SIP (2018), vol. 7, e9, page 1 of 16 © The Authors, 2018.

doi:10.1017/ATSIP.2018.11

ORIGINAL PAPER

Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue

KOJI INOUE, DIVESH LALA, KATSUYA TAKANASHI AND TATSUYA KAWAHARA

Engagement represents how much a user is interested in and willing to continue the current dialogue. Engagement recognition will provide an important clue for dialogue systems to generate adaptive behaviors for the user. This paper addresses engagement recognition based on multimodal listener behaviors of backchannels, laughing, head nodding, and eye gaze. In the annotation of engagement, the ground-truth data often differs from one annotator to another due to the subjectivity of the perception of engagement. To deal with this, we assume that each annotator has a latent character that affects his/her perception of engagement. We propose a hierarchical Bayesian model that estimates both engagement and the character of each annotator as latent variables. Furthermore, we integrate the engagement recognition model with automatic detection of the listener behaviors to realize online engagement recognition. Experimental results show that the proposed model improves recognition accuracy compared with other methods which do not consider the character such as majority voting. We also achieve online engagement recognition without degrading accuracy.

Keywords: Engagement, Multimodal, Listener behaviors, Latent variable model, Dialogue

Received 7 February 2018; Revised 11 August 2018

I. INTRODUCTION

Many spoken dialogue systems have been developed and practically used in a variety of contexts such as user assistants and conversational robots. The dialogue systems effectively interact with users in specific tasks including question answering [1, 2], board games [3], and medical diagnoses [4]. However, human behaviors observed during human-machine dialogues are much different from those of human–human dialogues. Our ultimate goal is to realize a dialogue system which behaves like a human being. It is expected that these systems will permeate many aspects of our daily lives in a symbiotic manner.

It is crucial for dialogue systems to recognize and understand the conversational scene which contains a variety of information such as dialogue states and users' internal states. The dialogue states can be objectively defined and have been widely modeled by various kinds of machine learning techniques [5, 6]. On the other hand, the users' internal states are difficult to define and measure objectively. Many researchers have proposed recognition models for various kinds of internal states such as the level of interest [7, 8], understanding [9], and emotion [10–12]. From the perspective of the relationship between dialogue participants (i.e. between a system and a user), other researchers have dealt with entrainment [13], rapport [14–16], and engagement.

In this paper, we address engagement which represents the process by which individuals establish, maintain, and end their perceived connection to one another [17]. Engagement has been studied primarily in the field of human–robot interaction and is practically defined as how much a user is interested in the current dialogue. Building and maintaining a high level of engagement leads to natural and smooth interaction between the system and the user. It is expected that the system can dynamically adapt its behavior according to user engagement, and increases the quality of the user experience through the dialogue. In practice, some attempts have been made to control turn-taking behaviors [18] and dialogue policies [19, 20].

Engagement recognition has been widely studied from the perspective of multimodal behavior analyses. In this study, we propose engagement recognition based on the scheme depicted in Fig. 1. At first, we automatically detect listener behaviors such as backchannels, laughing, head nodding, and eye gaze from signals of multimodal sensors. Recent machine learning techniques have been applied to this task and achieved sufficient accuracy [21]. According to the observations of the behaviors, the level of engagement

Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Corresponding author:
Koji Inoue
Email: inoue@sap.ist.i.kyoto-u.ac.jp

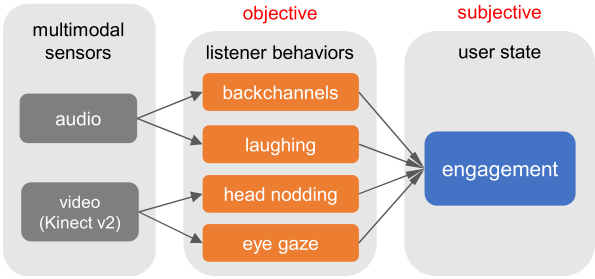

Fig. 1. Scheme of engagement recognition.

is estimated. Although the user behaviors are objectively defined, the perception of engagement is subjective and may depend on each perceiver (annotator). In the annotation of engagement, this subjectivity sometimes results in inconsistencies of the ground-truth labels between annotators. Previous studies integrated engagement labels among annotators like majority voting [22, 23]. However, the inconsistency among annotators suggests that each annotator perceives engagement differently. This inconsistency can be associated with the difference of the character (e.g. personality) of annotators. To deal with this issue, we propose a hierarchical Bayesian model that takes into account the difference of annotators, by assuming that each annotator has a latent character that affects his/her perception of engagement. The proposed model estimates not only engagement but also the character of each annotator as latent variables. It is expected that the proposed model more precisely estimates each annotator's perception by considering the character. Finally, we integrate the engagement recognition model with automatic detection of the multimodal listener behaviors to realize online engagement recognition which is vital for practical spoken dialogue systems. This study makes a contribution to studies on recognition tasks containing subjectivity, in that the proposed model takes into account the difference of annotators.

The rest of this paper is organized as follows. We overview related works in Section II. Section III introduces the human–robot interaction corpus used in this study and describes how to annotate user engagement. In Section IV, the proposed model for engagement recognition is explained based on the scheme of Fig. 1. We also demonstrate an online processing of engagement recognition for spoken dialogue systems in Section V. In Section VI, experiments of engagement recognition are conducted and analyzed. In Section VII, the paper concludes with suggestions for future directions of human–robot interaction with engagement recognition.

II. RELATED WORKS

In this section, we first summarize the definition of engagement. Next, previous studies on engagement recognition are described. Finally, several attempts to generate system behaviors according to user engagement are introduced.

A) Definition of engagement

The engagement was originally defined in a sociology study [24]. This concept has been extended and variously defined in the context of dialogue research [25]. We categorize the definitions into two types as follows. The first type focuses on cues to start and end the dialogue. Example definitions are "*the process by which two (or more) participants establish, maintain, and end their perceived connection*" [17] and "*the process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume, or terminate an interaction*" [26]. This type of engagement is related to *attention* and *involvement* [27]. The focus of studies based on these definitions was when and how the conversation starts and also ends. For example, one of the tasks was to detect the engaged person who wants to start the conversation with a situated robot [28, 29]. The second type of definition is about the quality of the connection between participants during the dialogue. For example, the engagement was defined as "*how much a participant is interested in and attentive to a conversation*" [30] and "*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*" [31]. This type of engagement is related to *interest* and *rapport*. The focus of studies based on these definitions was how the user state changes during the dialogue. Both types of definitions are important for dialogue systems to accomplish a high quality of dialogue. In this study, we focus on the latter type of engagement, so that the purpose of our recognition model is to understand the user state during the dialogue.

B) Engagement recognition

A considerable number of studies have been made on engagement recognition over the last decade. Engagement recognition has been generally formulated as a binary classification problem: engaged or not (disengaged), or a category classification problem like *no interest* or *following the conversation* or *managing the conversation* [32]. The useful features for engagement recognition have been investigated and identified from a multi-modal perspective. Non-linguistic behaviors are commonly used as the clue to recognize engagement because verbal information like linguistic features is specific to the dialogue domain and content, and speech recognition is error-prone. A series of studies on human–agent interaction found that user engagement was related to several features such as spatial information of the human (e.g. location, trajectory, distance to the robot) [18, 26, 33], eye-gaze behaviors (e.g. looking at the agent, mutual gaze) [18, 22, 34–36], facial information (e.g. facial movement, expression, head pose) [34, 36], conversational behaviors (e.g. voice activity, adjacency pair, backchannel, turn length) [18, 35, 37], laughing [38], and posture [39]. Engagement recognition modules based on the multi-modal features were implemented in agent systems and empirically tested with real users [36]. For human–human interaction, it was also revealed that the effective features in dyadic

conversations were acoustic information [30, 40, 41], facial information [40, 41], and low-level image features (e.g. local pattern, RGB data) [41]. Furthermore, the investigation was extended to multi-party conversations. They analyzed features including audio and visual backchannels [23], eye-gaze behaviors [23, 32], and upper body joints [42]. The recognition models using the features mentioned above were initially based on heuristic approaches [27, 33, 43]. Recent methods are based on machine learning techniques such as support vector machines (SVM) [18, 23, 36, 42], hidden Markov models [30], and convolutional neural networks [41]. In this study, we focus on behaviors when the user is listening to system speech, such as backchannels, laughing, head nodding, and eye-gaze.

We also find a problem of subjectivity in the annotation process of user engagement. Since the perception of engagement is subjective, it is difficult to define the annotation criteria objectively. Therefore, most of the previous studies conducted the annotation of engagement with multiple annotators. One approach is to train a few annotators to avoid disagreement among annotators [18, 32, 36, 41]. However, when we consider natural interaction, the annotation becomes more complicated and challenging to be consistent among annotators. In this case, another approach is based on 'wisdom of crowds' where many annotators are recruited. Eventually, the annotations were integrated using methods such as majority labels, averaged scores, and agreed labels [18, 22, 23]. In our proposed method, the different views of the annotators are taken into account. It is expected that we can understand the difference among the annotators. Our work is novel, in that the difference in annotation form the basis of our engagement recognition model.

C) Adaptive behavior generation according to engagement

Some attempts were made to generate system behaviors after recognizing user engagement. These works are essential to clarify the significance of engagement recognition. Although this is beyond this paper, our purpose of engagement recognition is similar to those of the studies.

Turn-taking behaviors are fine-grained and could be reflective of user engagement. An interactive robot was implemented to adjust its turn-taking behavior according to user engagement [18]. For example, if a user was engaged, the robot behaved to start a conversation with the user and give the floor to the user. As a result, subjective evaluations of both the effectiveness of communication and user experience were improved by this behavior strategy. Besides, in our preliminary study with a remote conversation, the analysis result implied that if a participant was engaged in the conversation, the duration of the participant's turn became longer than the case of not engaged, and the frequency of backchannels given by the counterpart was also higher [44].

