Latent Character Model for Engagement Recognition Based on Multimodal Behaviors

Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara

Abstract Engagement represents how much a user is interested in and willing to continue the current dialogue and is the important cue for spoken dialogue systems to adapt the user state. We address engagement recognition based on listener's multimodal behaviors such as backchannels, laughing, head nodding, and eye gaze. When the ground-truth labels are given by multiple annotators, they differ according to each annotator due to the different perspectives on the multimodal behaviors. We assume that each annotator has a latent character that affects its perception of engagement. We propose a hierarchical Bayesian model that estimates both the engagement level and the character of each annotator as latent variables. Furthermore, we incorporate other latent variables to map the input feature into a sub-space. The experimental result shows that the proposed model achieves higher accuracy than other models that do not take into account the character.

1 Introduction

A number of spoken dialogue systems have been developed and practically used in various kinds of contexts such as user assistants and conversational robots. The systems interact with the user in certain tasks such as question answering [10] and medical diagnoses [6]. In most cases, however, the interaction is human-machine specific and much different from the case of human-human dialogue. Our ultimate goal is to realize conversational robots which behave like human beings and pervade many aspects of our daily lives in a symbiotic manner. To this end, it is needed for the systems to recognize and understand the conversational scene such as the user state. In this paper, we focus on user engagement in human-robot interaction. Engagement represents the process by which dialogue participants establish, main-

All authors
Graduate School of Informatics, Kyoto University, Kyoto, Japan,
e-mail: [inoue][lala][takanasi][kawahara]@sap.ist.i.kyoto-u.ac.jp

tain, and end their interaction [25]. Practically, it has been defined as the user state which represents how much a user is interested in and willing to continue the current dialogue [29, 20]. By recognizing user engagement in dialogue, the system can generate adaptive behaviors, which contributes to smooth and natural interaction.

In this study, we address engagement recognition based on listener's multimodal behaviors such as verbal backchannels, laughing, head nodding, and eye gaze. Since these behaviors are used by listeners to express responses toward speakers, it is presumed that these are related to engagement. To obtain the ground-truth labels of engagement, we ask third-party people (annotators) to judge the user engagement of dialogue data. Since the perception of engagement is subjective, the annotation result often depends on each annotator. Previous studies integrated engagement labels among annotators like majority voting to train recognition models [15, 28, 16].

The difference among annotators suggests that each annotator has different perspective on multimodal behaviors and engagement. We assume that each annotator has its latent character, and the character affects his/her perspective for engagement. The latent character represents a kind of template for the perspective on engagement. We propose a latent character model which estimates not only the engagement level but also the character of each annotator as latent variables. The model can simulate each annotator's perception more precisely. This study contributes to a variety of recognition tasks containing subjectivity such as emotion recognition in that the proposed model takes into account the differences and commonalities of multiple annotators.

2 Related work

Engagement has been variously defined in different kinds of studies [8]. The definitions are mainly classified into two types. The first one focuses on the start and end of the interaction. For example, it is defined as "*the process by which two (or more) participants establish, maintain, and end their perceived connection*" [25]. This type is related to other concepts such as *attention* and *involvement* [19, 30]. The second type focuses on the quality of interaction. For example, engagement was defined as "*how much a participant is interested in and attentive to a conversation*" [29] 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*" [20]. This type is related to *interest* and *rapport*. In this study, we consider engagement in the context of the latter type.

Engagement recognition has been widely studied in previous studies. It was formulated as a binary classification problem: engaged or not (disengaged), or a category classification problem [1]. The used features were based on non-linguistic multimodal behaviors. Non-linguistic information is commonly used as features because linguistic information is specific to the dialogue domain and content, and speech recognition is error-prone. Previous studies investigated the relationship between engagement and multimodal behaviors such as spatial information (e.g. lo-

