Interspeech 2018
2-6 September 2018, Hyderabad

Engagement Recognition in Spoken Dialogue via Neural Network by Aggregating Different Annotators' Models

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

Graduate School of Informatics, Kyoto University, Japan
{inoue, lala, takanasi, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

This paper addresses engagement recognition based on four multimodal listener behaviors - backchannels, laughing, eye-gaze, and head nodding. Engagement is an indicator of how much a user is interested in the current dialogue. Multiple third-party annotators give ground truth labels of engagement in a human-robot interaction corpus. Since perception of engagement is subjective, the annotations are sometimes different between individual annotators. Conventional methods directly use integrated labels, such as those generated through simple majority voting, and do not consider each annotator's recognition. We propose a two-step engagement recognition where each annotator's recognition is modeled and the different annotators' models are aggregated to recognize the integrated label. The proposed neural network consists of two parts. The first part corresponds to each annotator's model which is trained with the corresponding labels independently. The second part aggregates the different annotators' models to obtain one integrated label. After each part is pre-trained, the whole network is fine-tuned through back-propagation of prediction errors. Experimental results show that the proposed network outperforms baseline models which directly recognize the integrated label without considering differing annotations.

Index Terms: engagement, multimodal, behaviors, different labels, neural network

1. Introduction

A number of spoken dialogue systems have been developed and deployed in conversational agents and robots. Some systems handle situated interactions such as guidance [1, 2] and quiz games [3, 4], while others are designed to conduct chatting [5]. In these dialogue scenarios, where the systems are not entirely reactive, the systems should recognize whether the user is being engaged in the current dialogue [6, 7]. In the case of human-human conversations, people can recognize engagement from listener behaviors such as backchannels.

Engagement recognition has been widely studied [8]. Engagement represents the process by which participants establish, maintain, and end their interaction [9]. In the field of human-robot interaction, engagement is defined as the user state which represents how much a user is interested in and willing to continue the current dialogue [10, 11]. For example, by recognizing user engagement, the systems can control turn-taking behaviors [12, 13] and dialogue policies [14, 15, 16], and increase the quality of user experience through the dialogue. For input features of engagement recognition, we can exploit non-verbal multimodal behaviors such as eye-gaze [17, 18, 19, 20, 12, 21, 15], backchannels (e.g., "*yeah*") [19, 21], laughing [22], head nodding [21], facial movement and direction [17, 15], spatial location and distance [23, 24, 12], and conversational interaction features like adjacency pairs [19].

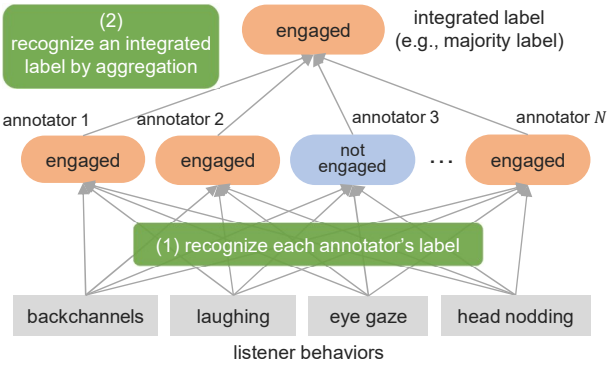

Figure 1: *Overview of the proposed method*

In addition, direct use of low-level signals such as acoustic and image features was explored [10, 25, 26, 27]. Although such recognition models were initially based on heuristic rules [9, 28, 23], recent approaches are based on machine learning techniques [10, 12, 21, 29, 26, 15, 27].

In this study, we address engagement recognition based on multimodal listener behaviors such as backchannels, laughing, eye-gaze, and head nodding. To obtain ground-truth labels of engagement, we ask third-party annotators to annotate the labels. However, perception of engagement is subjective and may depend on each annotator. Therefore, the ground truth label of engagement is sometimes inconsistent between annotators. Previous studies integrated the different labels by using simple methods such as majority voting, and trained a model with the integrated labels [18, 21]. They did not use the raw labels which differed from one annotator to another.