Dialogue strategy can also be adapted to user engagement. Topic selection based on user engagement was proposed [45]. The system was designed to predict user engagement on each topic, and select the next topic which maximizes both user engagement and the system's preference. A chatbot system was implemented to select a dialogue module according to user engagement [19]. For example, when the user was not engaged in the conversation, the system switched the current dialogue module into another one. Consequently, subjective evaluations such as the appropriateness of the system utterance were improved. Another system was designed to react to user disengagement [36]. In an interview dialogue, when the user (interviewee) was disengaged, the system said positive feedbacks to elicit more self-disclosure from the user. Another research group investigated how to handle user disengagement in a human–robot interaction [20]. They compared two kinds of system feedback: explicit and implicit. The result of a subject experiment suggested that the implicit strategy of inserting fillers was preferred by the users than the explicit one where the system directly asks a question such as "*Are you listening?*".

III. DIALOGUE DATA AND ANNOTATION OF ENGAGEMENT

In this section, we describe the dialogue data used in this study. We conducted an annotation of user engagement with multiple annotators. The annotation result is analyzed to confirm inconsistencies among the annotators on the perception of engagement.

A) Human–robot interaction corpus

We have collected a human–robot interaction corpus in which the humanoid robot ERATO intelligent conversational android (ERICA) [46, 47] interacted with human subjects. ERICA was operated by another human subject, called an operator, who was in a remote room. The dialogue was one-on-one, and the subject and ERICA sat on chairs facing each other. Figure 2 shows a snapshot of the dialogue. The dialogue scenario was as follows. ERICA works in a laboratory as a secretary, and the subject visited the professor. Since the professor was absent for a while, the subject talked with ERICA until the professor would come back.

The voice uttered by the operator was directly played with a speaker placed on ERICA in real time. When the operator spoke, the lip and head motions of ERICA were automatically generated from the prosodic information [48, 49]. The operator also manually controlled the head and eye-gaze motions of ERICA to express some behaviors such as eye-contact and head nodding. We recorded the dialogue with directed microphones, a 16-channel microphone array, RGB cameras, and a Kinect v2 sensor. After the recording, we manually annotated the conversation data including utterances, turn units, dialogue acts, backchannels, laughing, fillers, head nodding, and eye gaze (the object at which the participant is looking).

We use 20 dialogue sessions for an annotation of subject engagement in this paper. The subjects were 12 females and 8 males, with ages ranging from teenagers to over 70 years

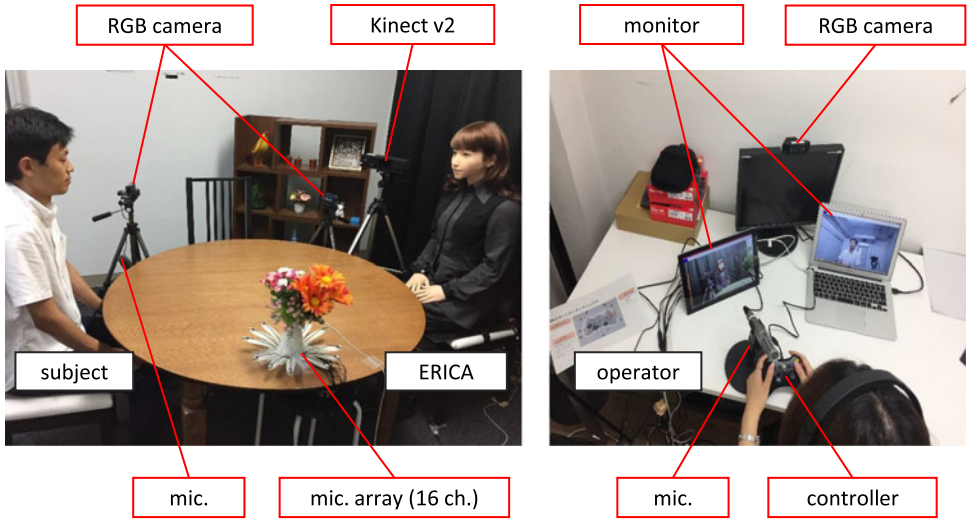

Fig. 2. Setup for dialogue collection.

old. The operators were six amateur actresses in their 20 and 30 s. Whereas each subject participated in only one session, each operator was assigned several sessions. Each dialogue lasted about 10 minutes. All participants were native Japanese speakers.

B) Annotation of engagement

There are several choices to annotate ground-truth labels of the subject engagement. The intuitive method is to ask the subject to evaluate his/her own engagement right after the dialogue session. However, in practice, we can obtain only one evaluation on the whole dialogue due to time constraints. It is difficult to obtain evaluations on fine-grained phenomena such as every conversational turn. Furthermore, we sometimes observe a bias where subjects tend to give positive evaluations of themselves. This kind of bias was observed in other works [50, 51]. Another method is to ask the ERICA's operators to evaluate the subject engagement. However, it was difficult to let the actresses participate in this annotation work due to time constraints. Similar to the first method, we would obtain only one evaluation on the whole dialogue, but this is not useful for the current recognition task. This problem often happens in other studies for building corpora because the dialogue recording and the annotation work are done separately. Most of the previous studies adopted a practical approach where they asked third-party people (annotators) to evaluate engagement. This approach is categorized into two types: training a small number of annotators [18, 32, 36, 41] and making use of the wisdom of crowds [22, 23]. The former type is valid when the annotation criterion is objective. On the other hand, the latter type is better when the criterion is subjective and when a large amount of data is needed. We took the latter approach in this study.

We recruited 12 females who had not participated in the dialogue data collection. Their gender was set to be same as those of the ERICA's operators. We instructed the annotators to take the point of view of the operator. Each dialogue session was randomly assigned to five annotators. The definition of engagement was presented as "*How much the subject is interested in and willing to continue the current dialogue with ERICA*". We asked the annotators to annotate the subject engagement based on its behaviors while the subject was listening to ERICA's talk. Therefore, the subject engagement can be interpreted as *listener engagement*. It also means that the annotators observe *listener behaviors* expressed by the subject. We showed a list of listener behaviors that could be related to engagement, with example descriptions. This list included facial expression, laughing, eye gaze, backchannels, head nodding, body pose, moving of shoulders, and moving of arms or hands. We instructed the annotators to watch the dialogue video by standing in the viewpoint of the ERICA's operator, and to press a button when all the following conditions were being met: (1) ERICA was holding the turn, (2) the subject was expressing any listener behaviors, and (3) the behavior means the high level of subject engagement. For condition (1), the annotators were notified of auxiliary information which showed the timing of when the conversational turn was changed between the subject and ERICA.

C) Analysis of annotation result

Across all annotators and sessions, the average number of button presses per session was 18.13 with a standard deviation of 12.88. Since each annotator was assigned some of the 20 sessions randomly, we tested one-way ANOVA for both inter-annotator and inter-session, respectively. As a result, we could see significant differences of the average numbers of button presses among both the annotators ($F(11, 88) = 4.64$, $p = 1.51 \times 10^{-5}$) and the sessions ($F(19, 80) = 2.56$, $p = 1.92 \times 10^{-3}$). There was a variation among not only sessions but also annotators.

In this study, we use ERICA's conversational turn as a unit for engagement recognition. The conversational turn is useful for spoken dialogue systems to utilize the result of engagement recognition because the systems typically make

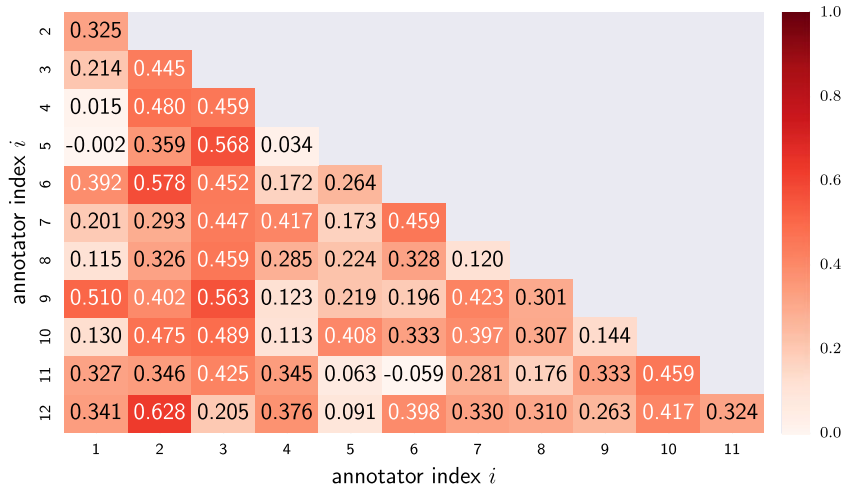

Fig. 3. Inter-annotator agreement score (Cohen's kappa) on each pair of the annotators.

a decision on their behaviors on a turn basis. If an annotator pressed the button more than once in a turn, we regarded that the turn was annotated as *engaged* by the annotator. We excluded short turns whose durations are smaller than 3 seconds, and also some turns corresponding to the greeting. As a result, the total number of ERICA's turns was 433 over 20 dialogue sessions. The numbers of engaged and not engaged turns from all annotators were 894 and 1271, respectively.

We investigated the agreement level among the annotators. The average value of Cohen's kappa coefficients on every pair of two annotators was 0.291 with a standard deviation of 0.229. However, as Fig. 3 shows, some pairs showed coefficients higher than the moderate agreement (larger than 0.400). The result suggests that the annotators can be clustered into some groups based on their tendencies to annotate engagement similarly. Each group regarded different behaviors as important and had different thresholds to accept the behaviors as engaged events.

We took a survey on which behaviors the annotators regarded as essential to judge the subject engagement. For every session, we asked the annotators to select all the essential behaviors to judge the subject engagement. Table 1 lists the results of this survey. Note that we conducted this survey in total 100 times (five annotators in 20 sessions). The result indicates that engagement could be related to some listener behaviors such as facial expression, backchannels, head nodding, eye gaze, laughing, and body pose.

