Annotation and analysis of listener's engagement based on multi-modal behaviors

Koji Inoue
Kyoto university
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Divesh Lala
Kyoto university
Kyoto, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Shizuka Nakamura
Kyoto university
Kyoto, Japan
shizuka@sap.ist.i.kyoto-u.ac.jp

Katsuya Takanashi
Kyoto university
Kyoto, Japan
takanasi@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Kyoto university
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

We address the annotation of engagement in the context of human-machine interaction. Engagement represents the level of how much a user is being interested in and willing to continue the current interaction. The conversational data used in the annotation work is a human-robot interaction corpus where a human subject talks with the android ERICA, which is remotely operated by another human subject. The annotation work was done by multiple third-party annotators, and the task was to detect the time point when the level of engagement becomes high. The annotation results indicate that there are agreements among the annotators although the numbers of annotated points are different among them. It is also found that the level of engagement is related to turn-taking behaviors. Furthermore, we conducted interviews with the annotators to reveal behaviors used to show a high level of engagement. The results suggest that laughing, backchannels and nodding are related to the level of engagement.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI);

Keywords

Engagement; Human-machine Interaction; Behavior; Annotation; Turn Taking; Multi-modal

1. INTRODUCTION

In recent years, many studies on human-machine interaction have been conducted in various scenarios such as user assistance systems and conversational robots. These systems can correctly reply to what they are asked [13, 32, 19]. Furthermore, some systems effectively interact with users in specific tasks like board games [30] and medical diagnoses [9]. However, the interaction is human-machine specific and is much different from that of human-human conversation. For example, the systems often do not start talking by themselves, and just wait for user input. The user utters slowly and carefully to make sure the utterance is correctly recognized by the system, and sometimes the utterance is simply a keyword command. To make the systems further pervade many aspects of our lives, the interaction between the system and user needs to be more natural like those of human beings.

To establish this communication between the system and user, they need to properly build engagement which represents the process where participants establish, maintain, and end their interaction [7]. Especially, in the field of human-machine interactions, this concept focuses on the user state (user engagement) which is how much a user is being interested in and willing to continue the interaction (as detailed in Section 2). By estimating the level of user engagement, the system can appropriately control its actions according to this level. For example, the system can change the conversational topic or its turn-taking behavior to increase the user's satisfaction. This function is one of abilities called "social skills" [6]. It is desirable for the system to have social skills which lead to symbiotic human-machine interaction.

In this paper we conduct an annotation of the level of engagement in natural one-on-one conversations. Through this annotation work, we comprehensively examine the relationship between the level of engagement and cues which trigger changes of engagement. For the cues we focus on behaviors generated by human subjects. These behaviors consist of not only what is spoken but also other behaviors such as laughing, backchannel, nodding, eye-contact and so on. In daily conversation we use various behaviors to convey our inner states each other. We aim to unveil the relationship between the level of engagement and its related behaviors. Our research goal is to estimate the level of engagement based on various behaviors generated by users. The result of the related behaviors is also useful for the system to represent

MA3HMI'16 , November 16 2016, Tokyo, Japan

ISBN 978-1-4503-4562-0/16/11...$15.00

DOI: http://dx.doi.org/10.1145/3011263.3011271

its own engagement toward the users through an embodied interface.

The rest of this paper is organized as follows. Section 2 briefly reviews related works on engagement from the perspective of both estimation and annotation. Section 3 describes the conversational data used in our annotation work and the annotation procedure. The annotation result and analysis on turn-taking behaviors are presented in Section 4. The interview results from the annotators are also discussed in the section. Section 5 concludes this paper with future work.

2. RELATED WORK

In this section, we briefly review related work which deals with engagement, staring with the definition of engagement. Engagement is originally defined in facial communication as "joining each other openly in maintaining a single focus of cognitive and visual attention" [12]. From the perspective of the communication process, it is also defined as "the process by which two (or more) participants establish, maintain, and end their perceived connection" [29]. In this manner, the meaning of engagement is close to that of attention and involvement [11]. Engagement has been frequently discussed and studied in the field of human-machine interaction, introduced as "the process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume or terminate an interaction" [2]. In this research field, many studies focus more on the user state, namely the user engagement. For example, it is described as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" [25] or "how much a participant is interested in and attentive to a conversation" [34].

Detection of engagement has been widely investigated in many studies. It was mostly formulated as a binary classification problem: engaged or not (disengaged). Since acoustic information is not robust against background noise, most of the methods are based on non-verbal information. The majority of the methods made use of eye-gaze information [3, 21, 24]. Eye-gaze behavior is frequently used to express engagement [18]. Besides, head/body pose and spatial features are also used in some models [17, 20]. Most of the methods to detect engagement are designed by heuristic rules. It is desirable that the detection model is learned from human conversational data. Machine learning is an effective way to model the complex structure between engagement and signal data captured by sensors. However, these require a large amount of data and effective annotation is needed to obtain labels of the level of engagement.