cation, trajectory, distance) [14, 2, 28], eye gaze (e.g. looking at a robot, mutual gaze) [3, 15, 22, 1, 28, 16, 31], facial information (e.g. facial movement, expression, head pose) [3, 4, 31], verbal backchannels [28, 22], head nodding [16], laughing [27], and posture [23, 7]. Additionally, low-level signals such as acoustic and visual features were considered [29, 4, 11]. The initial recognition models were based on heuristic rules [24, 19, 14]. The recent approach is based on machine learning techniques such as support vector machines (SVM) [28, 16, 7, 31], hidden Markov model (HMM) [29], and convolutional neural networks (CNN) [11]. Recently, some researchers have undertaken a study on system behaviors after recognizing user engagement. They found that the engagement level is related to turn-taking behaviors [28, 12]. Other researchers investigated how to handle user disengagement by changing the dialogue policy or changing the system responses [31, 26]. Our purpose of engagement recognition is similar to those of these studies.

In this study, we address a problem of subjectivity on the annotation of engagement. The perception of engagement is subjective and thus often results in disagreement among annotators. Earlier studies took an approach to train a few annotators to avoid disagreement [1, 28, 11, 31]. When the annotators have to consider multimodality, the annotation becomes more complicated and diverse. Besides, it is natural that there are various perspectives for understanding multimodal behaviors. To collect the various perspectives, we can use another approach based on "wisdom of crowds" where many annotators are recruited and asked to annotate user engagement. Previous study integrated the various labels given by the multiple annotators using majority voting [15, 28, 16].

We take into account the various perspectives of the annotators. We assume that each annotator has a latent character that affects his/her perception of engagement. Our proposed model estimates not only user engagement but also the character of each annotator. The model can simulate each annotator's perception of engagement. It is expected that we can understand the differences and common points among the annotators from the annotation data. The similar model considering the difference of annotators is a two-step conditional random fields (CRF), which was proposed for a backchannel prediction task [18, 17]. The prediction model was trained for each annotator, and the final result is determined by voting from the individual models. On the other hand, we train the model based on the character, not for each annotator. Therefore, more robust estimation is expected even if the amount of data for each annotator is small.

3 Annotation of listener's engagement

We have collected a human-robot interaction corpus where an autonomous android robot, named ERICA [13], interacted with a human subject. ERICA was operated by another human subject, called an operator, who was in a remote room. Fig. 1 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

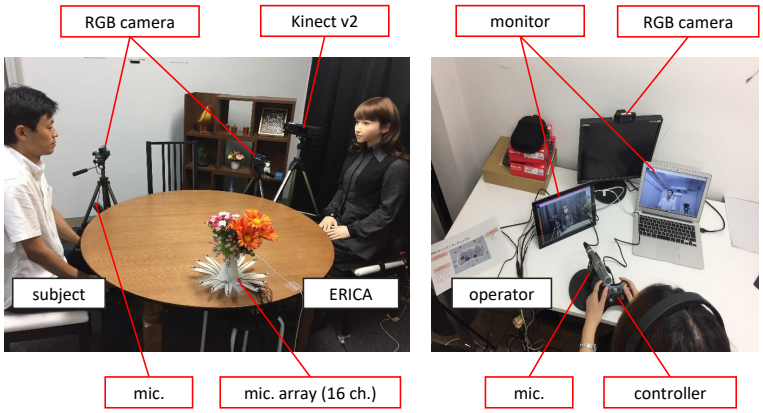

Fig. 1 Setup for conversation

professor was absent for a while, the subject talked with ERICA until the professor would come back. Each dialogue lasted about 10 minutes. The voice uttered by the operator was directly played with a speaker placed on ERICA in real time. We recorded the dialogue with directed microphones, a 16-channel microphone array, RGB cameras, and Kinect v2. We manually annotated utterances, turn units, and dialogue acts. From this corpus, we used 20 sessions for the annotation of subject engagement. The subjects were 12 females and 8 males, with ages ranging from teenagers to over 70 years old. The operators were 6 actresses in their 20s and 30s. One of the 6 actresses was assigned to each session. All the participants were native Japanese speakers.