We assume that each annotator's recognition can help to recognize the integrated label. Therefore, we use not only the integrated labels but also different annotators' labels. In our previous work, we proposed a hierarchical Bayesian model to recognize each annotator's label [30]. In this paper, we propose a neural network to predict the integrated label by considering the different annotators' labels. Figure 1 depicts the overview of the proposed method. The proposed neural network consists of two parts. The first part corresponds to recognition of each annotator's label. The second part aggregates the results of the first part to recognize the integrated label. Each part is pre-trained one by one, and then the whole network is fine-tuned. It is expected that the pre-training and fine-tuning lead to efficient training for the network so that we improve the recognition accuracy. The proposed network is more natural modeling in that each annotator's model is captured and aggregated to recognize the integrated labels, and this study contributes to studies on recognition tasks with human subjectivity such as emotion recognition.

10.21437/Interspeech.2018-2067

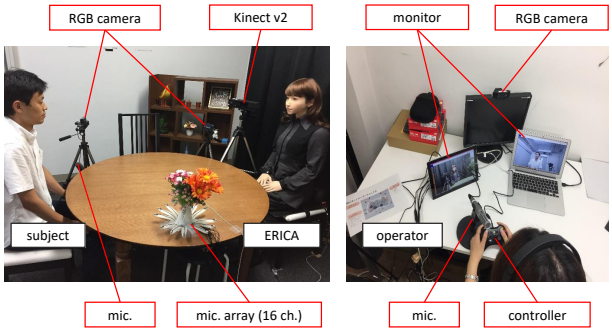

Figure 2: *Snapshot of data collection*

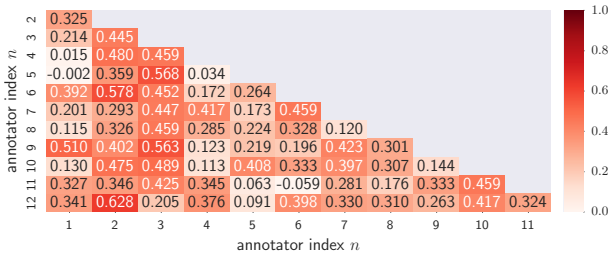

Figure 3: *Inter-rater agreement scores (Cohen's kappa coefficient) for each pair of the annotators*

2. Dialogue corpus and annotation of engagement

We have collected a human-robot interaction corpus in which the android (humanoid) robot ERICA [31, 32] interacted with a human subject. ERICA was operated by another human subject, called an operator, who was being in a remote room. Fig. 2 shows a snapshot of the corpus. We asked the robot operator to think of and utter her dialogue content by herself. The voice of the operator was directly played with a speaker placed in ERICA. We use 20 dialogue sessions for annotation of the subject engagement in this paper. 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. Whereas each subject participated in only one session, each operator participated some sessions. All participants were native Japanese speakers.

There are several methods to annotate the ground-truth data of the subject engagement. The direct method is to ask the subject of the dialogue to evaluate his/her own engagement right after the dialogue session. However, we observed some bias where the subjects tend to give positive evaluations on themselves. This kind of bias was also observed in other works [33, 34]. The second method is to ask the operator to evaluate the subject engagement. However, due to time constraints, the actresses could not participate in this annotation work. Many other studies adopt a practical method to ask third-party people (annotators) to evaluate engagement by watching a video of the dialogue. There are two approaches to this annotation: training a small number of annotators [20, 12, 26, 15] and making use of the wisdom of crowds [18, 21]. The latter is realistic for a large-scale annotation, thus we took this approach.

The annotators were 12 females who had not participated in the dialogue experiment and were recruited in our university. We randomly selected 5 of the 12 annotators for each dialogue session. The instructions given to them were as follows. The definition of engagement was presented as *how much the subject is interested in and willing to continue the current dialogue*. We also gave a list of listener behaviors that could be related to engagement. This list included facial expression, laughing, eye-gaze, backchannels, head nodding, body pose, moving of shoulders, and moving of arms or hands. Watching the dialogue video from the robot's viewpoint, each annotator was asked to press a button when both conditions were being met: the subject is expressing any listener behaviors and the annotator interprets that the behavior suggests a high level of engagement.