In the following experiments, we use four behaviors: backchannels, laughing, head nodding, and eye gaze. As we have seen in the section on the related works, these behaviors have been identified as indicators of engagement. We manually annotated the occurrences of the four behaviors. The definition of backchannel in this annotation was responsive interjections (such as '*yeah*' in English and '*un*' in Japanese), and expressive interjections (such as '*oh*' in English and '*he-*' in Japanese) [52]. The laughing was defined as vocal laughing, not including just smiling without any vocal utterance. We annotated the occurrence of head nodding based on the vertical movement of the head. The occurrence of eye-gaze behaviors is acknowledged when the subject was gazing at ERICA's face continuously for more than 10 seconds. We decided this 10-second threshold by confirming a reasonable balance between the accuracy in Table 2 and the recall of the engaged turns. It was challenging to annotate facial expression and body pose due to their ambiguity. We will consider the other behaviors which we do not use in this study as additional features in the future work. The relationship between

Table 1. The number of times selected by the annotators as meaningful behaviors

Listener behavior	#selected
Facial expression	77
Backchannels	67
Head nodding	65
Eye gaze	40
Laughing	39
Body pose	32
Moving of shoulders	3
Moving of arms or hands	2
Others	4

Table 2. Relationship between the occurrence of each behavior and the annotated engagement (1: occurred / engaged, 0: not occurred / not engaged)

		Engagement		
Behavior		1	0	Accuracy
Backchannel	1	790	830	0.569
	0	104	441	
Laughing	1	139	46	0.630
	0	755	1225	
Head nodding	1	653	532	0.643
	0	241	739	
Eye gaze	1	332	243	0.628
	0	562	1028	

Table 3. Accuracy scores of each annotator's engagement labels when the reference labels of each behavior are used

Behavior	Annotator index											
	1	2	3	4	5	6	7	8	9	10	11	12
Backchannel	0.751	0.588	0.584	0.487	0.301	0.553	0.647	0.441	0.749	0.576	0.542	0.553
Laughing	0.348	0.690	0.753	0.702	0.850	0.648	0.479	0.752	0.518	0.589	0.651	0.619
Head nodding	0.736	0.722	0.695	0.607	0.566	0.620	0.647	0.466	0.769	0.576	0.657	0.604
Eye gaze	0.488	0.717	0.689	0.696	0.659	0.581	0.599	0.596	0.559	0.677	0.627	0.660

the occurrences of the four behaviors and the annotated engagement is summarized in Table 2. Note that we used the engagement labels given by all individual annotators. The result suggests that these four behaviors are useful cues to recognize the subject engagement. We further analyzed the accuracy scores on each annotator in Table 3. The results show that each annotator has a different perspective on each behavior. For example, the engagement labels of the first annotator (index 1) are related to the labels of backchannels and head nodding. On the other hand, those of the second annotator (index 2) are related to those of laughing, head nodding, and eye gaze. This difference implies that we need to consider each annotator's different perspective on engagement.

IV. LATENT CHARACTER MODEL

In this section, we propose a hierarchical Bayesian model for engagement recognition. As we have shown, the annotators can be clustered into some groups based on their perception manners. We assume that each annotator has a character which affects his/her perception of engagement. The character is a latent variable estimated from the annotation data. We call the proposed model as a latent character model. This model is inspired by latent Dirichlet allocation [53] and the latent class model which estimates annotators' abilities for a decision task like diagnosis [54].

A) Problem formulation

At first, we define the problem formulation of this engagement recognition as follows. Engagement recognition is done for each turn of the robot (ERICA). The input is based on listener behaviors of the user during the turn: laughing, backchannels, head nodding, and eye gaze. Each behavior is represented as binary: occur or not as defined in the previous section. The input feature is a combination of the occurrences of the four behaviors and referred to as *behavior pattern*. In this study, since we use the four behaviors, the possible number of the behavior patterns is 16 ($= 2^4$). Although the number of the behavior patterns would be massive if we use many behaviors, the observed patterns are limited so that we can exclude the less-frequent patterns. The output is also binary: engaged or not, as annotated in the previous section. Note that this ground-truth label differs for each annotator.

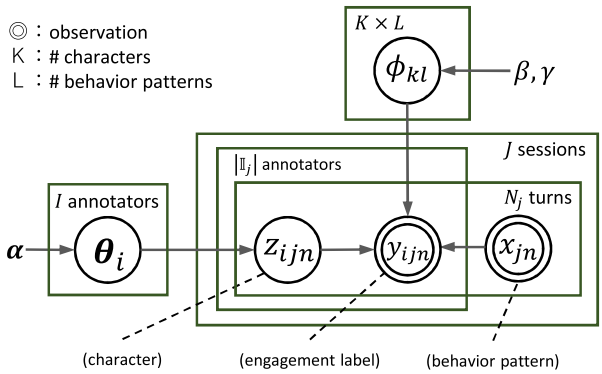

Fig. 4. Graphical model of latent character model.

B) Generative process

The latent character model is illustrated as the graphical model in Fig. 4. The generative process is as follows. For each annotator, parameters of a character distribution are generated from the Dirichlet distribution as

$$\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ik}, \ldots, \theta_{iK}) \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad 1 \leq i \leq I, \qquad (1)$$

where I, K, i, k denote the number of annotators, the number of characters, the annotator index, and the character index, respectively, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$ is a hyperparameter. The parameter θ_{ik} represents the probability that the i-th annotator has the k-th character. For each combination of the k-th character and the l-th behavior pattern, a parameter of an engagement distribution is generated by the beta distribution as

$$\phi_{kl} \sim \text{Beta}(\beta, \gamma), \quad 1 \leq k \leq K, \quad 1 \leq l \leq L, \qquad (2)$$

where L denotes the number of behavior patterns, and β and γ are hyperparameters. The parameter ϕ_{kl} represents the probability that annotators with the k-th character interpret the l-th behavior pattern as an *engaged* signal.

The number of the total dialogue sessions is represented as J. For the j-th session, the set of annotators who were assigned to this session is represented as $\mathbb{I}_j$. The number of conversational turns of the robot in the j-th session is represented as N_j. For each turn, a character is generated from the categorical distribution corresponding to the i-th annotator as

$$z_{ijn} \sim \text{Categorical}(\boldsymbol{\theta}_i), \quad 1 \leq j \leq J, \quad 1 \leq n \leq N_j, \quad i \in \mathbb{I}_j, \qquad (3)$$

where n denotes the turn index. The input behavior pattern in this turn is represented as $x_{jn} \in \{1, \ldots, L\}$. Note that the behavior pattern is independent from the annotator index i. The binary engagement label is generated from the Bernoulli distribution based on the character and the input behavior pattern as

$$y_{ijn} \sim \text{Bernoulli}(\phi_{z_{ijn}x_{jn}}). \tag{4}$$

Given the dataset of the above variables and parameters, the conditional distribution is represented as

$$\begin{aligned} &p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\boldsymbol{X}) \\ &\quad = p(\boldsymbol{Z}|\boldsymbol{\Theta})\ p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Phi})\ p(\boldsymbol{\Theta})\ p(\boldsymbol{\Phi}), \end{aligned} \tag{5}$$

where the bold capital letters represent the datasets of the variables written by the corresponding small letters. Note that $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are the model parameters, and the dataset of the behavior patterns $\boldsymbol{X}$ is given and regarded as constant.

C) Training

In the training phase, the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are estimated. The training datasets of the behavior patterns $\boldsymbol{X}$ and the engagement labels $\boldsymbol{Y}$ are given. We use collapsed Gibbs sampling which marginalizes the model parameters and efficiently samples only target variables. Here, we sample each character alternately and iteratively from its conditional probability distribution as

$$z_{ijn} \sim p(z_{ijn}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}_{\backslash ijn}, \boldsymbol{\alpha}, \beta, \gamma), \tag{6}$$

where the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are marginalized. Note that $\boldsymbol{Z}_{\backslash ijn}$ is the set of the characters without z_{ijn}. The conditional probability distribution is expanded in the same manner as the other work [55]. The distribution is proportionate to the product of two terms as

$$\begin{aligned} &p(z_{ijn} = k|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}_{\backslash ijn}, \boldsymbol{\alpha}, \beta, \gamma) \\ &\quad \propto p(z_{ijn} = k|\boldsymbol{Z}_{\backslash ijn}, \boldsymbol{\alpha}) \\ &\qquad \times p(y_{ijn}|\boldsymbol{X}, \boldsymbol{Y}_{\backslash ijn}, z_{ijn} = k, \boldsymbol{Z}_{\backslash ijn}, \beta, \gamma), \end{aligned} \tag{7}$$

where $\boldsymbol{Y}_{\backslash ijn}$ is the dataset of the engagement labels without y_{ijn}. The first term is calculated as

$$p(z_{ijn} = k|\boldsymbol{Z}_{\backslash ijn}\boldsymbol{\alpha}) = \frac{p(z_{ijn} = k, \boldsymbol{Z}_{\backslash ijn}|\boldsymbol{\alpha})}{p(\boldsymbol{Z}_{\backslash ijn}|\boldsymbol{\alpha})} \tag{8}$$

$$= \frac{D_{ik\backslash ijn} + \alpha_k}{D_i - 1 + \sum_{k'=1}^{K} \alpha_{k'}}. \tag{9}$$