We annotated subject engagement by recruiting other 12 females who had not participated in the above dialogue experiment. Note that we considered other methods asking the subjects or the operators to annotate subject engagement by themselves right after the dialogue. However, it was hard to make them annotate it due to a time constraint. Besides, we sometimes observe a bias where the subjects tend to give positive evaluations of themselves [21]. Each dialogue session was randomly assigned to 5 annotators. The instructions given to the annotators was as follows. Engagement was defined as "How much the subject is interested in and willing to continue the current dialogue with ERICA". We also explained a list of listener's behaviors which 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. The annotators were asked to watch the dialogue video from ERICA's viewpoint, and to judge subject engagement based on the subject's behaviors. Specifically, the annotators had to press a button when the following three conditions were being met: (1) the subject was being a listener, (2) the subject was expressing any listener's behaviors, and (3) the behavior shows the high level of engagement.

In this study, we use ERICA's conversational turns as a unit for engagement recognition. When an annotator pressed the button during an ERICA's turn more

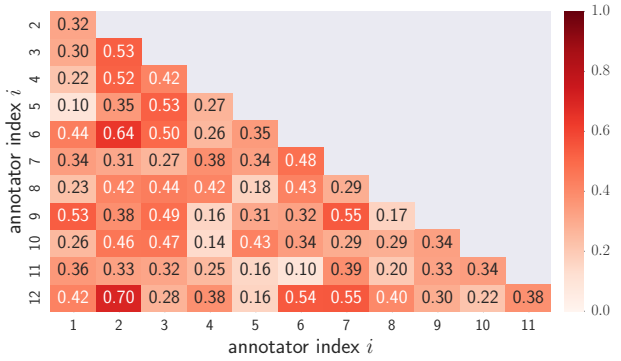

Fig. 2 Inter-annotator agreement scores on each pair of the annotators (Cohen's kappa)

than once, we regarded the turn was annotated as *engaged* by the annotator. Therefore, each turn has binary labels: engaged or not. There were 433 turns in the 20 sessions, and the number of the engaged labels was 894, and that of the not-engaged ones was 1,271. We investigated the agreement score among the annotators. The average value of Cohen's kappa on every pair of two annotators was 0.291 with a standard deviation of 0.229. Fig. 2 shows the matrix of the Cohen's kappa values on each pair among the annotators. Some pairs showed scores which were higher than the moderate agreement (larger than 0.4). This result suggests that the annotators could be clustered into some groups based on their perspectives on multimodal behaviors and engagement.

We also investigated which listener's behaviors were related to engagement. After the annotation work, we asked each annotator to select all meaningful behaviors to annotate subject engagement. As a result, the annotators mostly selected facial expression, laughing, eye gaze, backchannels, head nodding, and body pose. Among them, we use four behaviors, backchannels, laughing, head nodding, and eye gaze in the following experiment. We manually annotated the occurrence of these behaviors. The definition of backchannels was responsive interjections (such as "*huh*" in English and "*un*" in Japanese) and expressive interjections (such as "*oh*" in English and "*he-*" in Japanese) [5]. The laughing was defined as vocal laughing, not just smiling without any vocal utterance. The occurrence of head nodding was judged by the vertical movement of the head. The eye gaze of the subject was annotated as a binary state: the subject was gazing at ERICA's face. We defined the occurrence of eye-gaze behaviors as the event when the subject was gazing at ERICA's face continuously more than 10 seconds. We confirmed the histogram of the continuous gazing times, and then made sure of the certain number of occurrences with this criteria. It was difficult to annotate other behaviors such as facial expression and body pose due to its ambiguity. Note that these behaviors will be considered in the future work.

4 Latent character model using different annotator perspectives

It is essential for engagement recognition to consider various annotator perspectives. The annotation result suggests that each annotator has a different perspective on the multimodal behaviors for the perception of engagement. We assume that the different perspectives can be interpreted by a latent variable called *character*. This character represents a template for the perception of engagement. For example, annotators with one character tend to regard the laughing behavior as the engagement indicator. On the other hand, other annotators with another character tend to regard backchannels as the indicator. We introduce a hierarchical Bayesian model to estimate not only the engagement level but also the latent character from the annotation data. This model is called latent character model and enables us to simulate the various perspectives by considering the different characters.

