are the ones after applying OpenFace to videos recorded by THETA V. The extension of these videos are *.avi*. labels, face point cloud data and head motion data are in csv format.

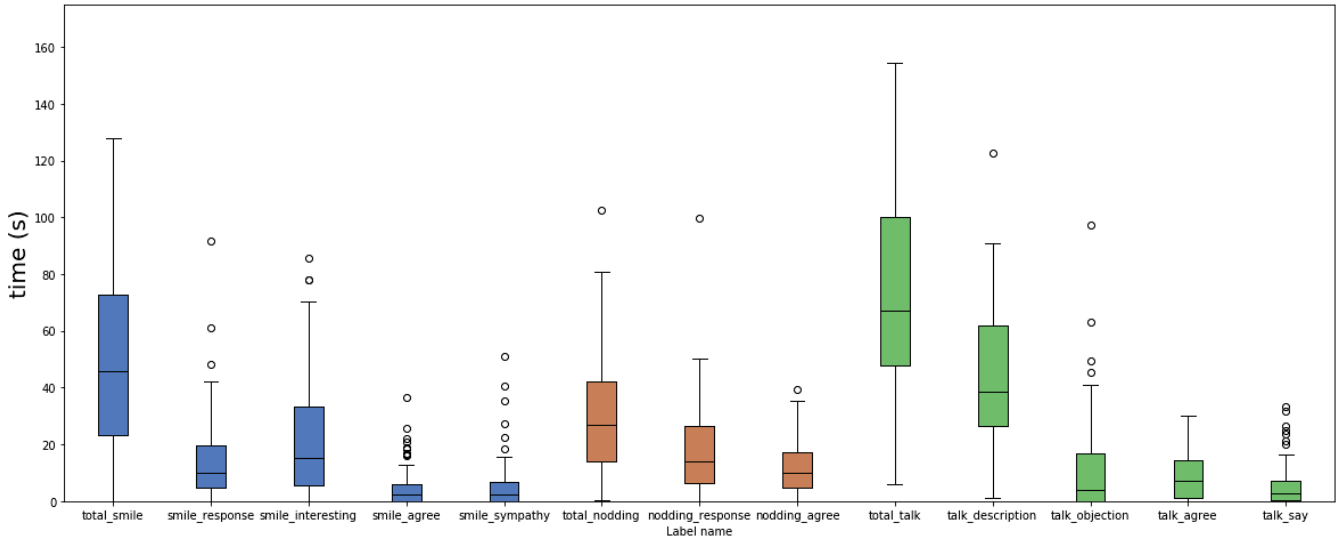

Figure 12: Statistical Labels

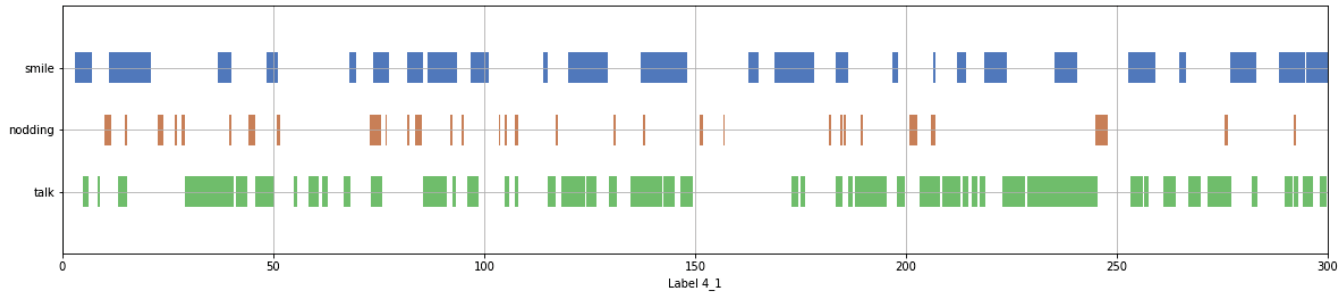

Figure 13: Labeling time series about one person

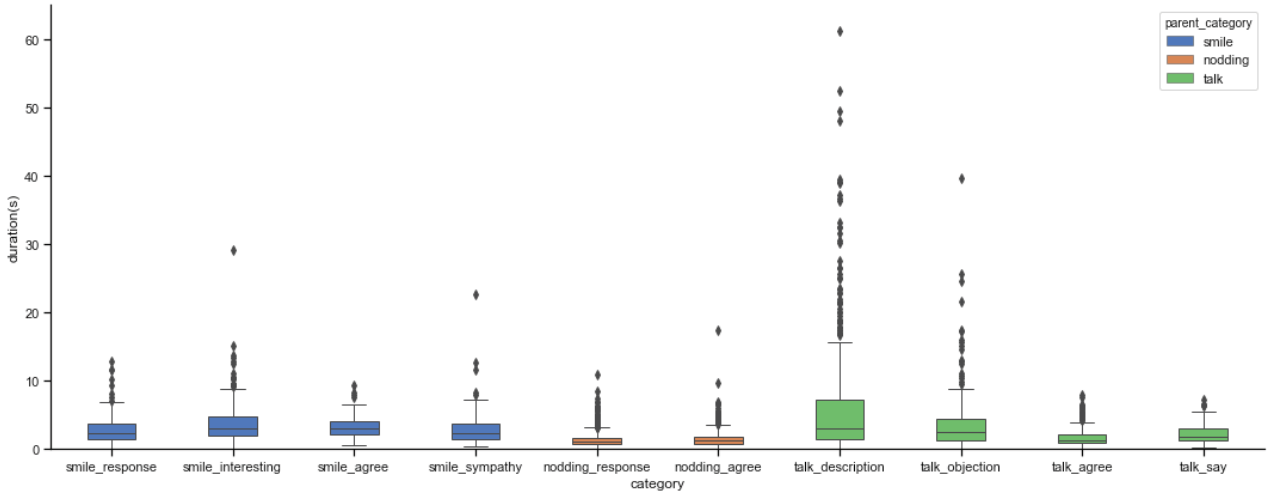

Figure 14: Duration of behavior

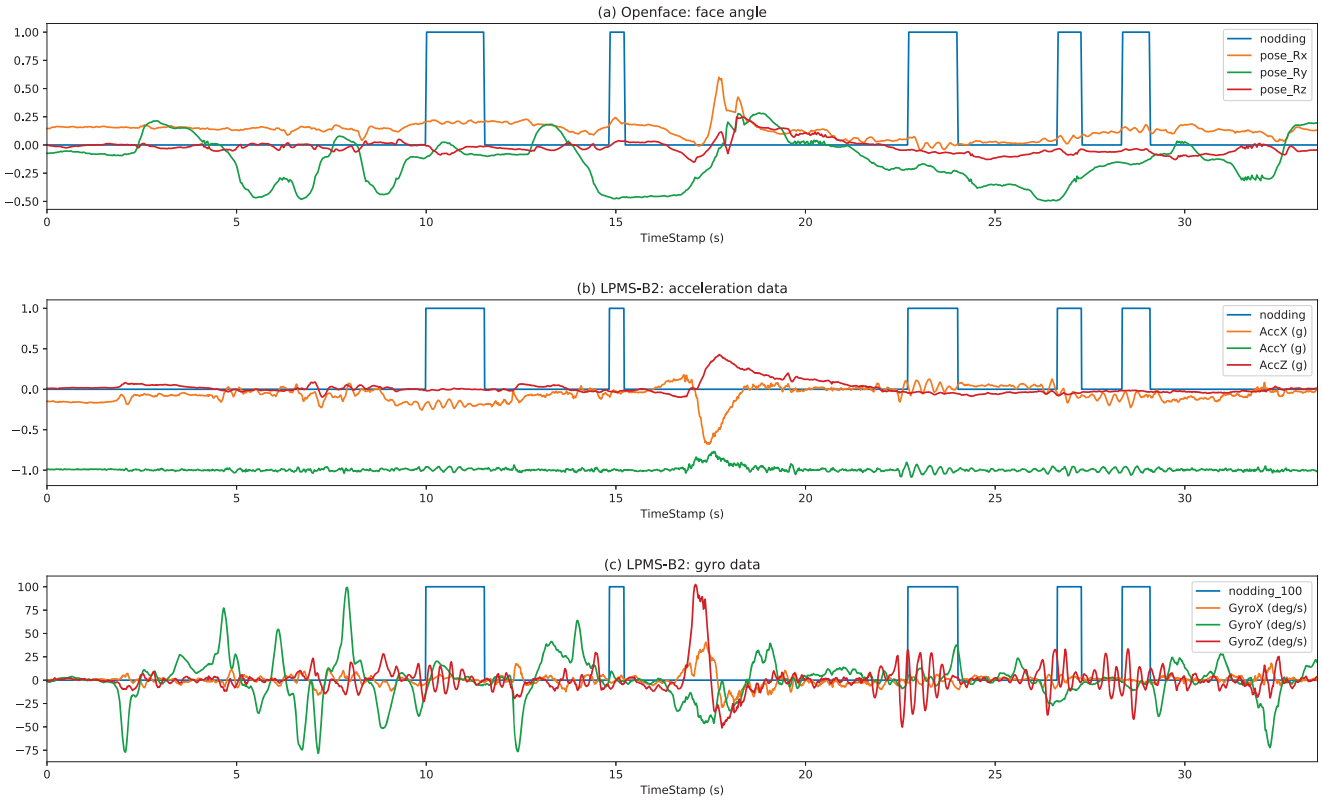

Figure 15: Participant no.1's data

6 FUTURE WORK

We will try to recognize micro-behavior that occur during group discussion among multiple people by using *M3B Corpus*. Furthermore, we will analyze the posture and gesture of the body and label the person who was watching from Pupils' video, and increase the