We use the robot's conversational turn as a unit for engagement recognition in this study. If an annotator pressed the button once or more in a turn, the turn is regarded as engaged. The total number of robot turns was 433 in 20 dialogue sessions. We excluded short turns whose durations are shorter than 3 seconds, and also turns corresponding to greetings. The numbers of engaged and not engaged turns annotated by individual annotators were 894 and 1,271 respectively. The average value of Cohen's kappa coefficients on every pair of two annotators was 0.291 with a standard deviation of 0.229. This shows the difficulty of this annotation work which has high subjectivity. However, as Fig. 3 shows, some annotator pairs showed higher coefficients than the moderate agreement (larger than 0.4). The result suggests that each annotator had a different perspective on multimodal behaviors for engagement recognition, but some annotators had similar tendencies. We integrated the labels by majority voting with the five annotators. If more than three annotators gave an engaged label, the integrated label was annotated as engaged. In terms of integrated labels, the number of engaged and not engaged turns were 166 and 267 respectively.

We also asked the subjects to complete a survey on which behaviors were related to engagement. For each dialogue session, we asked each annotator to select all meaningful behaviors in order to judge the engagement level. The result indicated that engagement could be related to multiple behaviors of backchannels, laughing, eye-gaze, head nodding, facial expression, and body pose. Among them, we use four behaviors: backchannels, laughing, eye-gaze, and head nodding in the following experiment. It was difficult to annotate facial expression and body pose due to their ambiguity.

3. Neural network aggregating different annotators' models

Each annotator's recognition should be individually modeled in the case where the recognition task has subjectivity. We propose a neural network that consists of two parts where the first part corresponds to each annotator's recognition model and the second part aggregates the different annotators' models to recognize the majority label. Similar models that aggregate different annotators' models were proposed based on this two-step approach [35, 36]. These models used conditional random fields (CRF), while our model uses neural networks, with the advantage being that we can fine-tune the whole network in the end-to-end manner.

3.1. Problem formulation

Engagement recognition is done for each turn of the robot. The input is based on listener behaviors of the user during the turn: backchannels, laughing, eye-gaze, and head nodding. Table 1 summarizes the used feature set. We distinguished backchan-

Table 1: *Feature set of listener behaviors (7 dimensions)*

behavior type	feature
backchannels	(1) # of responsive interjections
	(2) # of expressive interjections
laughing	(3) # of laughing
eye-gaze	(4) time ratio of gaze toward robot
	(5) total time of gaze toward robot
	(6) # of gaze switching toward robot
head nodding	(7) # of head nodding

nels into two types: responsive interjections (such as "*yeah*" in English and "*un*" in Japanese) and expressive interjections (such as "*oh*" in English and "*he-*" in Japanese) [37]. We manually annotated these features on the human-robot interaction corpus. Note that we are also working on automatic detection of these behaviors [38, 39]. In future work, we will integrate the engagement recognition model with these automatic detection methods for practical spoken dialogue systems. The output is binary, engaged or not, for each turn. The reference label is based on the majority which was voted from labels of the five annotators.

3.2. Network architecture

The proposed network consists of two parts as illustrated in Figure 4. Each part is realized by a linear interpolation of a GRU (gated recurrent unit) [40] and a linear transformation, inspired by the recurrent high-way network [41]. It is expected that the GRU captures context information through turns and the linear transformation captures local information of the current behavior input.