For machine learning, some works have conducted annotations of engagement with Wizard-of-Oz conversation data. During a conversation where a user interacts with an agent which introduces mobile phone devices, the disengagement period was annotated by multiple third-party annotators [23]. Disengagement is where the user feels bored and wishes to exit the interaction. In conversational systems, detection of disengagement is important for the systems to recover and keep the interaction as long as possible. The annotation data was used to learn an N-gram model of eye-gaze behavior to estimate user disengagement. Another scenario was a conversation where a human subject talked with a robot guiding them in a museum [33]. Third-party annotators were asked to judge the level of engagement, especially the existence of engagement intentions. Engagement intention is defined as a binary state where a positive value meant that the user wanted to start a conversation or intended to take the speaking floor. This data was used to learn a support vector machine to detect the engagement intention from visual information.

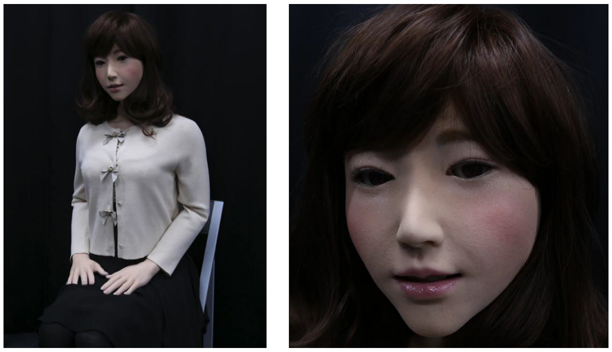


Figure 1: Outlook of android ERICA

A few studies have been conducted using human-human conversation data. Annotators were asked to detect the change point of the discrete level of engagement in a multiparty conversation among three levels: minus, neutral, and plus [5]. This annotation work was done for the whole group and for each participant. A study was also conducted for a remote conversation, and the level of engagement was categorized into six levels, ranging from no interest to governing or managing discussion [1]. The annotation was done in 15 seconds segments. To conduct annotation work for engagement, we need to consider several factors such as simplicity for the annotators, richness of the information, and consistency among annotators.

In this paper, we present annotation of engagement for conversations between a human and a teleoperated android. We examine various behaviors used to show the level of engagement. We assume that the behaviors are intermediate states between engagement and the signal information captured by sensors, and these behaviors are human-understandable, such as laughing, nodding and backchannels. This annotation work will enable us to understand the structure between engagement and its related behaviors. Note that the tendency of behaviors generated by a participant could be affected by their role, as a speaker or listener. We focus on the listener's behaviors because we can restrict the variety of the behaviors to only non-lexical information.

3. DATA

The conversation data used in this annotation work is a one-on-one conversation where a person talks with an android ERICA which is remotely operated. In this section, we introduce the android ERICA, followed by a description of the conversational data and the annotation procedure.

3.1 Android ERICA

An autonomous android, named ERICA (ERato Intelligent Conversation Android), who has the appearance of a human being has been developed [10]. ERICA is depicted in Figure 1. She has 19 active joints to express various behaviors such as facial expressions, gaze, and nodding. In our

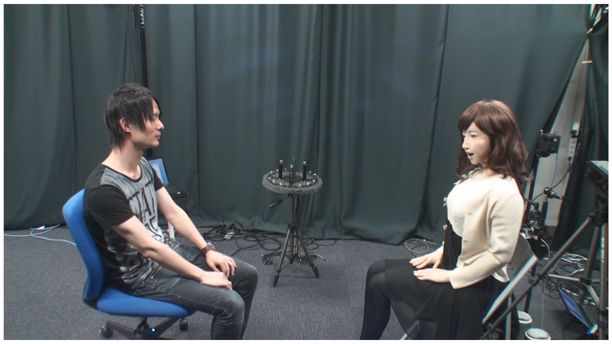


Figure 2: Scene of data collection (Camera 1)