4.1 Problem formulation

Engagement recognition is done on each system's dialogue turn. The input is based on the occurrences of the four behaviors: laughing, backchannels, head nodding, and eye gaze. Specifically, the input feature is a four-dimensional binary vector corresponding to the combination of the occurrences of the four behaviors, and this is called *behavior pattern*. Therefore, the number of possible states is exponential to the input behaviors ($16 = 2^4$ states in this case). Since this leads to the data sparseness problem, we introduce latent variables to map these behavior patterns into a smaller dimension. The latent variables are called *behavior class*. The output is also binary state: engaged or not. Note that each turn has several ground-truth labels annotated by the multiple annotators. Concretely, the engagement recognition model predicts each annotator's label individually.

4.2 Generative process

The graphical model is depicted in Fig. 3. The generative process is as follows. For each annotator, the character distribution is generated from the Dirichlet distribution as

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

where i, I, K denote the annotator index, the number of annotators, and the number of characters, respectively, and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k, \cdots, \alpha_K)$ is a hyperparameter. The model parameter θ_{ik} represents the probability that the i-th annotator has the k-th character. The behavior-class distribution is generated from the Dirichlet distribution as

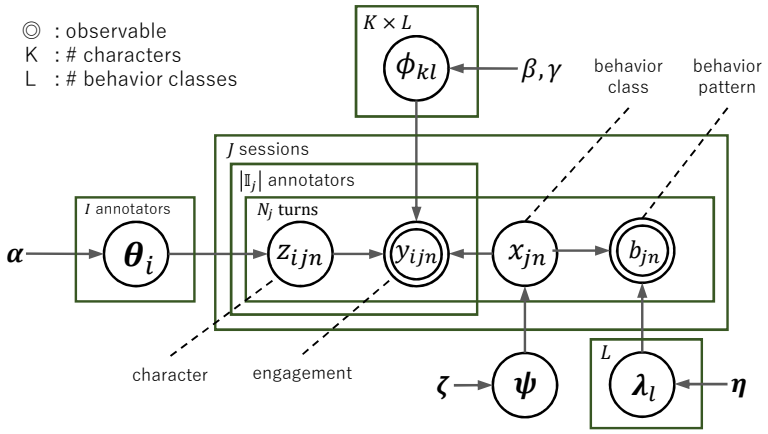

Fig. 3 Graphical model of the proposed model

$$\boldsymbol{\psi} = (\psi_1, \cdots, \psi_l, \cdots, \psi_L) \sim \text{Dirichlet}(\boldsymbol{\zeta}) \,, \tag{2}$$

where l, L denote the behavior-class index, the number of behavior classes, respectively, and $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_l, \cdots, \zeta_L)$ is a hyperparameter. The model parameter ψ_l represents the probability that the l-th behavior class is generated. For each combination of the character and the behavior class, the engagement distribution is generated from the beta distribution as

$$\phi_{kl} \sim \text{Beta}(\beta, \gamma) \,,\ 1 \leq k \leq K \,,\ 1 \leq l \leq L \,, \tag{3}$$

where β and γ are hyperparameters. For example, the parameter ϕ_{kl} represents the probability that annotators with k-th character give the engaged label when they observe the l-th behavior class. For each behavior class, the behavior-pattern distribution is generated from the Dirichlet distribution as

$$\boldsymbol{\lambda}_l = (\lambda_{l1}, \cdots, \lambda_{lm}, \cdots, \lambda_{lM}) \sim \text{Dirichlet}(\boldsymbol{\eta}) \,,\ 1 \leq l \leq L, \tag{4}$$

where m, M denote the behavior-pattern index, the number of behavior patterns, respectively, and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_m, \cdots, \eta_M)$ is a hyperparameter. For example, the parameter λ_{lm} represents the l-th behavior class generates the m-th behavior pattern. In the case of the current setting, the number of behavior patterns (M) is 16.