Note that D_i and $D_{ik\backslash ijn}$ represent the number of turns where the i-th annotator was assigned, and the number of turns where the i-th annotator had the k-th character without considering z_{ijn}, respectively. The above expansion from equations (8) to (9) is explained in the appendix. The second term is calculated as

$$\begin{aligned} &p(y_{ijn}|\boldsymbol{X}, \boldsymbol{Y}_{\backslash ijn}, z_{ijn} = k, \boldsymbol{Z}_{\backslash ijn}, \beta, \gamma) \\ &\quad = \frac{p(\boldsymbol{Y}|\boldsymbol{X}, z_{ijn} = k, \boldsymbol{Z}_{\backslash ijn}, \boldsymbol{X}, \beta, \gamma)}{p(\boldsymbol{Y}_{\backslash ijn}|\boldsymbol{X}, \boldsymbol{Z}_{\backslash ijn}, \boldsymbol{X}, \beta, \gamma)} \end{aligned} \tag{10}$$

$$\begin{aligned} &= \prod_{l=1}^{L} \left\{ \frac{\Gamma(N_{kl\backslash ijn} + \beta + \gamma)}{\Gamma(N_{kl\backslash ijn} + N_{ijnl} + \beta + \gamma)} \right. \\ &\qquad \times \frac{\Gamma(N_{kl1\backslash ijn} + N_{ijnl1} + \beta)}{\Gamma(N_{kl1\backslash ijn} + \beta)} \\ &\qquad \left. \times \frac{\Gamma(N_{kl0\backslash ijn} + N_{ijnl0} + \gamma)}{\Gamma(N_{kl0\backslash ijn} + \gamma)} \right\}, \end{aligned} \tag{11}$$

where $\Gamma(\cdot)$ is the gamma function. Note that $N_{kl\backslash ijn}$ represents the number of times when the l-th behavior pattern was observed by annotators with the k-th character without considering x_{ijn}. Among them, $N_{kl1\backslash ijn}$ and $N_{kl0\backslash ijn}$ are the number of times when the annotators gave the engaged and not engaged labels, respectively. Besides, N_{ijnl} represents the binary variable indicating if the i-th annotator observed the l-th behavior pattern in the n-th turn of the j-th session. Among them, N_{ijnl1} and N_{ijnl0} are binary variables indicating if the annotator gave the engaged and not engaged labels, respectively. The above expansion from equations (10) to (11) is also explained in the appendix.

After sampling, we select one of the sampling results where the joint probability of the variables is maximized as

$$\boldsymbol{Z}^* = \arg\max_{\boldsymbol{Z}^{(r)}}\ p(\boldsymbol{Y}, \boldsymbol{Z}^{(r)}|\boldsymbol{X}, \boldsymbol{\alpha}, \beta, \gamma), \tag{12}$$

where $\boldsymbol{Z}^{(r)}$ represents the r-th sampling result. The joint probability is expanded as

$$\begin{aligned} &p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\alpha}, \beta, \gamma) \\ &\quad = p(\boldsymbol{Z}|\boldsymbol{\alpha})p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \beta, \gamma), \end{aligned} \tag{13}$$

$$\begin{aligned} &\propto \prod_{i=1}^{I} \frac{\prod_k \Gamma(D_{ik} + \alpha_k)}{\Gamma(D_i + \sum_k \alpha_k)} \\ &\qquad \times \prod_{k=1}^{K} \prod_{l=1}^{L} \frac{\Gamma(N_{kl1} + \beta)\Gamma(N_{kl0} + \gamma)}{\Gamma(N_{kl} + \beta + \gamma)}. \end{aligned} \tag{14}$$

Note that N_{kl} is the number of times when annotators with k-th character annotated the l-th behavior pattern. Among them, N_{kl1} and N_{kl0} represent the number of times when the annotators gave the engaged and not-engaged labels, respectively. Besides, D_{ik} is the number of turns where the i-th annotator had the k-th character. The above expansion from equations (13) to (14) is also explained in the appendix. Finally, the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are estimated based on the sampling result $\boldsymbol{Z}^*$ as

$$\theta_{ik} = \frac{D_{ik} + \alpha_k}{D_i + \sum_{k'=1}^{K} \alpha_{k'}}, \tag{15}$$

$$\phi_{kl} = \frac{N_{kl1} + \beta}{N_{kl} + \beta + \gamma}, \tag{16}$$

where D_i, D_{ik}, N_{kl}, and N_{kl1} are counted up among the sampling result $\boldsymbol{Z}^*$.

D) Testing

In the testing phase, the unseen engagement label given by a target annotator is predicted based on the estimated model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. The input data are the behavior pattern $x_t \in \{1, \ldots, L\}$ and the target annotator index $i \in \{1, \ldots, I\}$. Note that t represents the turn index in the test data. The probability that the target annotator gives the engaged label on this turn is calculated by marginalizing the character as

$$p(y_{it} = 1 | x_t, i, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{k=1}^{K} \theta_{ik}\, \phi_{kx_t}. \tag{17}$$

The t-th turn is recognized as *engaged* when this probability is higher than a threshold.

E) Related models

We summarize similar models considering the difference of annotators. In a task of backchannel prediction, a two-step conditional random fields was proposed [56, 57]. They trained a prediction model per annotator. The final decision is based on voting by the individual models. Our method trains the model based on the character, not for each annotator, so that robust estimation is expected even if the amount of data for each annotator is not large, which is the case in many realistic applications. In a task of estimation of empathetic states among dialogue participants, a model classified annotators by considering both the annotators' tendencies of the estimation and their personalities [58]. The personalities correspond to the characters in our model. Their model was able to estimate the empathetic state based on a specific personality. It assumed that the personality and the input features such as the behavior patterns are independent. In our model, we assume that the character and the input features are dependent, meaning that how to perceive each behavior pattern is different for each character.

V. ONLINE PROCESSING

In order to use the engagement recognition model in spoken dialogue systems, we have to detect the behavior patterns automatically. In this section, we first explain automatic detection methods of the behaviors. Then, we integrate the detection methods with the engagement recognition model.

A) Automatic detection of behaviors

We detect backchannels and laughing from speech signals recorded by a directed microphone. Note that we will investigate making use of a microphone array in future work. This task has been widely studied in the context of social signal detection [59]. We proposed using bi-directional long short-term memory with connectionist temporal classification (BLSTM-CTC) [60]. The advantage of CTC is that we do not need to annotate the time-wise alignment of the social signal events. On each user utterance, we extracted the log-Mel filterbank features as a 40-dimension vector and also a delta and a delta-delta of them. The number of dimensions of the input features was 120. We trained the BLSTM-CTC model for backchannels and laughing independently. The number of hidden layers was 5 and the number of units on each layer was 256. For training, we used other 71 dialogue sessions recorded in the same manner as the dataset for engagement recognition. In the training set, the total number of user utterances was 14,704. Among them, the number of utterances containing backchannels was 3931, and the number of utterances containing laughing was 1003. Then, we tested the 20 sessions which were used for engagement recognition. In the test dataset, the total number of the subject utterances was 3517. Among them, the utterances containing backchannels was 1045, and the utterances containing laughing was 240. Precision and recall of backchannels were 0.780 and 0.865, and the F1 score was 0.820. For laughing, precision and recall were 0.772 and 0.496, and the F1 score was 0.604. The occurrence probabilities of backchannel and laughing are computed for every user utterance. Therefore, we take the maximum value during the turn as the input for engagement recognition.

We detect head nodding from visual information captured by the Kinect v2 sensor. Detection of head nodding has also been widely studied in the field of computer vision [61, 62]. We used LSTM for this task [63]. With the Kinect v2 sensor, we can measure the head direction in the 3D space. We calculated a feature set containing the instantaneous speeds of the yaw, roll, and pitch of the head. Other features were the average speed, average velocity, acceleration and range of the head pitch over the previous 500 milliseconds. We trained an LSTM model with these features whose number of dimension was 7. We used a single hidden layer with 16 units. The dataset was the same 20 sessions as the engagement recognition task, and 10-fold cross-validation was applied. The number of data frames per second was about 30, and we made a prediction every 5 frames. There were 29 560 prediction points in the whole dataset, and 3152 of them were manually annotated as points where the user was nodding. Note that we discarded the data frames on the subject's turn. For prediction-point-wise evaluation, precision and recall of head nodding frames were 0.566 and 0.589, and the F1 score was 0.577. For event-wise detection, we regarded a continuous sequence of detected nodding as a head nodding event where the duration is longer than 300 milliseconds. If the sequence overlapped with a ground-truth event, the event was correctly detected. On an event basis, there were 855 head nodding events. Precision and recall of head nodding events were 0.608 and 0.763, and the F1 score was 0.677. We compared the LSTM performance with several other models such as SVM and DNN and found that the LSTM model had the best score [63]. The occurrence probability of head nodding is estimated at every frame. We smoothed the output sequence

with a Gaussian filter where the standard deviation for Gaussian kernel was 3.0. Therefore, we also take the maximum value during the turn as the input for engagement recognition.

The eye-gaze direction is approximated by the head orientation given by the Kinect v2 sensor. We would be able to detect eye-gaze direction precisely if we used an eye tracker, but non-contact sensors such as the Kinect v2 sensor are preferable for spoken dialogue systems interacting with users on a daily basis. Eye gaze towards the robot is detected when the distance between the head-orientation vector and the location of the robot's head is smaller than a threshold. We set the threshold at 300 mm in our experiment. The number of eye-gaze samples per second was about 30. In the dataset of the 20 sessions, there were 300 426 eye-gaze samples in total, and 77 576 samples were manually annotated as *looking at the robot*. For frame-wise evaluation, precision and recall of the eye-gaze towards the robot were 0.190 and 0.580, and the F1 score was 0.286. This result implies that the ground-truth label of eye gaze based on the actual eye-gaze direction is sometimes different from the head direction. However, even if the frame-wise performance is low, it is enough if we can detect the eye-gaze behavior which is continuous gaze longer than 10 seconds. We also evaluated the detection performance on a turn basis. There were 433 turns in the corpus, and the continuous eye-gaze behavior was observed in 115 turns of them. For this turn-wise evaluation, precision and recall of the eye-gaze behavior were 0.723 and 0.704, and the F1 score was 0.713. This result means that this method is sufficient to detect the eye-gaze behavior for engagement recognition. Note that we ignored some *not looking* states if the duration is smaller than 500 milliseconds. To convert the estimated binary states (looking or not) to the occurrence probability of the eye-gaze behavior, we use a logistic function with a threshold of 10 seconds.

