INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

HRI-RNN: A User-Robot Dynamics-Oriented RNN for Engagement Decrease Detection

Asma Atamna, Chloé Clavel

LTCI, Télécom Paris, Institut Polytechnique de Paris. Palaiseau, France
firstname.lastname@telecom-paris.fr

Abstract

Natural and fluid human-robot interaction (HRI) systems rely on the robot's ability to accurately assess the user's *engagement* in the interaction. Current HRI systems for engagement analysis, and more broadly emotion recognition, only consider user data while discarding robot data which, in many cases, affects the user state. We present a novel recurrent neural architecture for online detection of user engagement decrease in a spontaneous HRI setting that exploits the robot data. Our architecture models the user as a distinct party in the conversation and uses the robot data as contextual information to help assess engagement. We evaluate our approach on a real-world highly imbalanced data set, where we observe up to 2.13% increase in F1 score compared to a standard gated recurrent unit (GRU).

1. Introduction

The ability to detect user engagement in a conversation is crucial for social robots, as it can be used to adapt the robot's dialogue strategy and, hence, improve the quality of the interaction. In social robotics, engagement can be defined as "the process by which two (or more) participants establish, maintain, and end their perceived connection to one another" [1]. Aside from few attempts in human-human interaction [2] and in human-virtual agent interaction [3, 4] at integrating second-party information in the analysis of user's socio-emotional behavior, current human-robot interaction (HRI) systems rely only on the user data without exploiting the contextual information offered by the robot data, despite the evident link between the robot's behavior and the user's socio-emotional state. We argue that architectures for automatic user engagement analysis can benefit from using the robot data, as well as from learning the interaction dynamics between the user and the robot.

We present HRI-RNN, a novel recurrent neural network (RNN) for on-the-fly detection of user engagement decrease in HRI[1] that explicitly models the user as an individual party in the interaction and exploits the robot data to encode a *context* that we use, with the user state, for prediction (our Python implementation of HRI-RNN is publicly available at github.com/asmaatamna/HRI-RNN). Our model, to the best of our knowledge, is the first one that exploits the robot data while modeling the user-robot interaction dynamics. It is based on the assumption that user engagement is determined by the user's behavior and the current context of the interaction given partly by the robot's behavior, both modeled using recurrent gated units (GRUs) [6] in HRI-RNN. Moreover, our architecture allows to assess two important research questions:

Q1. Does robot data provide additional context information that helps assess user's engagement?

Q2. How does distinguishing user data from robot data compare to treating them indistinctly in a single feature vector?

Our experiments, conducted on a real-world spontaneous HRI data set, show the benefit of using the robot data (speech features in our case) as additional contextual information to detect user's engagement decrease.

This paper is organized as follows: Section 2 discusses related work, Section 3 presents the data set used to evaluate our approach, Section 4 defines the task at hand and introduces our architecture, Section 5 details the experimental procedure, Section 6 discusses the results, and Section 7 concludes the paper.

2. Related Work

2.1. Engagement Analysis in HRI

Advances in deep learning have led to a surge in recurrent neural architectures for automatic engagement analysis [7, 8, 5, 9]. Features commonly used by these approaches to assess engagement include posture, distance to the robot, speech, head motion, gaze, and facial expressions. Studies have shown that combining features from different data modalities leads to a better prediction of user engagement breakdown [10, 9]. In [5], a long short-term memory (LSTM) [11] is used to detect in real-time user engagement decrease in a spontaneous HRI setting. User behavior is modeled by nonverbal features extracted from audiovisual data, and a comparative study investigates the optimal size of the observation window. In [12], a different approach is proposed to recognize engagement in child-robot interaction, where a convolutional neural network is trained directly on facial images extracted from interaction videos. In [8], the authors propose a hierarchical GRU-based architecture for engagement recognition in spoken dialogue that handles the problem of dissimilar data annotations. Engagement is investigated in user's listener mode based on seven manually annotated user features.

2.2. Emotion Recognition in Conversations

Neural architectures that handle conversations by explicitly modeling the different parties involved are very recent. A notable work is presented in [13] for multimodal emotion detection at the utterance level in conversation. The proposed architecture, DialogueRNN, uses three GRUs to model (i) the speaker state, which is maintained for each party, (ii) the context from the preceding utterances, and (iii) the emotion of the preceding utterances respectively, the assumption being that the emotion of an utterance depends on the three aforementioned aspects. In [14], a related architecture is presented with the main difference that the emotion of the preceding utterances is not used to predict the emotion of the current utterance. The work in [14] builds on [15], where a similar architecture with more parameters to estimate is proposed. More recently, [16, 17] have

[1] We refer the reader to [5] for a more in-depth definition of user engagement decrease.

http://dx.doi.org/10.21437/Interspeech.2020-1261

used graph convolutional networks [18] for emotion recognition in conversation to address context propagation issues usually associated with RNNs on long sequences. A conversation is modeled as a graph whose nodes and edges respectively represent utterances and temporal dependencies between them, and emotion recognition is modeled as a node classification task.