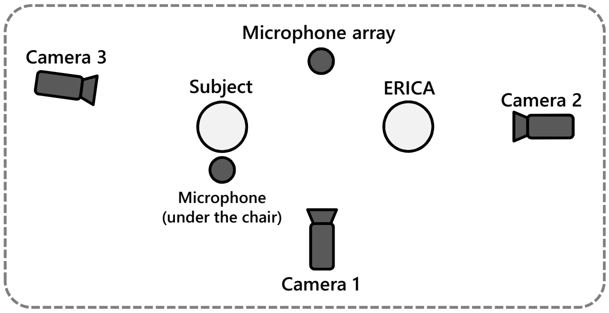


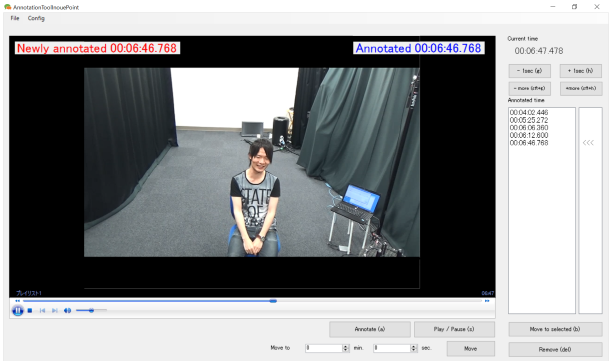
Figure 3: Top view of sensor position

data collection, some of these behaviors were controlled by a human operator in a remote place. We also defined a character of ERICA in order to smoothly interact with human subjects. ERICA is a 23 year-old woman, and is given a specific social role according to the conversational situation [14]. In this study, her role is a secretary of a laboratory.

3.2 Corpus of human-robot interaction

We first collected conversation data where ERICA interacts with a human subject. ERICA is operated by a human operator who remained the same throughout the data collection period. The voice uttered by the operator was played in real time with a speaker deployed in next to the body of ERICA. The lip motion of ERICA is automatically generated from the prosodic information of the operator's voice [16]. The operator also controlled the head and eye-gaze motion of ERICA to express eye-contact and nodding. During conversation the subject and ERICA sat on chairs facing each other. A scene of the conversation is shown in Figure 2.

The conversational scenario was as follows. ERICA meets the subject for the first time and they greet each other. We asked the operator to follow two conversational phases. After the greeting, they would chat about personal topics such as their hometowns and hobbies. This phase of ice-breaking is frequently observed in daily conversations. We provided the operator with a list of topics which they could chat about. Afterwards, they moved to the second phase where they talked more specifically about android robots. In this phase, they talked about their impressions of android robots or the functions of ERICA. We set the total time of the interaction to about 10 minutes, but did not restrict the exact time, and the end of the conversation was decided by the operator. The operator also decided when to shift from the first to the second phase.



Newly annotated 00:06:46.768

Annotated 00:06:46.768

Current time
00:06:47.478

Annotated time
00:04:02.446
00:05:25.272
00:06:06.360
00:06:12.600
00:06:46.768

Move to 0 min. 0 sec.

Figure 4: Annotation tool (with Camera 2)

We recruited 15 undergraduate students as subjects (4 females and 11 males), who were all Japanese native speakers. All conversations were recorded by multi-modal sensors including a single-channel microphone, a 16-channel microphone array, and three HD cameras. Figure 3 illustrates the sensor positions. This data was used to annotate utterances, turns, topics, backchannels [8] and clause boundaries which represent semantic pauses inside the utterances [31].

After the conversation, we took a survey from each subject about their impressions of the conversation and ERICA. The survey consisted of 32 questions in 7-point Likert scale (from 1 to 7). The surveys provided some results which are thought to be related to engagement. Firstly, the level of interest to the conversation was positive (M=5.73, SD=0.85). The level of continuity also showed a similar tendency (M=5.40, SD=1.36). For the emotional aspects, positive scores were obtained for enjoyment towards the conversation (M=6.07, SD=0.57) and empathy (M=5.53, SD=0.81). These results suggest that the subjects were engaged in the conversations. On the other hand, naturalness of behaviors generated by ERICA were mixed. Although there was no awkwardness about the spoken content (M=6.20, SD=0.75) and the turn taking (M=5.93, SD=0.68), eye-gaze behavior (M=3.60, SD=1.58) and the face direction (M=3.86, SD=1.50) were seen as unnatural. Since these behaviors were manually controlled by the operator, this could be due to a high task load for the operator. In the future, we plan to make these behaviors automatically operated by coordinating with multiple sensors.

3.3 Annotation procedure

For the annotation of engagement, we chose 8 out of 15 sessions described in the previous section. We recruited 6 annotators (1 female and 5 males) who had not participated in the conversations. Each session was assigned to three annotators, and each annotator worked with some sessions (at most 5).