B) Integration with engagement recognition

We use the behavior detection models in the test phase of engagement recognition. At first, the occurrence probability of the l-th behavior pattern in the t-th turn of the test dataset is calculated as

$$p_t(l) = \prod_{m=1}^{M} p_t(m)^{b_m} (1 - p_t(m))^{(1-b_m)}, \qquad (18)$$

where M, m, $p_t(m)$, and $b_m \in \{0, 1\}$ denote the number of behaviors, the behavior index, the output probability of the behavior m, and the occurrence of the behavior m, respectively. Note that $p_t(m)$ corresponds to the output of each behavior detection model, and the binary value b_m is based on the given behavior pattern l. For example, when the given behavior pattern l represents the case where both laughter ($m = 2$) and eye-gaze ($m = 4$) occur, the binary values are represented as $(b_1, b_2, b_3, b_4) = (0, 1, 0, 1)$. As the behavior pattern l is represented by the combination of the occurrences, $l = \sum_{m=1}^{M} b_m \cdot 2^{m-1}$. The probability of engaging (equation (17)) is reformulated by marginalizing not only the character but also the behavior pattern with its occurrence probability as

$$p(y_{it} = 1 | P_{tl}, i, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \theta_{ik}\, \phi_{kl}\; p_t(l), \qquad (19)$$

where P_{tl} denotes the set of the occurrence probabilities of all possible behavior patterns calculated by equation (18).

VI. EXPERIMENTAL EVALUATIONS

In this section, the latent character model is compared with other models that do not consider the difference of the annotators. Besides, we evaluate the accuracy of the online implementation. Furthermore, we investigate the effectiveness of each behavior to identify important behaviors in this recognition task. In this experiment, the task is to recognize each annotator's labels. Since we observed a low agreement among the annotators in Section III C, it does not make sense to recognize a single overall label such as majority voting. In real-life applications, we can select an annotator or a character appropriate for the target system. At the end of this section, we suggest a method to select a character distribution for engagement recognition based on a personality trait expected for a system such as a humanoid robot.

A) Experimental setup

We conducted the cross-validation with the 20 dialogue sessions: 19 for training and the rest for testing. In the proposed model, the number of sampling was 10 000, and all prior distributions were the uniform distribution. The evaluation was done for each annotator one by one where five annotators individually annotated each dialogue session. Given the annotator index i, the probability of the engaged label (equations (17) or (19)) was calculated for each turn. Setting the threshold at 0.5, we obtained the accuracy score which is a ratio of the number of the correctly recognized turns to the total number of turns. The final evaluation was made by averaging the accuracy scores for all annotators and also the cross-validation. The chance level was 0.579 (=1,271 / 2,165).

B) Effectiveness of character

At first, we compared the proposed model with two methods to see the effectiveness of the character. In this experiment, we used the input behavior patterns which were manually annotated. For the proposed model, we explored an appropriate number of characters (K) by changing from 2 to 5 on a trial basis. The first compared model was the same as the proposed model other than a unique character ($K = 1$). The second compared models were based on other machine learning methods. We used logistic regression, SVM, and a multilayer perceptron (neural network). For each model, two types of training are considered: *majority* and *individual*. In the *majority* type, we integrated the training labels of the five annotators by majority voting and trained a unique model which was independent of the

annotators. In the *individual* type, we trained an individual model for each annotator with his/her data only and used each model according to the input annotator index i in the test phase. Although the *individual* type can learn the different tendency of each annotator, the amount of training data decreases. Furthermore, we divided the training data into training and validation datasets on a session basis so that it corresponds to 9:1. We trained each model with the training dataset, and then tuned the parameter of each model with the validation dataset. For logistic regression, we tuned the weight parameter of the $l2$-norm regularization. For SVM, we used the radial basis function kernel and tuned the penalty parameter of the error term. For the multilayer perceptron, we tuned the weight parameter of the $l2$-norm regularization for each unit. We also needed to decide on other settings for the multilayer perceptron such as the number of hidden layers, the number of hidden units, the activation function, the optimization method, and the batch size. We tried many settings and report the best result among them.

Table 4 summarizes the accuracy scores. Among the conventional machine learning methods, the multilayer perceptron showed the highest score on both of the *majority* and *individual* types. For the *majority* type, the best setting of the multilayer perceptron was 5 hidden layers and 16 hidden units. For the *individual* type, the best setting was 1 hidden layer and 16 hidden units. The difference in the number of hidden layers can be explained by the number of available training data.

Considering the character ($K \geq 2$), we improved the accuracy compared with the w/o-character models including the multilayer perceptron. The highest accuracy was 0.711 with the four characters ($K = 4$). We conducted a paired t-test between the cases of the unique character ($K = 1$) and the four characters ($K = 4$) and found a significant difference between them ($t(99) = 2.55$, $p = 1.24 \times 10^{-2}$). We also performed a paired t-test between the proposed model with the four characters and the multilayer perceptron models. There was a significant difference between the proposed model ($K = 4$) and the *majority* type of the multilayer perceptron ($t(99) = 2.34$, $p = 2.15 \times 10^{-2}$). We also found a significant difference between the proposed model ($K = 4$) and the *individual* type of the multilayer perceptron ($t(99) = 2.55$, $p = 1.24 \times 10^{-2}$).

These results indicate that the proposed model simulates each annotator's perception of engagement more accurately than the others by considering the character. Apparently, the *majority* voting is not enough for this kind of recognition task that contains subjectivity. Although the *individual* model has the potential to simulate each annotator's perception, it fails to address the data sparseness problem in model training. This means that there was not enough training data for each annotator. We often face this problem when we use data that was collected by the wisdom of crowd approach where a large number of annotators are available but the amount of data of each annotator is small.

C) Evaluation with automatic behavior detection

We evaluated the online processing described in Section V. We compared two types of the input features in the test phase: manually annotated and automatically detected. Note that we used the manually annotated features for training in both cases. We also tested the number of characters (K) at only one and four. Table 5 shows the difference between the manual and automatic features. The accuracy is not degraded so much even when we use the automatic detection. We performed the paired t-test on the proposed model with the four characters ($K = 4$), and there was no significant difference between the cases of the manual and automatic features ($t(99) = 1.45$, $p = 1.51 \times 10^{-1}$). This result indicates that we can apply our proposed model to live spoken dialogue systems. Note that all detection models can run in real time with short processing time which does not affect the decision-making process in spoken dialogue systems.

D) Identifying important behaviors

We examined the effectiveness of each behavior by eliminating one of the four behaviors from the feature set. We again tested the number of characters (K) at only one and four. Table 6 reports the results on both cases of the manual and automatic features. From this table, laughing and eye-gaze

Table 4. Engagement recognition accuracy (K is the number of characters.)

Method		Accuracy
Chance level		0.579
Logistic regression	*Majority*	0.670
	Individual	0.681
SVM	*Majority*	0.667
	Individual	0.683
Multilayer perceptron	*Majority*	0.678
	Individual	0.690
Latent character (proposed)	$K = 1$ (no character)	0.674
	$K = 2$	0.697
	$K = 3$	0.703
	$K = 4$	0.711
	$K = 5$	0.688

Table 5. Engagement recognition accuracy of the online processing

Method		Behavior	
		Manual	Automatic
Logistic regression	*Majority*	0.670	0.663
	Individual	0.681	0.678
SVM	*Majority*	0.667	0.668
	Individual	0.683	0.673
Multilayer Perceptron	*Majority*	0.678	0.653
	Individual	0.690	0.681
Latent character (proposed)	$K = 1$	0.674	0.663
	$K = 4$	0.711	0.700

Table 6. Recognition accuracy without each behavior of the proposed method

	Manual		Automatic	
Used behavior	$K = 1$	$K = 4$	$K = 1$	$K = 4$
All	0.674	0.711	0.663	0.700
w/o backchannels	0.669	0.699	0.657	0.684
w/o laughing	0.606	0.684	0.654	0.689
w/o head nodding	0.664	0.700	0.658	0.699
w/o eye gaze	0.650	0.681	0.602	0.669

behaviors are more useful for this engagement recognition task. This result is partly consistent with the analysis of Table 2. It is assumed that backchannels and head nodding were indicating engagement, but those can be used more frequently than the others. While backchannel and head nodding behaviors play a role to acknowledge that the turn-taking floor will be held by the current speaker, laughing and eye-gaze behaviors express the reaction towards the spoken content. Therefore, laughing and eye-gaze behaviors are more related to the high level of engagement. However, it is thought that some backchannels such as *expressive interjections* (such as '*oh*' in English and '*he-*' in Japanese) [52] are used to express the high level of engagement. From this perspective, there is room for further investigation to classify each behavior into some categories which are correlated with the level of engagement.

E) Example of parameter training

We analyzed the result of parameter training. The following parameters were trained using all 20 sessions. In this example, the number of characters was four ($K = 4$).

The parameters of the character distribution (θ_{ik}) is shown in Fig. 5. The vertical axis represents the probability that each annotator has each character. It is observed that some annotators have common patterns. We clustered the annotators based on this distribution by using the hierarchical clustering with the unweighted pair-group method with arithmetic mean algorithm. From the generated tree diagram, we extracted three clusters that are reported in Table 7. Four annotators were independent (the annotators 5, 7, 8, 12). The table also shows the averaged agreement scores of the engagement labels among the annotators inside

Table 7. Clustered annotators based on character distribution and averaged in-cluster agreement scores

Cluster	Annotator index	Cohen's kappa
A	1, 9	0.510
B	2, 3, 4, 6	0.431
C	10, 11	0.459

Table 8. Averaged agreement scores between-clusters

Cluster pair	Cohen's kappa
A - B	0.321
A - C	0.239
B - C	0.137

the same cluster. All scores are over the moderate agreement (larger than 0.400) and also higher than the whole averaged score (0.291) reported in Section III C. We further analyzed the averaged agreement scores between the clusters. Table 8 reports the scores that are lower than the in-cluster agreement scores.