There are J dialogue sessions, and the set of annotator indices who annotated the j-th session is represented as $\mathbb{I}_j$. Besides, there are N_j system's dialogue turns in the j-th session. For each turn, the character of the i-th annotator is generated from the categorical distribution as

$$z_{ijn} \sim \text{Categorical}(\boldsymbol{\theta}_i) \,,\ i \in \mathbb{I}_j \,,\ 1 \leq j \leq J \,,\ 1 \leq n \leq N_j \,, \tag{5}$$

where n denotes the turn index. In addition, the behavior class is generated from the categorical distribution as

$$x_{jn} \sim \text{Categorical}(\boldsymbol{\psi}) \ , \ 1 \leq j \leq J \ , \ 1 \leq n \leq N_j \ . \tag{6}$$

Based on the generated behavior class, the behavior pattern is observed from the categorical distribution as

$$b_{jn} \sim \text{Categorical}(\boldsymbol{\lambda}_{x_{jn}}) \ , \ 1 \leq j \leq J \ , \ 1 \leq n \leq N_j \ . \tag{7}$$

The behavior patterns correspond to the input features. When the i-th annotator with the character z_{ijn} perceives the behavior class x_{jn}, the engagement label is observed based on the Bernoulli distribution as

$$y_{ijn} \sim \text{Bernoulli}(\phi_{z_{ijn}x_{jn}}) \ , \ i \in \mathbb{I}_j \ , \ 1 \leq j \leq J \ , \ 1 \leq n \leq N_j \ . \tag{8}$$

The engagement labels correspond to the outputs of the model. Among the above variables, the characters and the behavior classes are latent variables, and the engagement labels and the behavior patterns are observable.

Given the data set of the above variables and parameters, the joint distribution is represented as

$$\begin{aligned} &p(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{B},\boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Psi},\boldsymbol{\Lambda}) = \\ &\quad p(\boldsymbol{X}|\boldsymbol{\Psi})p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Phi})p(\boldsymbol{Z}|\boldsymbol{\Theta})p(\boldsymbol{B}|\boldsymbol{X},\boldsymbol{\Lambda})p(\boldsymbol{\Theta})p(\boldsymbol{\Phi})p(\boldsymbol{\Psi})p(\boldsymbol{\Lambda}) \ , \end{aligned} \tag{9}$$

where the bold capital letters represent the data sets of the variables written by those small letters. Note that $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Lambda}$ are the model parameters.

4.3 Training

In the training phase, the model parameters $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Lambda}$ are estimated. We use the collapsed Gibbs sampling which marginalizes the model parameters and iteratively and alternatively samples the latent variables. Here, we sample the character z_{ijn} and the behavior class x_{jn} from those conditional probability distributions as

$$z_{ijn} \sim p(z_{ijn}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}_{\backslash ijn},\boldsymbol{\alpha},\beta,\gamma) \ , \tag{10}$$

$$x_{jn} \sim p(x_{jn}|\boldsymbol{X}_{\backslash jn},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{B},\beta,\gamma,\boldsymbol{\zeta},\boldsymbol{\eta}) \ , \tag{11}$$

where the model parameters $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Lambda}$ are marginalized. Note that $\boldsymbol{Z}_{\backslash ijn}$ and $\boldsymbol{X}_{\backslash jn}$ are the set of characters without z_{ijn} and the set of behavior classes without x_{jn}, respectively. The detail of sampling formulas is omitted here, but it can be obtained in the same manner as the other work [9]. After sampling, we select one of the sampling results as $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$ where the joint probability $p(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{B}|\boldsymbol{\alpha},\beta,\gamma,\boldsymbol{\zeta},\boldsymbol{\eta})$ is maximized. The model parameters $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Lambda}$ are estimated based on the

sampling result $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$ as

$$\theta_{ik} = \frac{D_{ik} + \alpha_k}{\sum_{k'=1}^{K} (D_{ik'} + \alpha_{k'})} , \tag{12}$$

$$\phi_{kl} = \frac{N_{kl1} + \beta}{N_{kl1} + N_{kl0} + \beta + \gamma} , \tag{13}$$

$$\psi_l = \frac{T_l + \zeta_l}{\sum_{l'=1}^{L} (T_{l'} + \zeta_{l'})} , \tag{14}$$

$$\lambda_{lm} = \frac{S_{lm} + \eta_m}{\sum_{m'=1}^{M} (S_{lm'} + \eta_{m'})} . \tag{15}$$