The instruction given to the annotators was as follows. At first, the definition of engagement in this study was presented. Following the past studies reviewed in Section 2, we defined it as "How much a user is being interested in and willing to continue the interaction". The annotators were asked to watch the video and press a button when they noticed that the level of the subject's engagement switched to high while listening to the talk given by ERICA. Since this

Table 1: The number of annotated points

session	duration	annotator						
		A	B	C	D	E	F	ave.
20150528-01	8m 48s	10	17	-	6	-	-	11.0
20150528-02	11m 22s	11	-	13	-	7	-	10.3
20150528-03	9m 41s	-	16	9	-	-	25	16.7
20150528-04	11m 34s	6	18	-	-	5	-	9.7
20150528-05	10m 06s	-	25	-	6	9	-	13.3
20150528-06	10m 29s	-	-	18	6	10	-	11.3
20150528-07	11m 41s	-	23	23	5	-	-	17.0
20150528-08	7m 27s	8	-	-	6	6	-	6.7
ave.		8.8	19.8	15.8	5.8	7.4	25.0	

Table 2: The number of agreement turns

session	#agreement			#ERICA
	three	two	one	turn
20150528-01	2	6	14	64
20150528-02	1	4	17	71
20150528-03	7	4	10	53
20150528-04	1	3	18	74
20150528-05	3	6	15	67
20150528-06	0	9	14	77
20150528-07	3	10	16	59
20150528-08	3	2	4	46
ave.	2.5	5.5	13.5	63.9

judgement was done for the period when the subject was listening to the talk, the annotators needed to focus on non-verbal information. Therefore, we gave a list of non-verbal behaviors expected to be generated by the subject. The list includes facial expression, eye-gaze, backchannel, nodding, body pose, and gesture. The video the annotators observed was the view covering the front of subject (Camera 2 in Figure 3) so that the annotators can focus on the behaviors of the subject. The video view and annotation software used for this work are shown in Figure 4. After the annotation work on each session, we took a survey and interview from each annotator to understand which behavior affected their judgment for engagement. In the interview, each annotator reviewed the annotation result, and was asked the reason why the annotator pressed the button at that point. We recorded the answer given by the annotator, especially the behaviors which the annotator referred to determine high-level engagement.

This methodology has strengths and limitations. This work is simple binary classification and the annotators did not care about timing when the level of engagement switches back from high to not high. Hence this work can be easily done by annotators, which makes us possible to conduct this work with many annotators and many sessions. This is important when we use the result in machine learning approaches. On the other hand, simple binary classification cannot represent various levels of engagement. From surveying the annotators (also 7-point Likert scale, from 1 to 7), it was found that the annotation work was easy (SD=5.67, SD=1.40), but the confidence in their results was slightly decreased (M=4.08, SD=1.55).

4. RESULT

In this section, we report the results obtained through the annotation work.

4.1 Agreement among annotators

The number of annotated points given by each annotator on each session is summarized in Table 1. It is observed that these numbers differ among annotators. For example, although the annotators B and C tended to frequently find engagement, annotators D and E rarely responded. On the other hand, the difference among the sessions was smaller than the inter-annotator differences. From this result, it is clear that each annotator has different criterion and thresholds for engagement.

The time-wise distributions of annotated points on four sessions (150528-01, 150528-02, 150528-03, and 150528-08) are shown in Figure 5, 6, 7, and 8, respectively. Note that "phase" in the figures represents the two conversational phases: chatting and talking about android robots. In these sessions, it can be seen that there are some agreements among annotators. In the two session (150528-02 and 150528-03), the distribution of annotated points is substantially uniform, while in the other two sessions (150528-01 and 150528-08), the annotated points are frequently observed in the second phase. The observations in the latter two sessions (150528-01 and 150528-08) can be explained by the effect of the ice-breaking process. At first the subject could be nervous and passively attended to the conversation. After they talked about personal topics, the subject becomes more relaxed and becomes more active in the conversation.

We also tabulated the agreement points among the annotators. To identify agreement, we use the conversational turn as an unit. If there are multiple annotated points given by different annotators in the same turn held by ERICA, those points are regarded as an agreement. We also relate an annotated engagement point in the turn held by a subject was associated to those of the previous ERICA's turn. In a few cases the same annotator pressed the button more than twice in the same turn, so there is an inconsistency between the total numbers of Table 1 and 2. The number of agreement points are shown in Table 2. It is observed that there are measurable agreements among the annotators.

4.2 Relationship with turn-taking behaviors

We further analyzed the relationship between the level of subject's engagement and his/her actions. Previous works have found that engagement is related to the motivations