The parameters of the engagement distribution (ϕ_{kl}) is shown in Fig. 6. The vertical axis represents the probability that each behavior pattern is recognized as engaged by each character. Note that we excluded some behavior patterns which appear less than five times in the corpus. We also show the number of times when each behavior pattern is observed. The proposed model can obtain the different distribution for each character. Although the first character ($k = 1$) seems to be reactive to some behavior patterns, it is also reactive to behaviors other than the four behaviors because the high probability is estimated against the empty behavior pattern (nothing) where no behavior was observed. The second and third characters ($k = 2, 3$) show a similar tendency, but some are different (e.g. BL, BG, and BNG). The fourth character ($k = 4$) is reactive to all behavior patterns except the patterns of empty (nothing) and eye gaze only (G). Among all characters, the co-occurrence of multiple behaviors leads to higher probability (the right side of the figure). Especially, when all behaviors are observed (BLNG), the probability becomes very high in all characters. This tendency indicates that the co-occurrence of multiple behaviors expresses the high level of engagement.

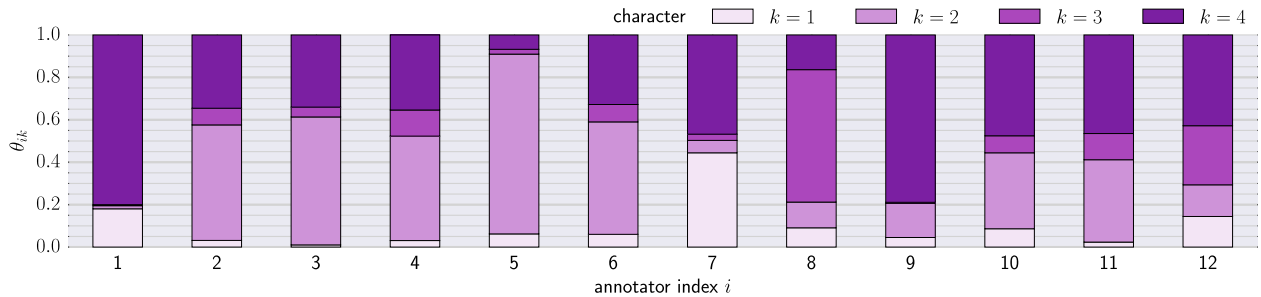

Fig. 5. Estimated parameter values of character distribution (Each value corresponds to the probability that each annotator has each character.).

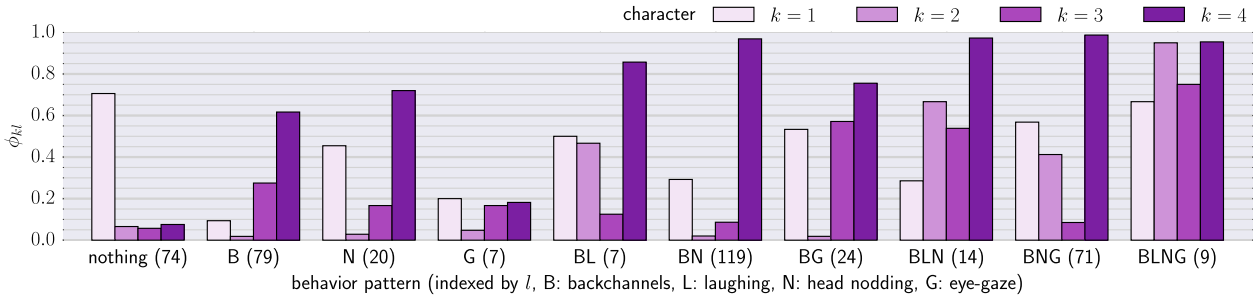

Fig. 6. Estimated parameter values of engagement distribution (Each value corresponds to the probability that each behavior pattern is recognized as engaged by each character. The number in parentheses next to the behavior pattern is the frequency of the behavior pattern in the corpus.).

F) How to determine character distribution

The advantage of the proposed model is to simulate various kinds of perspectives for engagement recognition by changing the character distribution. However, when we use the proposed model in a spoken dialogue system, we need to determine one character distribution to be simulated. We suggest a method based on the personality given to the dialogue system. Specifically, once we set the personality for the dialogue system, the character distribution for engagement recognition is decided. Here, as a proxy of the personality, we use the Big Five traits: *extroversion*, *neuroticism*, *openness to experience*, *conscientiousness*, and *agreeableness* [64]. For example, given a social role of a laboratory guide, the dialogue system is expected to be extroverted.

In the annotation work, we also measured the Big Five scores of each annotator [65]. We then trained a softmax single-layer linear regression model which maps the Big Five scores to the character distribution as shown in Fig. 5. Note that the weight parameters of the regression are constrained by the l_1-norm regularization and to be non-negative. The bias term was not added. Table 9 shows the regression weights and indicates that some characters are related to some personality traits. For example, *extroversion* is related to the first and fourth characters, followed by the second character.

We also tested the regression with some social roles. When a dialogue system plays a role of a laboratory guide, we set the input score on *extroversion* at the maximum value among the annotators, and the other scores are set to the average values. The output of the regression was $\boldsymbol{\theta} = (0.147, 0.226, 0.038, 0.589)$, which means the fourth character is weighted in this social role. For another social role such as a counselor, we set the input scores on *conscientiousness* and *agreeableness* at the maximum values among the annotators, and the other scores are set to the average values. The output was $\boldsymbol{\theta} = (0.068, 0.464, 0.109, 0.359)$, which means the second and fourth characters are weighted in this social role Further investigation is required on the effectiveness of this personality control.

Table 9. Regression weights for mapping from Big Five scores to character distribution

	Character index (k)			
Big Five factor	1	2	3	4
Extroversion	4.11	1.95	0.00	4.00
Neuroticism	0.71	0.00	0.00	1.94
Openness to experience	1.52	0.00	5.10	0.00
Conscientiousness	0.00	3.00	1.74	0.00
Agreeableness	0.00	3.25	0.00	2.25

VII. CONCLUSION

We have addressed engagement recognition using listener behaviors. Since the perception of engagement is subjective, the ground-truth labels depend on each annotator. We assumed that each annotator has a character that affects his/her perception of engagement. The proposed model estimates not only user engagement but also the character of each annotator as latent variables. The model can simulate each annotator's perception by considering the character. To use the proposed model in spoken dialogue systems, we integrated the engagement recognition model with the automatic detection of the listener behaviors. In the experiment, the proposed model outperforms the other methods that use either the majority voting for label generation in training or individual training for each annotator. Then, we evaluated the online processing with the automatic detection of the listener behaviors. As a result, we achieved online engagement recognition without degrading accuracy. The proposed model that takes into account the difference of annotators will contribute to other recognition tasks that contain subjectivity such as emotion recognition. We also confirmed that the proposed model can cluster the annotators based on their character distributions and that each character has a different perspective on behaviors for engagement recognition. From the analysis result, we can learn the traits of each annotator or character. Therefore, we can choose an annotator or a character that a live spoken dialogue system wants to imitate. We also presented another method to select a character distribution based on the Big Five personality traits expected for the system.

A further study on adaptive behavior generation of spoken dialogue systems should be conducted. Dialogue systems should consider the result of engagement recognition and appropriately change their dialogue policy for the users. As we have seen in Section II C, a little study has been made

Fig. 7. Real-time engagement visualization tool.

on this issue. We are also studying methods for utilizing the result of engagement recognition. One possible way is to change the dialogue policy according to user engagement adaptively. For example, when a system is given a social role of information navigation such as a laboratory guide, the system mostly takes the dialogue initiative. In this case, it is expected that, when the system recognizes low-level user engagement, the system gives a feedback response that attracts the user's attention. Moreover, it is also possible that the system adaptively changes the explanation content according to the level of user engagement. The explanation content can be elaborated for users with high-level engagement. On the other hand, for users with low-level engagement, the system should make the content more understandable. Another way of utilizing engagement is to use it as an evaluation metric for dialogue. Studies on non-task oriented dialogue such as casual chatting have tried to establish evaluation metrics including the length of dialogue, linguistic appropriateness, and human judgment. However, there is still no clear metric to evaluate dialogue. We will be able to use engagement as an evaluation metric or reference labels for training models.

Finally, we will realize the adaptive behavior generation in our humanoid robot ERICA by utilizing engagement recognition. To this end, we have applied real-time engagement recognition into a dialogue system of ERICA, and also implemented a visualization tool for real-time engagement recognition, as shown in Fig. 7. After we implement the system behavior strategy as mentioned above, we will also conduct dialogue experiments to evaluate the effect of the awareness of user engagement.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI (Grant Number 15J07337) and JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401).

REFERENCES

[1] Higashinaka, R. *et al.*: Towards an open-domain conversational system fully based on natural language processing, in *COLING*, 2014, 928–939.

[2] Hori, C.; Hori, T.: End-to-end conversation modeling track in DSTC6, in *Dialog System Technology Challenges 6*, 2017.

[3] Skantze, G.; Johansson, M.: Modelling situated human-robot interaction using IrisTK, in *SIGDIAL*, 2015, 165–167.

[4] DeVault, D. *et al.*: A virtual human interviewer for healthcare decision support, in *AAMAS*, 2014, 1061–1068.

[5] Young, S.; Gašić, M.; Thomson, B.; Williams, J.D.: POMDP-based statistical spoken dialog systems: A review. *Proc. IEEE.*, **101** (5) (2013), 1160–1179.

[6] Perez, J.; Boureau, Y.-L.; Bordes, A.: Dialog system & technology challenge 6 overview of track 1 - END-to-end goal-oriented dialog learning. In *Dialog System Technology Challenges 6*, 2017.