The network architecture is as follows. The input vector is represented as:

$$\boldsymbol{X}_t = (x_{1t}, \cdots, x_{bt}, \cdots, x_{Bt})^T , \tag{1}$$

where x_{bt} represents the feature value of the b-th behavior in the t-th robot turn. B is the number of behavior features (B=7 in this case). In the first part, a combination of a GRU and a linear transformation is prepared for each annotator, and it is trained to recognize each annotator's label. For the n-th annotator's model, the input vector is fed into the three functions as:

$$Z_{tn} = \sigma(\mathrm{GRU}_n(\boldsymbol{X}_t)) , \tag{2}$$
$$T_{tn} = \sigma(W_{Tn}\boldsymbol{X}_t + b_{Tn}) , \tag{3}$$
$$C_{tn} = \sigma(W_{Cn}\boldsymbol{X}_t + b_{Cn}) , \tag{4}$$

where $n \in \{1, \cdots, N\}$ represents the annotator index, $\sigma(\cdot)$ is the sigmoid function, and GRU$(\cdot)$ is the gated recurrent unit that stores a context hidden state inside it. N is the number of total annotators in the training data (N=12 in this case). Afterwards, the outputs of the GRU and the linear transformation is linearly interpolated by the weight parameter as:

$$Y_{tn} = C_{tn} \cdot T_{tn} + (1 - C_{tn}) \cdot Z_{tn} . \tag{5}$$

This output corresponds to the recognition result of the n-th annotator in the t-th turn. In the second part, the outputs of all the annotators' models are concatenated as:

$$\boldsymbol{Y}_t = (Y_{t1}, \cdots, Y_{tn}, \cdots, Y_{tN})^T . \tag{6}$$

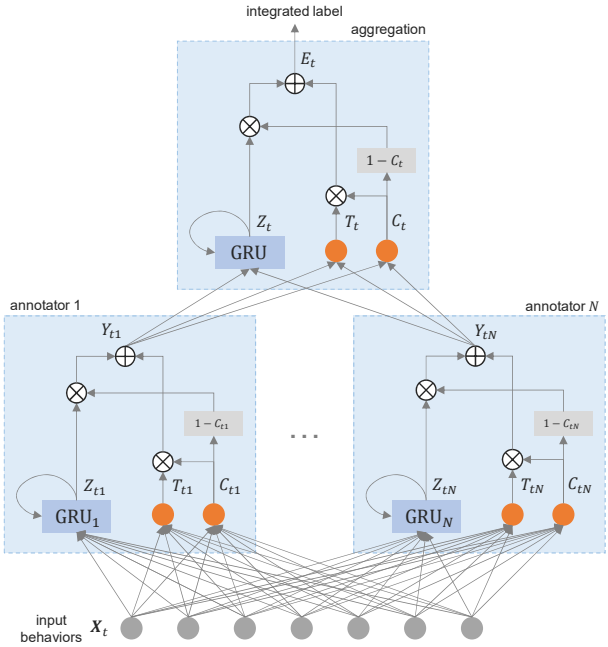

Figure 4: *Proposed network architecture*

The same combination of a GRU and a linear transformation is applied to the concatenated vector to aggregate the different outputs as:

$$Z_t = \sigma(\mathrm{GRU}(\boldsymbol{Y}_t)) , \tag{7}$$
$$T_t = \sigma(W_T\boldsymbol{Y}_t + b_T) , \tag{8}$$
$$C_t = \sigma(W_C\boldsymbol{Y}_t + b_C) . \tag{9}$$

Finally, the probability of engaged in the turn is obtained in the similar way as the first part as:

$$E_t = C_t \cdot T_t + (1 - C_t) \cdot Z_t . \tag{10}$$

If the output E_t is larger than a threshold, the t-th turn is recognized as *engaged* otherwise *not engaged*.

3.3. Pre-training and fine-tuning

To make the first part of the proposed method realize each annotator's recognition and to make the second part aggregate different annotators' models, we pre-train each part one by one. Although we can directly train the whole network using the integrated labels, it is expected that the pre-training leads to more efficient learning. Furthermore, the direct training sometimes falls into local minima. Similar to the current case, it is often the case that each annotator partially gives ground-truth labels. In other words, some annotations of individual annotators are missing. In this case, the separated pre-training and fine-tuning would be a practical approach.