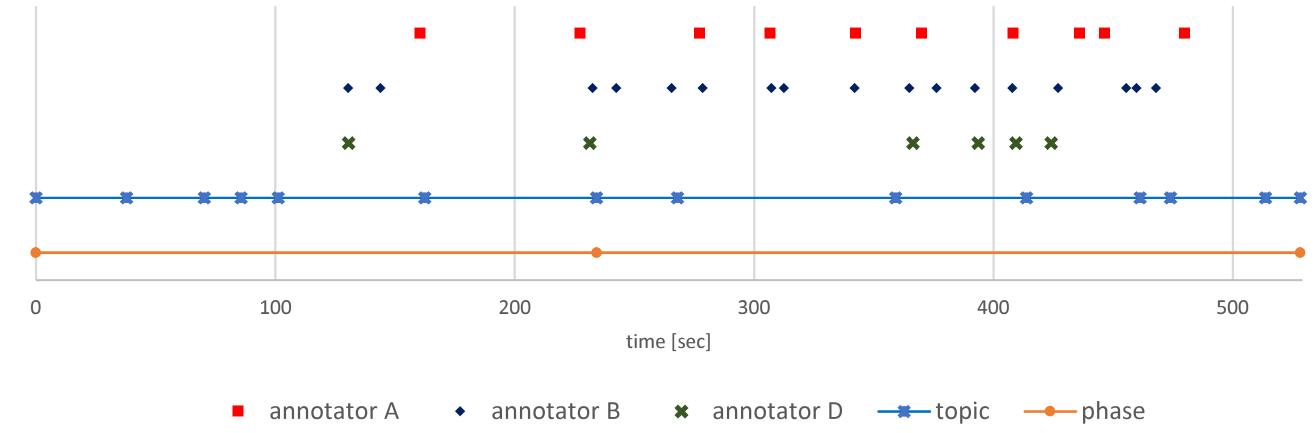


Figure 5: Time distribution of annotation result (session 150528-01)

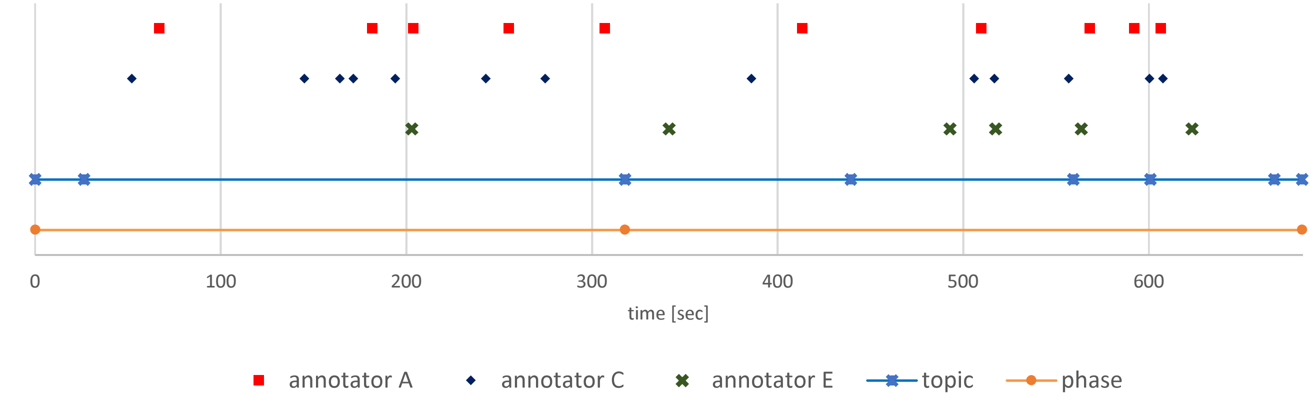


Figure 6: Time distribution of annotation result (session 150528-02)

of turn-taking [4, 33]. If the level of the subject's engagement is high, the subject may be willing to take the conversational turn. On the other hand, if the level is low, the system needs to continue to speak and attract the subject's attention. To verify this assumption, we compared two sets, engaged and not engaged, from all the sessions together. The set of engaged samples were annotated as the high level of engagement by more than two annotators. The set of not engaged samples consists of all the turns where no annotator marked them as engaged. For these two sets we examined the relationship between the level of engagement and the next action as illustrated in Figure 9. Firstly, the duration of the following subject's turn is summarized in Table 3. The duration of the next turn in the engaged set was about twice as long than that of the not engaged set. We then compared the numbers of backchannels given by ERICA in Table 4 and found that the engaged set had a larger proportion of backchannels from ERICA. These results suggest that the level of engagement could reflect the subject's willingness to speak more and hold the turn, and also ERICA's intention to not take the turn from them.

To examine turn-taking behaviors more reliably, we should take into account the transition relevance place (TRP) where the turn can be exchanged [26]. In this analysis, transition relevance places are approximated by clause boundaries. Here, we use absolute and strong boundaries from the definition of clause boundaries [31]. We tabulated the number of clause boundaries inside the subject's utterance, where the subject held the turn and ERICA did not take the turn in the TRP. Table 5 indicates that the proportion of clause boundaries is larger when the subject was engaged. This result therefore supports the assumption that engagement is related to turn-taking behaviors.