Note that D_{ik} is the number of turns where the i-th annotator has the k-th character. N_{kl1} is the number of times when annotators with the k-th character gave the engaged labels for the l-th behavior class. Similarly, N_{kl0} is the number of the not-engaged labels. T_l is the number of times when the l-th behavior class was generated. Finally, S_{lm} is the number of times when the m-th behavior pattern was observed from the l-th behavior class. These numbers are counted up among the sampling results $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$ and also the observable datasets $\boldsymbol{Y}$ and $\boldsymbol{B}$

4.4 Testing

In the testing phase, the unseen engagement label given by a target annotator is predicted by using the estimated model parameters. Specifically, the model is given the estimated model parameters $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Lambda}$, the input behavior pattern b_t, and the target annotator index i. Note that t represents the turn index in the test data. Given the input behavior pattern, the probability of each behavior class is calculated as

$$p(l|b_t, \boldsymbol{\Psi}, \boldsymbol{\Lambda}) = \frac{1}{\Xi} \ \psi_l \ \lambda_{lb_t} , \tag{16}$$

where Ξ is the partition function. The probability that the target annotator gives the engaged label is calculated by marginalizing both the characters and the behavior classes as

$$p(y_{it} = 1|b_t, i, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Lambda}) = \sum_{k=1}^{K} \theta_{ik} \sum_{l=1}^{L} \phi_{kl} \ p(l|b_t, \boldsymbol{\Psi}, \boldsymbol{\Lambda}) . \tag{17}$$

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

Table 1 Recognition result (average accuracy)

K (#character)	L (#behavior class)				
	2	4	8	12	16
1	0.669	0.667	0.667	0.662	0.667
2	0.695	0.698	0.702	0.705	0.702
3	0.698	0.714	0.708	0.702	0.712
4	0.697	0.709	0.705	0.712	0.707
5	0.689	0.707	0.708	0.711	0.703

5 Experimental evaluation

We compared the proposed model with other methods which do not consider the different annotator perspectives. We conducted the cross-validation with the 20 dialogue sessions: 19 for training and the rest for testing. In this experiment, we used the input behavior patterns which were manually annotated. The output ground-truth labels were the annotation results described in Section 3. In the proposed model, the number of sampling was 3,000, and all prior distributions were the uniform distribution. The number of characters (K) was changed from 1 to 5 on a trial basis. The unique character ($K = 1$) means the case where we do not consider the different perspectives. Besides, the number of behavior classes (L) was chosen from $\{2,4,8,12,16\}$.

The evaluation metric is as follows. Each session has different ground-truth labels given by 5 annotators. We evaluated each annotator's labels individually. Given the target annotator index i, the engaged probability (Eq. 17) was calculated for each turn. Setting the threshold at 0.5, we calculated the accuracy which is a ratio of the number of the correct turns to the total number. We averaged the accuracy scores for all five annotators and also among the cross-validation. The chance level was 0.579 (= 1,271 / 2,165).

Compared methods are based on the logistic regression. We considered two types of training: *majority* and *individual*. In the *majority* type, we integrated the training labels of the five annotators by the majority voting and trained an unique model which is 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 target annotator index i in the test phase. Although the *individual* type can learn the different perspective of each annotator, the amount of training data is much smaller.

Table. 1 summarizes the recognition accuracy with the proposed model. Note that the accuracies of the compared methods are 0.670 and 0.681 for *majority* and *individual* types, respectively. The proposed method achieves the accuracy by 0.714 which is higher than those of the compared methods. The best accuracy was achieved when the number of characters (K) is 3, and the number of behavior classes (L) is 4. This result suggests that the character can be substantially represented in 3 dimensions and the behavior patterns are potentially classified by 4 variables.

6 Conclusion

We have addressed engagement recognition from listener's multimodal behaviors in spoken dialogue. The different perspectives of multiple annotators are represented by the latent characters in the proposed model. Besides, the input behavior pattern is classified into smaller meaningful classes. The proposed latent character model achieved the higher accuracy than the compared methods which do not consider the character. In future work, we will implement a spoken dialogue system utilizing the engagement recognition model. To this end, the engagement recognition model will be integrated with automatic behavior detection methods. Furthermore, we will design the system behaviors after the system recognizes the user engagement.