2.3. Positioning

Our model is inspired by DialogueRNN [13] which we adapt for a real-time prediction task in a spontaneous HRI setting. However, unlike DialogueRNN that operates at the utterance level where only one party is involved at a time (the speaker), our approach handles settings where data from both the user and the robot are available simultaneously. Additionally, HRI-RNN is *user-centered* in that we model the user state—in speaker and listener modes—and use the robot data only to encode the context of the interaction. More importantly, our model considers user and user-robot dynamics to assess user's engagement—both crucial in engagement analysis [19]. This sets it apart from existing engagement models that do not take into account interparty dynamics. Our work is also related to [5] in that we tackle the same task, adopt the same problem formulation introduced by the authors, and evaluate our model on the same public data set, the UE-HRI [20]. Unlike [5] where the results are reported on a rebalanced test set, however, we evaluate our model on test sets that reflect the real data distribution. As for our data, we make the assumption that nonverbal cues are the most relevant for our task and use audio and video modalities.

3. Data Set

We assess our approach for detecting user engagement decrease on the UE-HRI data set [20]. This data set contains 278 audiovisual recordings of adult users spontaneously interacting with the humanoid robot Pepper (see doc.aldebaran.com/2-5/home_pepper.html). While videos only feature users, audio data corresponds to either the user or the robot. The space in front of the robot is divided into three engagement zones of increasing diameter. The interaction is initiated by the robot whenever a user is detected in the first engagement zone (about 1.5 meters away from the robot), who can then interact and leave whenever they wish. Although the users were warned that only one user should be in the first engagement zone, 69 multiparty interactions were recorded, 32 of which started as multiparty and ended as single-user. The average duration of an interaction is 7 ± 5 minutes.

3.1. Engagement Annotation

Two annotators annotated the data set using ELAN [21]. Videos were annotated segment by segment, according to whether the user shows *signs of engagement decrease* (SED) based on verbal and nonverbal behaviors expressed by the user (1 if SED, 0 otherwise). The overall Cohen's kappa coefficient is $\kappa = 0.73$, reflecting a substantial agreement. In this work, we only use data segments on which the two annotators agree. Note that in most interactions, the user exhibits SED in less than 10% of the interaction. That is, the data classes are heavily imbalanced.

3.2. Feature Extraction

The features we use to detect SED can be grouped in four categories: distance, gaze, head and face, and speech features. The first three categories are extracted from video data that only shows the user. Speech features are extracted from either the user or the robot audio data, depending on the user's mode (speaker or listener) as summarized in Table 1. These multimodal features are extracted from raw audio and video data collected by Pepper as detailed below.

Table 1: *Multimodal features used to detect user's SED per category. Columns 2 & 3 show whether the features characterize the user or the robot depending on the user's mode.*

Feature category	User's mode	
	Speaker	Listener
Distance (front sonar, face distance, head position, engagement zone)	User	User
Gaze (direction, is looking at robot)	User	User
Head & Face (head angles, 17 face AUs)	User	User
Speech (voicing probability, F0 loudness, log-energy, 12 MFCCs, is robot speaking, speech duration)	User	Robot

Distance. User's distance to the robot is measured using Pepper's front sonar. The face distance to the robot camera, as well as the 3D head position with respect to the robot's torso, are extracted using the NAOqi People Perception module of Pepper. User's position with respect to the three engagement zones is determined using the user's 3D head coordinates in the robot frame.

Gaze. User's gaze direction with respect to Pepper's face plane along the (yaw, pitch) axes is extracted with OpenFace [22]. We also use Pepper's ALGazeAnalysis module, which provides information on the user's face orientation, to determine whether the user is looking at the robot.

Head and Face. We use OpenFace [22] to estimate the user's head pose along the (yaw, pitch, roll) axes and to recognize the occurrence and intensity of 17 facial action units (AUs).

Speech. We extract voicing probability, fundamental frequency (F0), loudness, log-energy, and 12 Mel-frequency cepstral coefficients (MFCCs) from one audio channel with openSMILE [23] over 50 ms windows at a frame rate of 100 Hz. We also use a binary feature to represent the speaker (1 for the robot, 0 for the user) and extract speech duration from Pepper's data directly.