[7] Schuller, B.; Köhler, N.; Müller, R.; Rigoll, G.: Recognition of interest in human conversational speech, in *INTERSPEECH*, 2006, 793–796.

[8] Wang, W.Y.; Biadsy, F.; Rosenberg, A.; Hirschberg, J.: Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. *Comput. Speech. Lang.*, **27** (1) (2013), 168–189.

[9] Kawahara, T.; Hayashi, S.; Takanashi, K.: Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations, in *INTERSPEECH*, 2013, 1882–1885.

[10] Han, K.; Yu, D.; Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine, in *INTERSPEECH*, 2014, 223–227.

[11] Valstar, M. *et al.*: Depression, mood, and emotion recognition workshop and challenge, in *AVEC 2016: International Workshop on Audio/Visual Emotion Challenge*, 2016, 3–10.

[12] Kahou, S.E. *et al.*: Multimodal deep learning approaches for emotion recognition in video. *J. Multimodal User Interfaces*, **10** (2) (2016), 99–111.

[13] Mizukami, M.; Yoshino, K.; Neubig, G.; Traum, D.; Nakamura, S.: Analyzing the effect of entrainment on dialogue acts, in *SIGDIAL*, 2016, 310–318.

[14] Lubold, N.; Pon-Barry, H.: Acoustic-prosodic entrainment and rapport in collaborative learning dialogues, in *ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014, 5–12.

[15] Matsuyama, Y.; Bhardwaj, A.; Zhao, R.; Akoju, S.; Cassell, J.: Socially-aware animated intelligent personal assistant agent, in *SGIDIAL*, 2016, 224–227.

[16] Müller, P.; Huang, M.X.; Bulling, A.: Detecting low rapport during natural interactions in small groups from non-verbal behaviour, in *IUI*, 2018.

[17] Sidner, C.L.; Lee, C.; Kidd, C.D.; Lesh, N.; Rich, C.: Explorations in engagement for humans and robots. *Artificial Intelligence*, **166** (1–2) (2005), 140–164.

[18] Xu, Q.; Li, L.; Wang, G.: Designing engagement-aware agents for multiparty conversations, in *CHI*, 2013, 2233–2242.

[19] Yu, Z.; Nicolich-Henkin, L.; Black, A.W.; Rudnicky, A.I.: A Wizard-of-Oz study on a non-task-oriented dialog systems that reacts to user engagement, in *SIGDIAL*. 2016, 55–63.

[20] Sun, M.; Zhao, Z.; Ma, X.: Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices, in *CHI*, 2017, 556–567.

[21] Rudovic, O.; Nicolaou, M.A.; Pavlovic, V.: Machine learning methods for social signal processing, in *Social Signal Processing*, Cambridge University Press, 2017, 234–254.

[22] Nakano, Y.I.; Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations, in *IUI*, 2010, 139–148.

[23] Oertel, C.; Mora, K.A.F; Gustafson, J.; Odobez, J.-M.: Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions, in *ICMI*, 2015.

[24] Goffman, E.: Behavior in Public Places: Notes on the Social Organization of Gatherings. Simon and Schuster, USA, 1966.

[25] Glas, N.; Pelachaud, C.: Definitions of engagement in human-agent interaction, in *Int. Workshop on Engagement in Human Computer Interaction*, 2015, 944–949.

[26] Bohus, D.; Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings, in *SIGDIAL*, 2009, 244–252.

[27] Peters, C.: Direction of attention perception for conversation initiation in virtual environments, in *Int. Workshop on Intelligent Virtual Agents*, 2005, 215–228.

[28] Yu, Z.; Bohus, D.; Horvitz, E.: Incremental coordination: Attention-centric speech production in a physically situated conversational agent, in *SIGDIAL*, 2015, 402–406.

[29] Bohus, D.; Andrist, S.; Horvitz, E.: A study in scene shaping: Adjusting F-formations in the wild. In *AAAI Fall Symp. on Natural Communication for Human-Robot Collaboration*, 2017.

[30] Yu, C.; Aoki, P.M.; Woodruff, A.: Detecting user engagement in everyday conversations, in *ICSLP*, 2004, 1329–1332.

[31] Poggi, I.: Mind, Hands, Face, Body: A Goal and Belief View of Multimodal Communication. Weidler, Germany, 2007.

[32] Bednarik, R.; Eivazi, S.; Hradis, M.: Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement, in *ICMI Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2012, 10.

[33] Michalowski, M.P.; Sabanovic, S.; Simmons, R.: A spatial model of engagement for a social robot, in *Int. Workshop on Advanced Motion Control*. 2006, 762–767.

[34] Castellano, G.; Pereira, A.; Leite, I.; Paiva, A.; McOwan, P.W.: Detecting user engagement with a robot companion using task and social interaction-based features, in *ICMI*. 2009, 119–126.

[35] Rich, C.; Ponsler, B.; Holroyd, A.; Sidner, C.L.: Recognizing engagement in human-robot interaction, in *HRI*, 2010, 375–382.

[36] Yu, Z.; Ramanarayanan, V.; Lange, P.; Suendermann-Oeft, D.: An open-source dialog system with real-time engagement tracking for job interview training applications, in *IWSDS*, 2017.

[37] Chiba, Y.; Nose, T.; Ito, A.: Analysis of efficient multimodal features for estimating user's willingness to talk: Comparison of human-machine and human-human dialog, in *APSIPA ASC*, 2017.

[38] Türker, B.B.; Buçinca, Z.; Erzin, E.; Yemez, Y.; Sezgin, M.: Analysis of engagement and user experience with a laughter responsive social robot, in *INTERSPEECH*. 2017, 844–848.

[39] Sanghvi, J.; Castellano, G.; Leite, I.; Pereira, A.; McOwan, P.W.; Paiva, A.: Automatic analysis of affective postures and body motion to detect engagement with a game companion, in *HRI*, 2011, 305–311.

[40] Chiba, Y.; Ito, A.: Estimation of user's willingness to talk about the topic: Analysis of interviews between humans, in *IWSDS*, 2016.

[41] Huang, Y.; Gilmartin, E.; Campbell, N.: Conversational engagement recognition using auditory and visual cues, in *INTERSPEECH*, 2016.

[42] Frank, M.; Tofighi, G.; Gu, H.; Fruchter, R.: Engagement detection in meetings. *arXiv preprint*, 2016. arXiv: 1608.08711.

[43] Sidner, C.L.; Lee, C.: Engagement rules for human-robot collaborative interactions, in *ICSMC*, 2003, 3957–3962.

[44] Inoue, K.; Lala, D.; Nakamura, S.; Takanashi, K.; Kawahara, T.: Annotation and analysis of listener's engagement based on multi-modal behaviors, in *ICMI Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, 2016.

[45] Glas, N.; Prepin, K.; Pelachaud, C.: Engagement driven topic selection for an information-giving agent, in *SemiDial*, 2015, 48–57.

[46] Glas, D.F.; Minaot, T.; Ishi, C.T.; Kawahara, T.; Ishiguro, H.: E.R.I.C.A: The ERATO intelligent conversational android, in *ROMAN*, 2016, 22–29.

[47] Inoue, K.; Milhorat, P.; Lala, D.; Zhao, T.; Kawahara, T.: Talking with ERICA, an autonomous android, in *SIGDIAL*, 2016, 212–215.

[48] Ishi, C.T.; Ishiguro, H.; Hagita, N.: Evaluation of formant-based lip motion generation in tele-operated humanoid robots, in *IROS*, 2012, 2377–2382.

[49] Sakai, K.; Ishi, C.T.; Minato, T.; Ishiguro, H.: Online speech-driven head motion generating system and evaluation on a tele-operated robot, in *ROMAN*, 2015, 529–534.

[50] Ramanarayanan, V.; Leong, C.W.; Suendermann-Oeft, D.: Rushing to judgement: How do laypeople rate caller engagement in thin-slice videos of human-machine dialog? in *INTERSPEECH*, 2017, 2526–2530.

[51] Ramanarayanan, V.; Leong, C.W.; Suendermann-Oeft, D.; Evanini, K.: Crowdsourcing ratings of caller engagement in thin-slice videos of human-machine dialog: Benefits and pitfalls, in *ICMI*, 2017, 281–287.

[52] Den, Y.; Yoshida, N.; Takanashi, K.; Koiso, H.: Annotation of japanese response tokens and preliminary analysis on their distribution in three-party conversations, in *Oriental COCOSDA*, 2011, 168–173.

[53] Blei, D.M.; Ng, A.Y.; Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3** (2003), 993–1022.

[54] Dawid, A.P.; Skene, A.M.: Maximum likelihood estimation of observer error-rates using the EM algorithm. *Appl. Stat.*, **28** (1) (1979), 20–28.

[55] Griffiths, T.L.; Steyvers, M.: Finding scientific topics. *Proc. Natl. Acad. Sci.*, **101** (suppl 1) (2004), 5228–5235.

[56] Ozkan, D.; Sagae, K.; Morency, L.P.: Latent mixture of discriminative experts for multimodal prediction modeling, in *COLING*, 2010, 860–868.

[57] Ozkan, D.; Morency, L.P.: Modeling wisdom of crowds using latent mixture of discriminative experts, in *ACL*, 2011, 335–340.

[58] Kumano, S.; Otsuka, K.; Matsuda, M.; Ishii, R.; Yamato, J.: Using a probabilistic topic model to link observers' perception tendency to personality, in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, 588–593.

[59] Schuller, B. *et al.*: The INTERSPEECH 2013 computational paralinguistics challenge social signals, conflict, emotion, autism, in *INTERSPEECH*, 2013, 148–152.

[60] Inaguma, H.; Inoue, K.; Mimura, M.; Kawahara, T.: Social signal detection in spontaneous dialogue using bidirectional LSTM-CTC, in *INTERSPEECH*, 2017, 1691–1695.