Acknowledgments

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

References

1. 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: Proc. ICMI Workshop on Eye Gaze in Intelligent Human Machine Interaction (2012)
2. Bohus, D., Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings. In: Proc. SIGDIAL, pp. 244–252 (2009)
3. 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: Proc. ICMI, pp. 119–126 (2009)
4. Chiba, Y., Ito, A.: Estimation of users willingness to talk about the topic: Analysis of interviews between humans. In: Proc. IWSDS (2016)
5. Den, Y., Yoshida, N., Takanashi, K., Koiso, H.: Annotation of japanese response tokens and preliminary analysis on their distribution in three-party conversations. In: Proc. Oriental COCOSDA, pp. 168–173 (2011)
6. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.P.: SimSensei kiosk: A virtual human interviewer for healthcare decision support. In: Proc. Autonomous Agents and Multi-Agent Systems, pp. 1061–1068 (2014)
7. Frank, M., Tofighi, G., Gu, H., Fruchter, R.: Engagement detection in meetings. arXiv preprint arXiv:1608.08711 (2016)
8. Glas, N., Pelachaud, C.: Definitions of engagement in human-agent interaction. In: Proc. International Workshop on Engagement in Human Computer Interaction, pp. 944–949 (2015)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(suppl 1), 5228–5235 (2004)

10. Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing. In: Proc. COLING, pp. 928–939 (2014)
11. Huang, Y., Gilmartin, E., Campbell, N.: Conversational engagement recognition using auditory and visual cues. In: Proc. INTERSPEECH (2016)
12. Inoue, K., Lala, D., Nakamura, S., Takanashi, K., Kawahara, T.: Annotation and analysis of listener's engagement based on multi-modal behaviors. In: Proc. ICMI Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (2016)
13. Inoue, K., Milhorat, P., Lala, D., Zhao, T., Kawahara, T.: Talking with ERICA, an autonomous android. In: Proc. SIGDIAL, pp. 212–215 (2016)
14. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot. In: Proc. International Workshop on Advanced Motion Control, pp. 762–767 (2006)
15. Nakano, Y.I., Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: Proc. IUI, pp. 139–148 (2010)
16. 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: Proc. ICMI (2015)
17. Ozkan, D., Morency, L.P.: Modeling wisdom of crowds using latent mixture of discriminative experts. In: Proc. ACL, pp. 335–340 (2011)
18. Ozkan, D., Sagae, K., Morency, L.P.: Latent mixture of discriminative experts for multimodal prediction modeling. In: Proc. COLING, pp. 860–868 (2010)
19. Peters, C.: Direction of attention perception for conversation initiation in virtual environments. In: Proc. International Workshop on Intelligent Virtual Agents, pp. 215–228 (2005)
20. Poggi, I.: Mind, hands, face and body: A goal and belief view of multimodal communication. Weidler (2007)
21. 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, pp. 2526–2530 (2017)
22. Rich, C., Ponsler, B., Holroyd, A., Sidner, C.L.: Recognizing engagement in human-robot interaction. In: Proc. HRI, pp. 375–382 (2010)
23. 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: Proc. HRI, pp. 305–311 (2011)
24. Sidner, C.L., Lee, C.: Engagement rules for human-robot collaborative interactions. In: Proc. ICSMC, pp. 3957–3962 (2003)
25. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence **166**(1-2), 140–164 (2005)
26. Sun, M., Zhao, Z., Ma, X.: Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices. In: CHI, pp. 556–567 (2017)
27. 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, pp. 844–848 (2017)
28. Xu, Q., Li, L., Wang, G.: Designing engagement-aware agents for multiparty conversations. In: Proc. CHI, pp. 2233–2242 (2013)
29. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. In: Proc. ICSLP, pp. 1329–1332 (2004)
30. 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: Proc. SIGDIAL, pp. 55–63 (2016)
31. 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: Proc. IWSDS (2017)