contents of the corpus. Finding out if there is a correlation between the micro-behavior during the discussion and the degree of engagement, and we aim to construct a system that supports group discussions in the future.

7 CONCLUSION

The need for analysis on multi-modal communication such as group discussion is considered to be required in the future. We created a corpus of 16 different discussions from 5 minutes of discussion in a group of 4 people, including labeling data. We named this corpus "M3B." This corpus will be published for research purpose in the form of concealing personal information. Based on the statistics and sensor data time-series waveform, we checked the micro-behavior that occurred during the meeting and investigated how they corresponded actually with the nodding. By using *M3B Corpus*, we expect that the analysis of multi-modal communication will be performed and the discussion on human-to-human communication will take place more.

ACKNOWLEDGMENTS

This research is partially supported by JST PRESTO and Initiative for Life Design Innovation (iLDi) Platform for Society 5.0.

REFERENCES

[1] K. Murano A. Onishi and T. Terada. 2019. A method for structuring meeting logs using wearable sensors. *Internet of Things* (2019), 140–152. https://doi.org/10.1016/j.iot.2019.01.005
[2] The Language Archive. 2002. ELAN. Retrieved June 17, 2019 from https://tla.mpi.nl/tools/tla-tools/elan/
[3] Peter.E Bull. 1987. *Posture and Gesture.* Pergamon Press.
[4] Arpan. C. 2017. Active Learning through Discussion. (2017), 1–4.