At first, we train the first part by using each annotator's labels. Using the n-th annotator's labels, we train the n-th annotator's model. The cost function is the squared error as $|Y_{tn} - Y'_{tn}|^2$, where $Y'_{tn} \in \{0, 1\}$ is the reference label of the n-th annotator in the t-th turn. Second, after fixing the first-part parameters, we train only the second part parameters by using the integrated labels. The cost function is also the squared error as $|E_t - E'_t|^2$ where $E'_t \in \{0, 1\}$ is the integrated reference label (the majority label) in the t-th turn. Finally, we fine-tune the

Table 2: *Recognition accuracy*

method	accuracy
(A) majority baseline	0.617
(B) one-layer	0.639
(C) two-layers (only fine-tuning)	0.686
(D) two-layers (only pre-training)	0.685
(E) proposed	0.728

whole network by using the integrated labels in the end-to-end manner. The cost function is same as the second part.

4. Experimental evaluations

We conducted cross validation of 20 dialogue sessions, with 19 used for training and the rest for testing per fold. The input behavior features were manually annotated. The settings of the proposed model are as follows. We used the Adam algorithm as the optimizer [42]. The minibatch size corresponded to one session. The number of epochs were 30 and 100 for the fine-tuning and the pre-training, respectively. The dropout rate was set at 0.2. We implemented the neural networks with Chainer 3.5.0[1] The reference labels were the majority labels defined in Section 2. The probability of engaged (Eq. (10)) was calculated for each robot turn. Setting the decision threshold at 0.5, we calculated the accuracy score which is a ratio of the number of the correctly recognized turns to the total number of the turns. The final evaluation was the averaged accuracy among all folds in the cross validation. The accuracy score of a majority baseline was 0.617 (=267/433).

4.1. Effect of aggregating different annotators' models

The proposed model is compared with some methods which do not consider the different annotators' models. The first method is a single-layer network using the combination of a GRU and a linear transformation, referred to as *one-layer*. This method was directly trained with the integrated labels. The second method is the same architecture of the proposed model but conducted only the fine-tuning directly without the pre-training, referred to as *two-layers (only fine-tuning)*. Additionally, we tested a method without the fine-tuning, referred to as *two-layers (only pre-training)*. This method partially corresponds to the earlier studies [35, 36].

The results are summarized in Table 2. The proposed method achieved the higher accuracy than the others. Comparing the one-layer model (B) with the two-layer models (C-E), the two-layer models had higher accuracies. This suggests that the more complicated network is effective to recognize the overall integrated label. Comparing the proposed method (E) with the fine-tuning only model (C), the pre-training is effective for this task and aggregating different annotators' models is important. Finally, the proposed method (E) showed higher accuracy than the pre-training only model (D), and this gain is derived from the advantage of using neural networks where we can fine-tune the whole network.

4.2. Identifying important features

We also examined the effect of each behavior feature. We eliminated each of the used features and tested the proposed model with the same cross validation. The results are reported in Table 3. Note that the greater the decrease in accuracy, the more effective that feature is for the model. For backchannels, the responsive interjections (1), such as "*yeah*", was not so useful for the model. On the other hand, expressive interjections (2), such as "*oh*", are important because they are reactions toward what was said and better reflects the interest of the listener. laughing (3) is the most effective, so this behavior is important for engagement recognition. With regard to eye-gaze, while time ratio (4) was not effective, the total time duration (5) and number of gaze switching (6) were effective. Finally, head nodding (7) was also effective.

Table 3: *Recognition accuracy without each feature on the proposed method*

eliminated feature	accuracy
nothing	0.728
(1) # of responsive interjections	0.730 (▲ 0.002)
(2) # of expressive interjections	0.712 (▼ 0.016)
(3) # of laughing	0.696 (▼ 0.032)
(4) time ratio of gaze toward robot	0.724 (▼ 0.004)
(5) total time of gaze toward robot	0.698 (▼ 0.030)
(6) # of gaze switching toward robot	0.697 (▼ 0.031)
(7) # of head nodding	0.713 (▼ 0.015)

5. Conclusions

We have proposed a neural network for engagement recognition in spoken dialogue. Since perception of engagement is subjective, the ground-truth data depends on each annotator. A network aggregating different labels may help recognition of integrated labels (e.g., majority voting). The proposed method consists of two parts - recognition of each annotator's label, and aggregation of different annotators' models to recognize the overall integrated label. Each part is realized by the linear combination of the GRU and the linear transformation. The first part was pre-trained with the different annotators' labels, and the second part was also pre-trained with the integrated labels by fixing the parameters of the first part. Afterward, the whole network was fine-tuned. The experimental result shows that pre-training both each annotator's model and aggregation of different annotators' models is effective. The result also indicates that laughing and some eye-gaze features are informative in this engagement recognition task, followed by expressive interjections and head nodding.