4.3 Relationship with multi-modal behaviors

We interviewed each annotator to know which behaviors were the cues to trigger a high level of engagement. This interview work was done with 2 sessions (150528-03 and 150528-08). For every annotated point, each annotator was asked to explain the reason he/she pressed the button, then we identified referred behaviors from the explanation after the interview. Table 6 and 7 show the number of referred behaviors by each annotator. Table 8 and 9 indicate the number of agreements. These agreement numbers represent how many annotators were affected by the same behavior to annotate the high level of engagement. Note that the list of behaviors in the tables is based on the behavior list which we gave to the annotators, but some behaviors which were not referred at all in the interview are omitted in the tables. Most of the time, facial expression was meant to laughing in the explanations given by the annotators.

The following discussion refers to the first session (Ta-

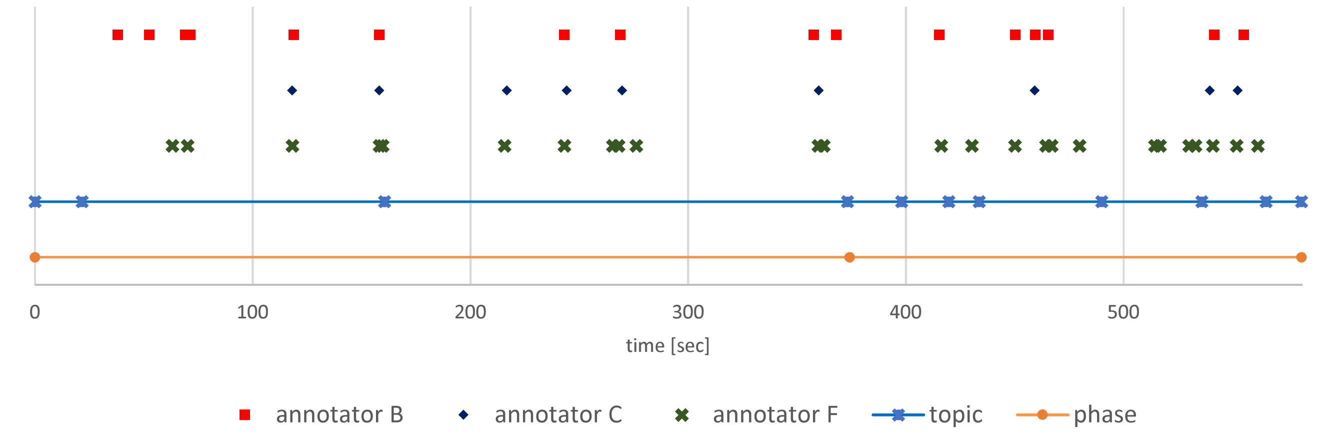


Figure 7: Time distribution of annotation result (session 150528-03)

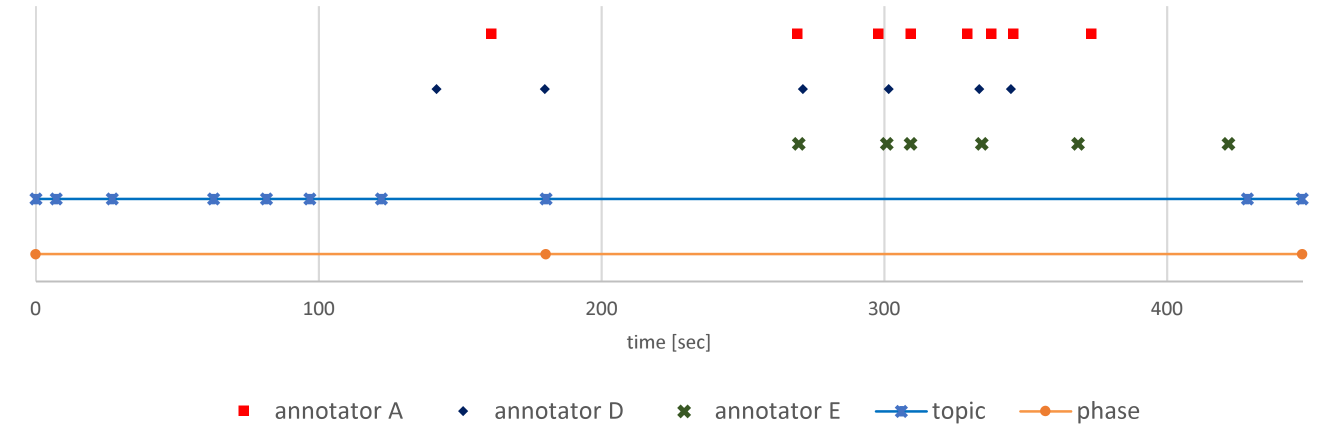


Figure 8: Time distribution of annotation result (session 150528-08)

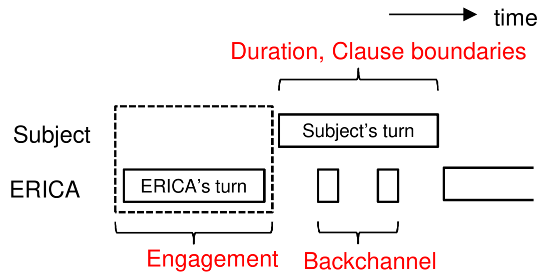


Figure 9: Analysis scope

Table 3: Average duration of subject's turn [sec.]