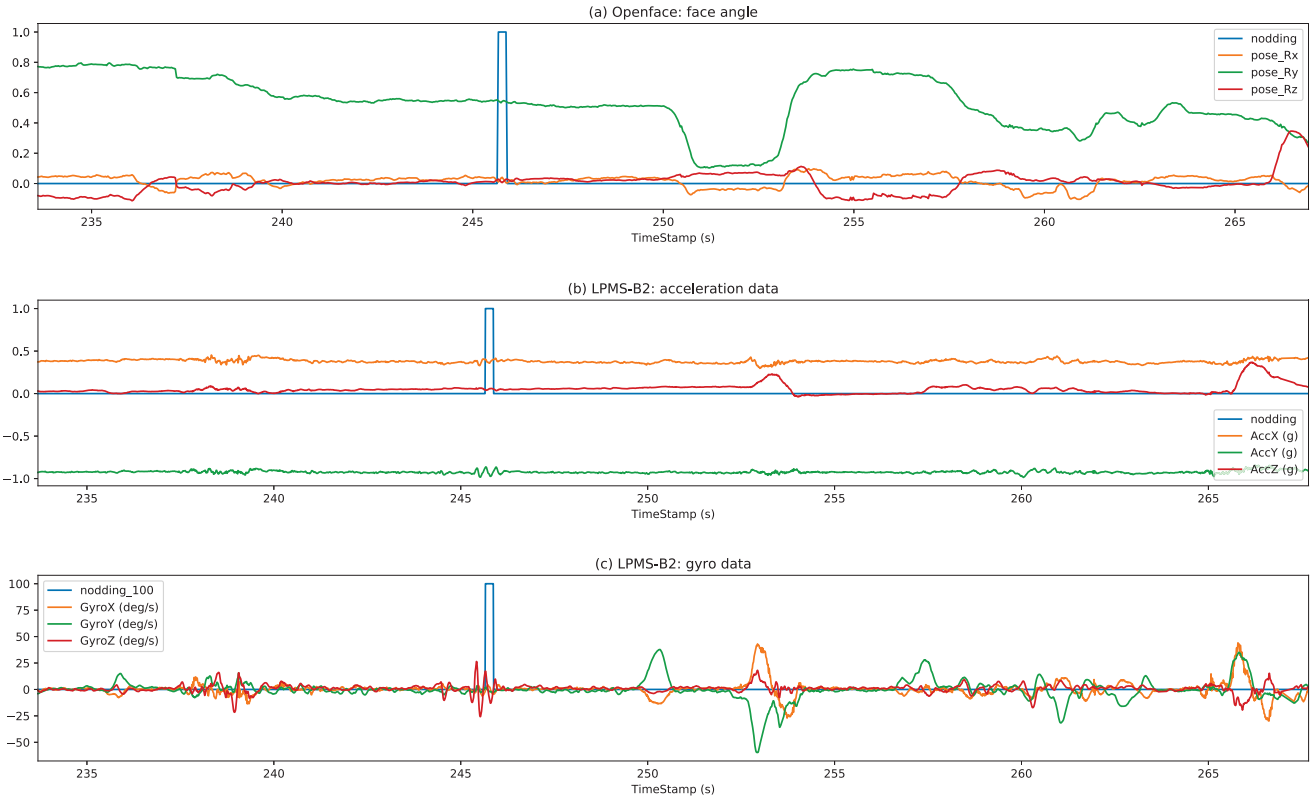

Figure 16: Participant no.2's data

Figure 17: pupil works success

Figure 18: pupil works failed

[5] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing* (2017).

[6] LP-RESEARCH Inc. 2019. LPMS-B2. Retrieved June 17, 2019 from https://lp-research.com/lpms-b2/

[7] Masashi Inoue, Toshio Irino, Nobuhiro Furuyama, Ryoko Hanada, Takako Ichinomiya, and Hiroyasu Massaki. 2011. Manual and Accelerometer Analysis of Head Nodding Patterns in Goal-oriented Dialogues. In *International Conference on Human-Computer Interaction*. Springer, 259–267.

[8] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. 2013. The vernissage corpus: A conversational human-robot-interaction dataset. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 149–150.

[9] Pupil Labs. 2017. Pupil Mobile Eye Tracking Headset. Retrieved June 17, 2019 from https://pupil-labs.com/pupil/

[10] P.Ekman and L. W. Friesen. 1969. The repertoire of nonverbal behavior. *Semiotica* (1969), 49–98. https://doi.org/10.1515/semi.1969.1.1.49

[11] RICOH. 2017. THETA V. Retrieved June 17, 2019 from https://theta360.com/en/about/theta/v.html

[12] Philipp M Scholl and Kristof Van Laerhoven. 2016. A multi-media exchange format for time-series dataset curation. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 715–721.

[13] T. Yoshioka M. Fujimoto S. Watanabe T. Oba A. Ogawa K. Otsuka D. Mikami K. Kinoshita T. Nakatani A. Nakamura T. Hori, S. Araki and J. Yamato. 2012. Low-latency Real-Time Meeting Recognition and Understanding Using Distant Microphones and Omni-directional Camera. *IEEE Transaction on Audio, Speech, and Language Processing* 20, 2 (2012), 499–513. https://doi.org/10.1109/TASL.2011.2164527

[14] Peter Robinson Tadas BaltruÅąaitis and Louis-Philippe Morency. 2016. OpenFace: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision* (2016). https://doi.org/10.1109/WACV.2016.7477553