As future work, we are now incorporating this engagement recognition model into practical spoken dialogue systems. The recognition model is being integrated with automatic detection models of multimodal behaviors [38, 39] to realize online engagement recognition. We are also developing a dialogue system using the online engagement recognition and also designing the system actions according to user engagement. We will conduct dialogue experiments to evaluate the effect of awareness of user engagement.

6. Acknowledgements

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

[1] https://chainer.org

7. References

[1] D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis, A. Leuski, D. Noren, and W. Swartout, "Ada and grace: Direct interaction with museum visitors," in *IVA*, 2012, pp. 245–251.

[2] D. Bohus, C. W. Saw, and E. Horvitz, "Directions robot: in-the-wild experiences and lessons learned," in *AAMAS*, 2014, pp. 637–644.

[3] G. Skantze and M. Johansson, "Modelling situated human-robot interaction using IrisTK," in *SIGdial*, 2015, pp. 165–167.

[4] I. Leite, A. Pereira, A. Funkhouser, B. Li, and J. F. Lehman, "Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction," in *ICMI*, 2016, pp. 13–20.

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

[6] L. Cerrato and N. Campbell, "Engagement in dialogue with social robots," in *IWSDS*, 2016.

[7] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions," in *ICMI*, 2017, pp. 464–472.

[8] N. Glas and C. Pelachaud, "Definitions of engagement in human-agent interaction," in *International Workshop on Engagement in Human Computer Interaction*, 2015, pp. 944–949.

[9] C. L. Sidner and C. Lee, "Engagement rules for human-robot collaborative interactions," in *ICSMC*, 2003, pp. 3957–3962.

[10] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," in *ICSLP*, 2004, pp. 1329–1332.

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

[12] Q. Xu, L. Li, and G. Wang, "Designing engagement-aware agents for multiparty conversations," in *CHI*, 2013, pp. 2233–2242.

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

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

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

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

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

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

[19] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *HRI*, 2010, pp. 375–382.

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

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

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

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

[24] D. Bohus and E. Horvitz, "Learning to predict engagement with a spoken dialog system in open-world settings," in *SIGdial*, 2009, pp. 244–252.

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

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

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

[28] C. Peters, "Direction of attention perception for conversation initiation in virtual environments," in *International Workshop on Intelligent Virtual Agents*, 2005, pp. 215–228.

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

[30] K. Inoue, D. Lala, K. Takanashi, and T. Kawahara, "Latent character model for engagement recognition based on multimodal behaviors," in *IWSDS*, 2018.

[31] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara, "Talking with ERICA, an autonomous android," in *SIGdial*, 2016.

[32] D. F. Glas, T. Minaot, C. T. Ishi, T. Kawahara, and H. Ishiguro, "ERICA: The ERATO intelligent conversational android," in *RO-MAN*, 2016.

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

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

[35] D. Ozkan, K. Sagae, and L. P. Morency, "Latent mixture of discriminative experts for multimodal prediction modeling," in *COLING*, 2010, pp. 860–868.

[36] D. Ozkan and L. P. Morency, "Modeling wisdom of crowds using latent mixture of discriminative experts," in *ACL*, 2011, pp. 335–340.

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

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

[39] D. Lala, K. Inoue, P. Milhorat, and T. Kawahara, "Detection of social signals for recognizing engagement in human-robot interaction," in *AAAI Fall Sympo. Natural Communication for Human-Robot Collaboration*, 2017.

[40] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[41] G. Pundak and T. N. Sainath, "Highway-LSTM and recurrent highway networks for speech recognition," in *INTERSPEECH*, 2017, pp. 1303–1307.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.