	duration
engaged	6.08
not engaged	3.54

ble 6 and 8). Annotators B and F gave many annotated points, but annotator C gave less annotated points than the others. There could be a difference in the sensitivity to behaviors related to a high level of engagement among annotators. Annotators B and C uniformly referred to all the

Table 4: The number of backchannel given by ERICA (ratio to the number of subject's turns)

	#backchannel	#subject turn
engaged	87 (1.36)	64
not engaged	270 (0.80)	339

Table 5: The number of clause boundaries (CBs) inside the subject's turn (the ratio to the number of all the CBs appeared in the subject's turn)

	#CBs inside	#all CBs
engaged	41 (0.48)	86
not engaged	89 (0.27)	334

behaviors except gesture, with laughing being an effective indicator of engagement. On the other hand, the annotator F frequently referred to laughing, nodding, and body pose, with gaze and backchannels not seen as important. For the number of agreements (Table 8), laughing seems to be a crucial cue to indicate a high level of engagement. Nodding and body pose could be also influential, with spoken content sometimes being relevant. Backchannels were referred

Table 6: The number of referred behaviors (session 150528-03)

behavior	annotator			
	B	C	F	total
facial expression (laughing)	10	7	17	34
gaze	5	4	1	10
backchannels	4	2	1	7
nodding	6	2	15	23
body pose	4	2	14	20
gesture	1	1	6	8
spoken content	8	4	2	14
#annotated points	16	9	25	-

Table 7: The number of referred behaviors (session 150528-08)

behavior	annotator			
	A	D	E	total
facial expression (laughing)	2	2	4	8
gaze	0	2	0	2
backchannels	6	3	4	13
nodding	4	3	5	12
body pose	1	1	0	2
spoken content	0	1	2	3
#annotated points	8	6	6	-

to a few times, but there is a three-agreement point which might indicate a specific backchannel to show a high level of engagement.

In the second session (Table 7 and 9), the annotators A and E show a similar tendency. They frequently referred to laughing, backchannels, and nodding except that annotator E was also affected by spoken content twice. Gaze was not referred by them. On the other hand, annotator D referred to gaze as well as the above three behaviors. In the number of agreements (Table 9), backchannels and nodding showed high frequency among the annotators. Laughing has only one three-agreement point, but it could be a crucial cue.

Comparing the results of the two sessions, we can see some intersections: laughing, backchannels, and nodding could be effective behaviors for showing a high level on engagement. On the other hand, the effect of gaze and body pose was different among the annotators. It is not clear if this difference is caused by the difference of the annotators or the subjects. To verify this matter, we need further annotation and analysis.

5. CONCLUSION

In this paper, we have addressed the annotation of high engagement. Our annotation procedure was quite simple, so we could conduct the annotation with many sessions and annotators. In this study we obtained the annotation results from 24 annotation trials with 8 conversational sessions, and the interview results from 6 trials with 2 sessions. The annotation results showed that the level of engagement is related to turn-taking behaviors. A subject in a high level of engagement tends to take the turn, and the interlocutor also tends not to take the turn but gives backchannels. These results support the hypothesis that the estimation of engagement could contribute towards realizing smooth turn-taking between a system and a user. Furthermore, the interview results indicated that some behaviors are effective to convey a high level of engagement: laughing, backchannels, and nodding. In future work, these behaviors could be useful to automatically estimate the level of engagement. We also need to implement automatic detection of the behaviors themselves from the signal obtained from sensors [15, 27, 28, 22].

Table 8: The number of agreements of referred behaviors among annotators (session 150528-03)

behavior	#agreement		
	three	two	one
facial expression (laughing)	4	7	8
gaze	0	1	8
backchannels	1	0	4
nodding	0	6	11
body pose	1	2	13
gesture	0	0	8
spoken content	0	4	6

Table 9: The number of agreements of referred behaviors among annotators (session 150528-08)

behavior	#agreement		
	three	two	one
facial expression (laughing)	1	0	5
gaze	0	0	2
backchannels	2	3	1
nodding	3	1	1
body pose	0	1	0
spoken content	0	0	3

6. ACKNOWLEDGMENTS

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project and JSPS KAKENHI Grant Number 15J07337.

7. REFERENCES

[1] R. Bednarik, S. Eivazi, and M. Hradis. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proc. workshop on eye gaze in intelligent human machine interaction*, page 10, 2012.

[2] D. Bohus and E. Horvitz. Models for multiparty engagement in open-world dialog. In *Proc. SIGDIAL*, pages 225–234, 2009.