[61] Fujie, S.; Ejiri, Y.; Nakajima, K.; Matsusaka, Y.; Kobayashi, T.: A conversation robot using head gesture recognition as para-linguistic information, in *ROMAN*, 2004, 159–164.

[62] Morency, L.P.; Quattoni, A.; Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition in *CVPR*, 2007.

[63] Lala, D.; Inoue, K.; Milhorat, P.; Kawahara, T.: Detection of social signals for recognizing engagement in human-robot interaction. In *AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.

[64] Barrick, M.R.; Mount, M.K.: The Big Five personality dimensions and job performance: A meta-analysis. *Pers. Psychol.*, **44** (1) (1991), 1–26.

[65] Wada, S.: Construction of the Big Five scales of personality trait terms and concurrent validity with NPI. *Japanese J. Psychol.*, **67** (1) (1996), 61–67.

APPENDIX A: SAMPLING FORMULAS OF LATENT CHARACTER MODEL

We explain how to obtain the sampling formulas of the latent character model (equations (9) and (11)) as follows. The joint probability of the datasets of the characters $\boldsymbol{Z}$ and engagement labels $\boldsymbol{Y}$ where the datasets of the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are marginalized is represents as

$$p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\alpha}, \beta, \gamma) = p(\boldsymbol{Z}|\boldsymbol{\alpha})\ p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \beta, \gamma). \tag{A.1}$$

The first term is expanded by marginalizing the parameter set $\boldsymbol{\Theta}$ as

$$\begin{aligned} p(\boldsymbol{Z}|\boldsymbol{\alpha}) &= \int p(\boldsymbol{Z}|\boldsymbol{\Theta})\ p(\boldsymbol{\Theta}|\boldsymbol{\alpha})\ d\boldsymbol{\Theta} && \text{(A.2)} \\ &= \int \prod_{i=1}^{I}\prod_{k=1}^{K} \theta_{ik}^{D_{ik}} \prod_{i=1}^{I} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \\ &\quad \times \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1} d\boldsymbol{\Theta} && \text{(A.3)} \\ &= \frac{\Gamma(\sum_k \alpha_k)^I}{\prod_k \Gamma(\alpha_k)^I} \int \prod_{i=1}^{I}\prod_{k=1}^{K} \theta_{ik}^{D_{ik}+\alpha_k - 1} d\boldsymbol{\Theta} && \text{(A.4)} \\ &= \frac{\Gamma(\sum_k \alpha_k)^I}{\prod_k \Gamma(\alpha_k)^I} \prod_{i=1}^{I} \frac{\prod_k \Gamma(D_{ik} + \alpha_k)}{\Gamma(D_i + \sum_k \alpha_k)}. && \text{(A.5)} \end{aligned}$$

Note that equation (A.3) is derived from the definitions of the categorical and Dirichlet distributions. The last expansion is based on the partition function of the Dirichlet distribution as

$$\int \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1} d\boldsymbol{\theta}_i = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma\left(\sum_k \alpha_k\right)}. \tag{A.6}$$

Similarly, the second term of equation (A.1) is also expanded by marginalizing the parameter set $\boldsymbol{\Phi}$ as

$$\begin{aligned} & p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \beta, \gamma) \\ &= \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Phi})\ p(\boldsymbol{\Phi}|\beta, \gamma)\ d\boldsymbol{\Phi} && \text{(A.7)} \\ &= \int \prod_{k=1}^{K}\prod_{l=1}^{L} \phi_{kl}^{N_{kl1}} (1 - \phi_{kl})^{N_{kl0}} \\ &\quad \times \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{\Gamma(\beta + \gamma)}{\Gamma(\beta)\Gamma(\gamma)} \phi_{kl}^{\beta - 1} (1 - \phi_{kl})^{\gamma - 1} d\boldsymbol{\Phi} && \text{(A.8)} \\ &= \frac{\Gamma(\beta + \gamma)^{KL}}{\Gamma(\beta)^{KL}\Gamma(\gamma)^{KL}} \\ &\quad \times \int \phi_{kl}^{N_{kl1} + \beta - 1} (1 - \phi_{kl})^{N_{kl0} + \gamma - 1} d\boldsymbol{\Phi} && \text{(A.9)} \\ &= \frac{\Gamma(\beta + \gamma)^{KL}}{\Gamma(\beta)^{KL}\Gamma(\gamma)^{KL}} \prod_{k=1}^{K}\prod_{l=1}^{L} \\ &\quad \times \frac{\Gamma(N_{kl1} + \beta)\Gamma(N_{kl0} + \gamma)}{\Gamma(N_{kl} + \beta + \gamma)}. && \text{(A.10)} \end{aligned}$$

Note that equation (A.8) is derived from the definitions of the Bernoulli and beta distributions. The last expansion is also based on the partition function like equation (A.6).

Using equation (A.5), the expansion from equations (8) to (9) is developed as

$$\begin{aligned} & \frac{p(z_{ijn} = k, \boldsymbol{Z}_{\backslash ijn}|\boldsymbol{\alpha})}{p(\boldsymbol{Z}_{\backslash ijn}|\boldsymbol{\alpha})} \\ &= \frac{\Gamma(D_{ik\backslash ijn} + 1 + \alpha_k) \prod_{k' \neq k} \Gamma(D_{ik'\backslash ijn} + \alpha_{k'})}{\Gamma(D_i + \sum_{k'} \alpha_{k'})} \\ &\quad \Big/ \frac{\prod_{k'} \Gamma(D_{ik'\backslash ijn} + \alpha_{k'})}{\Gamma(D_i - 1 + \sum_{k'} \alpha_{k'})} && \text{(A.11)} \\ &= \frac{\Gamma(D_{ik\backslash ijn} + 1 + \alpha_k)}{\Gamma(D_{ik\backslash ijn} + \alpha_k)} \frac{\Gamma(D_i - 1 + \sum_{k'} \alpha_{k'})}{\Gamma(D_i + \sum_{k'} \alpha'_k)} && \text{(A.12)} \\ &= \frac{D_{ik\backslash ijn} + \alpha_k}{D_i - 1 + \sum_{k'} \alpha_{k'}}. && \text{(A.13)} \end{aligned}$$

Note that we used a property of the gamma function as

$$\Gamma(x+1) = x\Gamma(x). \tag{A.14}$$

Using equation (A.10), the expansion from equations (10) to (11) is also developed as

$$\begin{aligned}
&\frac{p(\boldsymbol{Y}|\boldsymbol{X}, z_{ijn}=k, \boldsymbol{Z}_{\backslash ijn}, \beta, \gamma)}{p(\boldsymbol{Y}_{\backslash ijn}|\boldsymbol{X}, \boldsymbol{Z}_{\backslash ijn}, \beta, \gamma)} \\
&= \prod_{l=1}^{L} \frac{\Gamma(N_{kl1\backslash ijn} + N_{ijnl1} + \beta)}{\Gamma(N_{kl0\backslash ijn} + N_{ijnl0} + \gamma)} \\
&\quad \times \prod_{\substack{k'=1 \\ k' \neq k}}^{K} \prod_{l=1}^{L} \frac{\Gamma(N_{k'l1\backslash ijn} + \beta)\Gamma(N_{k'l0\backslash ijn} + \gamma)}{\Gamma(N_{k'l\backslash ijn} + \beta + \gamma)} \\
&\quad \Big/ \prod_{k'=1}^{K} \prod_{l=1}^{L} \frac{\Gamma(N_{k'l1\backslash ijn} + \beta)\Gamma(N_{k'l0\backslash ijn} + \gamma)}{\Gamma(N_{k'l\backslash ijn} + \beta + \gamma)} \qquad \text{(A.15)} \\
&= \prod_{l=1}^{L} \left\{ \frac{\Gamma(N_{kl\backslash ijn} + \beta + \gamma)}{\Gamma(N_{kl\backslash ijn} + N_{ijnl} + \beta + \gamma)} \right. \\
&\quad \times \frac{\Gamma(N_{kl1\backslash ijn} + N_{ijnl1} + \beta)}{\Gamma(N_{kl1\backslash ijn} + \beta)} \\
&\quad \left. \times \frac{\Gamma(N_{kl0\backslash ijn} + N_{ijnl0} + \gamma)}{\Gamma(N_{kl0\backslash ijn} + \gamma)} \right\}. \qquad \text{(A.16)}
\end{aligned}$$

Koji Inoue received M.S. degree in 2015 and is currently pursuing a Ph.D. degree at Graduate School of Informatics, Kyoto University, Japan. His research interests include multimodal signal processing, human–robot interaction, and spoken dialogue systems.

Divesh Lala received Ph.D. in Graduate School of Informatics in 2015, from Kyoto University, Kyoto, Japan. Currently, he is a Researcher in Graduate School of Informatics, Kyoto University. His research interests include human–agent interaction and multimodal signal processing.

Katsuya Takanashi received B.A. in Faculty of Letters in 1995, M. of Human and Environmental Studies in Graduate School of Human and Environmental Studies in 1997, and Ph.D. in Graduate School of Informatics in 2014, from Kyoto University, Kyoto, Japan. From 2000 to 2005 he was a Researcher at National Institute of Information and Communications Technology, Kyoto, Japan. Currently, he is a Researcher in Graduate School of Informatics, Kyoto University. He has been studying multimodal and multiparty human–human and human–robot interaction from both cognitive scientific and sociological perspectives. He has been a representative of a project on field studies in 'situated' multiparty conversation and also a member of several other projects on the modeling of multimodal interaction in Japan.

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including speech recognition software Julius and the automatic transcription system for the Japanese Parliament (Diet). Dr. Kawahara received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. He was a General Chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He is an editorial board member of Elsevier Journal of Computer Speech and Language, APSIPA Transactions on Signal and Information Processing, and IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is VP-Publications of APSIPA, a board member of ISCA, and a Fellow of IEEE.