To synchronize the feature vectors extracted from different data streams with different sampling frequencies, we perform temporal integration [24] using non-overlapping integration windows of 500 ms, with the mean and variance as integration functions. The resulting synchronized unimodal feature vectors are then concatenated (early fusion) to obtain the final feature vectors. Consequently, the UE-HRI data set can be defined formally as $\{\mathrm{X}_1, \ldots, \mathrm{X}_I\}$, where I is the number of interactions and $\mathrm{X}_i = (\mathrm{x}_i^{(1)}, \ldots, \mathrm{x}_i^{(T_i)})$ is an interaction of length T_i, with $\mathrm{x}_i^{(t)}$ being a feature vector.

4. Methodology

4.1. Problem Definition

Let $\mathrm{X}_i = (\mathrm{x}_i^{(1)}, \ldots, \mathrm{x}_i^{(T_i)})$ be an entire interaction of length T_i between a user and the robot, where $\mathrm{x}_i^{(t)} \in \mathbb{R}^d$ is the feature vector representing the interaction at time step t and d is the

dimension of our data. Our aim is to detect the decrease in user engagement in real time. That is, given a short sequence $(\mathrm{x}_i^{(t-\tau)}, \ldots, \mathrm{x}_i^{(t)})$ of X_i taken over an observation window of length τ, that potentially contains both user and robot speaking turns, we want to detect whether this sequence presents SED from the user. The restriction to a short span is crucial for the online detection setting we aim for.

For the binary classification task thus defined, we consider all sequences of length τ in the UE-HRI data set, $(\mathrm{x}_i^{(t-\tau)}, \ldots, \mathrm{x}_i^{(t)})$, $t = \tau, \ldots, T_i$, $i = 1, \ldots, I$, that we label with 1 (resp. 0) if there are SED (resp. no SED) in the last η seconds of the sequence. Note that the choice of τ and η values depends on the application and the problem design. In this work, we set τ and η to 5 and 2 sec respectively, which corresponds to 10 and 4 time steps respectively.

4.2. Proposed Model

We make the assumption that the perceived user engagement in a sequence $(\mathrm{x}_i^{(t-\tau)}, \ldots, \mathrm{x}_i^{(t)})$ depends on two key aspects:

1. The user's multimodal behavior,
2. The context given by the robot's behavior.

We introduce the Human-Robot Interaction Recurrent Neural Network (HRI-RNN), shown in Fig. 1,[2] that implements these two aspects by maintaining a *user state* that is updated throughout the sequence in order to capture the user state dynamics, which reflect the user engagement level. Additionally, a *context* is computed at each step of the interaction from the robot (audio) data and the previous user state. This context, which captures information that is not provided by the user data, is used jointly with the latter to update the user state. Finally, the last user state is used for *SED detection* (classification). We use GRU cells [6], which have proven to be as effective as LSTMs yet with fewer parameters, to update the user state and the context. The update scheme of a general GRU cell is given by

$$h^{(t)} = \mathrm{GRU}(h^{(t-1)}, z^{(t)}) \ , \tag{1}$$

where $z^{(t)}$ is the current input, $h^{(t-1)}$ and $h^{(t)}$ are the previous and current GRU hidden states respectively, and $h^{(0)} = 0$. Internal update equations are omitted due to space limitations. Classification, on the other hand, is performed using a multilayer perceptron (MLP).

4.2.1. Context

At each step of the interaction, HRI-RNN starts by computing a context using the module GRU_C (see Fig. 1). The aim of this module is to extract additional information from the robot data that may be correlated to the user engagement level and, hence, could help detect SED. Given input $\mathrm{x}_i^{(k)} \in (\mathrm{x}_i^{(t-\tau)}, \ldots, \mathrm{x}_i^{(t)})$, the context is computed by jointly using the robot data, the user state, and the previous context as follows:

$$\mathrm{c}_i^{(k)} = \mathrm{GRU}_C(\mathrm{c}_i^{(k-1)}, \mathrm{x}_{i,R}^{(k)} \oplus \mathrm{u}_i^{(k-1)}) \ , \tag{2}$$

where $\mathrm{c}_i^{(k-1)} \in \mathbb{R}^{d_c}$ is the previous context, d_c is the size of context vectors, $\mathrm{x}_{i,R}^{(k)} \in \mathbb{R}^{d_a}$ is the robot audio extracted from $\mathrm{x}_i^{(k)}$ if the robot is speaking and the zero vector otherwise, d_a is the number of audio features, $\mathrm{u}_i^{(k-1)} \in \mathbb{R}^{d_u}$ is the previous user state, d_u is the size of user state vectors, and $\oplus$ denotes vector concatenation. The initial context vector $\mathrm{c}_i^{(0)}$ is set to zero. The update in Eq. (2) allows to capture temporal dependencies between the robot audio data and the user state and, therefore, models the case where the user's current state depends on the past robot (vocal) behavior.