[3] D. Bohus and E. Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proc. ICMI Workshop on Machine Learning for Multimodal Interaction*, page 5, 2010.

[4] D. Bohus and E. Horvitz. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proc. SIGDIAL*, pages 98–109, 2011.

[5] F. Bonin, R. Bock, and N. Campbell. How do we react to context? annotation of individual and group engagement in a video corpus. In *Proc. PASSAT and SocialCom*, pages 899–903, 2012.

[6] C. Breazeal. Social interactions in HRI: the robot view. *IEEE Trans. on Systems, Man, and Cybernetics*, 34(2):181–186, 2004.

[7] L. Cerrato and N. Campbell. Engagement in dialogue with social robots. In *Proc. IWSDS*, 2016.

[8] P. Clancy, S. Thompson, R. Suzuki, and H. Tao. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of pragmatics*, 26(3):355–387, 1996.

[9] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L. P. Morency. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proc. Autonomous Agents and Multi-Agent Systems*, pages 1061–1068, 2014.

[10] D. Glas, T. Minaot, C. Ishi, T. Kawahara, and H. Ishiguro. Erica: The erato intelligent conversational android. In *Proc. ROMAN*, 2016.

[11] N. Glas and C. Pelachaud. Definitions of engagement in human-agent interaction. In *Proc. International Workshop on Engagment in Human Computer Interaction*, pages 944–949, 2015.

[12] E. Goffman. *Behavior in public places: Notes on the social organization of gatherings*. Simon and Schuster, 1966.

[13] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pages 928–939, 2014.

[14] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara. Talking with erica, an autonomous android. In *Proc. SIGDIAL*, 2016.

[15] K. Inoue, Y. Wakabayashi, H. Yoshimoto, K. Takanashi, and T.Kawahara. Enhanced speaker diarization with detection of backchannels using eye-gaze information in poster conversations. In *Proc. INTERSPEECH*, pages 3086–3090, 2015.

[16] C. Ishi, H. Ishiguro, and N. Hagita. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In *Proc. IROS*, pages 2377–2382, 2012.

[17] Y. Kuno, K. Sadazuka, M. Kawashima, K. Yamazaki, A. Yamazaki, and H. Kuzuoka. Museum guide robot based on sociological interaction analysis. In *Proc. CHI*, pages 1191–1194, 2007.

[18] S. R. H. Langton, R. J. Watt, and V. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in cognitive sciences*, 4(2):50–59, 2000.

[19] Y. Ma, P. A. Crook, R. Sarikaya, and E. Fosler-Lussier. Knowledge graph inference for spoken dialog systems. In *Proc. ICASSP*, pages 5346–5350, 2015.

[20] M. P. Michalowski, S. Sabanovic, and R. Simmons. A spatial model of engagement for a social robot. In *Proc. International Workshop on Advanced Motion Control*, pages 762–767, 2006.

[21] L. P. Morency, C. M. Christoudias, and T. Darrell. Recognizing gaze aversion gestures in embodied conversational discourse. In *Proc. ICMI*, pages 287–294, 2006.

[22] L. P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Proc. CVPR*, 2007.

[23] Y. Nakano and R. Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proce. IUI*, pages 139–148, 2010.

[24] C. Peters. Direction of attention perception for conversation initiation in virtual environments. In *Proc. International Workshop on Intelligent Virtual Agents*, pages 215–228, 2005.

[25] I. Poggi. *Mind, hands, face and body. A goal and belief view of multimodal communication.* Weidler, 2007.

[26] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735, 1974.

[27] H. Salamin, A. Polychroniou, and A. Vinciarelli. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *Proc. SMC*, pages 4282–4287, 2013.

[28] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, and G. Palm. Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Trans. on Interactive Intelligent Systems*, 2(4), 2012.

[29] L. Sidner and M. Dzikovska. Human-robot interaction: Engagement between humans and robots for hosting activities. In *Proc. ICMI*, page 123, 2002.

[30] G. Skantze and M. Johansson. Modelling situated human-robot interaction using IrisTK. In *Proc. SIGDIAL*, pages 165–167, 2015.

[31] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. Identification of "sentences" in spontaneous japanese-detection and modification of clause boundaries. In *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[32] G. Wilcock and K. Jokinen. Multilingual WikiTalk: Wikipedia-based talking robots that switch languages. In *Proc. SIGDIAL*, pages 162–164, 2015.

[33] Q. Xu, L. Li, and G. Wang. Designing engagement-aware agents for multiparty conversations. In *Proc. CHI*, pages 2233–2242, 2013.

[34] C. Yu, P. Aoki, and A. Woodruff. Detecting user engagement in everyday conversations. In *Proc. ICSLP*, pages 1329–1332, 2004.