[2] Our Python implementation of HRI-RNN and the HRI data after feature extraction are available at github.com/asmaatamna/HRI-RNN.

4.2.2. User State

The user state is modeled by the module GRU_U of HRI-RNN (see Fig. 1). It is updated at each step of the interaction by jointly using the user data and the context, along with the previous user state. Formally, for input $\mathrm{x}_i^{(k)} \in (\mathrm{x}_i^{(t-\tau)}, \ldots, \mathrm{x}_i^{(t)})$, the user state $\mathrm{u}_i^{(k)}$ is updated according to

$$\mathrm{u}_i^{(k)} = \mathrm{GRU}_U(\mathrm{u}_i^{(k-1)}, \mathrm{x}_{i,U}^{(k)} \oplus \mathrm{c}_i^{(k)}) \ , \tag{3}$$

where $\mathrm{u}_i^{(k-1)} \in \mathbb{R}^{d_u}$ is the previous user state, $\mathrm{x}_{i,U}^{(k)}$ is $\mathrm{x}_i^{(k)}$ where audio features are replaced with 0 if the user is listening, and $\mathrm{c}_i^{(k)} \in \mathbb{R}^{d_c}$ is the current context. The initial state $\mathrm{u}_i^{(0)}$ is set to the zero vector. The update in Eq. (3) allows to capture user state dynamics while taking into consideration the current context of the interaction.

4.2.3. SED Detection

To classify a sequence $(\mathrm{x}_i^{(t-\tau)}, \ldots, \mathrm{x}_i^{(t)})$ as presenting signs of user engagement decrease or not, we use the last user state, $\mathrm{u}_i^{(t)}$, that we feed to a MLP as shown in Fig. 1. Due to using a GRU cell to model the user state (GRU_U), $\mathrm{u}_i^{(t)}$ contains information on the past states and, consequently, on the entire interaction.

5. Experimental Procedure

5.1. Data Properties

We evaluate HRI-RNN on the UE-HRI data set [20], where we consider sequences of length $\tau = 5$ sec annotated over the last $\eta = 2$ sec following the procedure described in Sec. 4.1. The resulting data set contains 215 658 labeled sequences, and is heavily imbalanced with approx. 90% of sequences labeled 0 (no SED) versus approx. 10% labeled 1 (SED). Robot's audio features are found in around 80% of the sequences, with a an average speaking time per sequence of 40% (resp. 60%) for the robot (resp. the user). The total number of features is $d = 66$ with $d_a = 32$ audio features.

5.2. Architecture of HRI-RNN

We set the sizes of user state vectors $\mathrm{u}_i^{(k)}$ and context vectors $\mathrm{c}_i^{(k)}$ to $d_u = d_c = 32$. For classification, we use a MLP with one hidden layer of 16 neurons defined as follows:

$$\hat{y}_i^{(t)} = \mathrm{sigmoid}\left(\mathrm{W}_o \cdot \mathrm{ReLU}(\mathrm{W}_h \mathrm{u}_i^{(t)} + \mathrm{b}_h) + \mathrm{b}_o\right) \ , \tag{4}$$

where $\mathrm{W}_h \in \mathbb{R}^{16 \times d_u}$ and $\mathrm{b}_h \in \mathbb{R}^{16}$ (resp. $\mathrm{W}_o \in \mathbb{R}^{1 \times 16}$ and $\mathrm{b}_o \in \mathbb{R}$) define the linear mapping and the bias for the hidden (resp. output) layer, and $\hat{y}_i^{(t)} \in [0, 1]$ is the predicted label.

5.3. Baseline

We compare our model to a classical GRU (see Eq. (1)) followed by a MLP with one hidden layer for classification. The aim of this experiment is to answer Q1 and Q2. To this end, we evaluate two GRUs: $\mathrm{GRU}_{\text{user-only}}$ on sequences

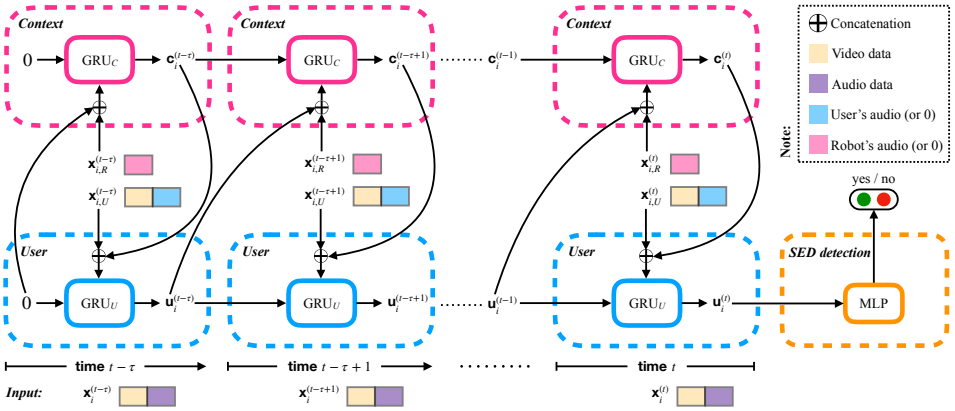

Figure 1: *Architecture of HRI-RNN.*

$(\mathrm{x}_{i,U}^{(t-\tau)}, \ldots, \mathrm{x}_{i,U}^{(t)})$ of user-only data (Q1) and $\mathrm{GRU}_{\text{user-robot}}$ on sequences $(\mathrm{x}_{i}^{(t-\tau)}, \ldots, \mathrm{x}_{i}^{(t)})$ of mixed user-robot data (Q2). Note that $\mathrm{GRU}_{\text{user-robot}}$ is equivalent to an ablation study where we remove the context module GRU_C from HRI-RNN (see Fig. 1). Similarly to HRI-RNN, GRU's last hidden state is used for classification. We set the size of the hidden states to 32.

5.4. Training

We train HRI-RNN and the baseline GRUs using Adam optimizer [25] on batches of size 5000, with weighted binary cross entropy as the loss function, where the weights are inversely proportional to class sizes. This allows to tackle class imbalance at a lesser computational cost than over- or down-sampling. Adam's hyperparameters are set upon experimentation as follows. The models are trained for 50 epochs while the learning rate and the L2 regularization factor (weight decay) are set to 10^{-3}. To estimate the performance, we conduct 10-fold cross validation using 9 folds for training and one fold for testing. We also use 10% of training data as validation set to save the model with the highest F1 score. We create train, validation, and test sets in a user-independent way, i.e. sequences from one particular user are found in only one set at a time.

6. Results

Table 2 summarizes the performance of HRI-RNN, the baseline on user-only data, $\mathrm{GRU}_{\text{user-only}}$, and the baseline on mixed user-robot data, $\mathrm{GRU}_{\text{user-robot}}$. Reported are the mean and standard deviation over 10 runs of the F1 score, which is a more reliable performance indicator than accuracy on imbalanced data sets, the recall and precision, the area under the ROC curve (AUC), and the accuracy for reference. Results are expressed in % and the best performances are highlighted in bold.

HRI-RNN shows the best performance under all metrics with the exception of recall, where $\mathrm{GRU}_{\text{user-only}}$ has the highest recall. HRI-RNN, however, achieves the highest precision and the best compromise between recall and precision, as reflected by the F1 score, where a 2.13% (resp. 1.36%) increase is observed in comparison to $\mathrm{GNN}_{\text{user-only}}$ (resp. $\mathrm{GNN}_{\text{user-robot}}$). These results show that using the robot audio data improves the detection of SED. A possible interpretation is that the extracted audio features—in particular MFCCs—may indirectly encode verbal content information, as well as prosodic features (e.g. pitch, loudness), that reflect robot's socio-emotional behavior influencing user's engagement. By using features that are strongly correlated to user's engagement level, our model has additional key information to efficiently detect SED. Additionally, separating robot data from user data, and using it in a context vector as done in HRI-RNN, leads to a better performance than when user and robot data are fused in the same input feature vector.

Note that following DialogueRNN [13], we experimented with an attention mechanism over the history of context vectors $\mathrm{c}_i^{(k)}$ without observing a significant improvement in HRI-RNN's performance (these results are omitted here for space limitations). Our explanation is that on short input sequences like ours, earlier information propagates well through GRU_C, which manages to compute good enough context representations without the need for an attention mechanism.

Table 2: *Performance (in %) of HRI-RNN and the GRU baseline on user-only data and on mixed user-robot data.*

Model	F1 score	Recall	Precision	AUC	Accuracy
HRI-RNN	**46.04± 4.90**	63.83± 8.34	**36.31± 5.04**	**85.20± 4.08**	**85.86± 2.29**
$\mathrm{GRU}_{\text{user-only}}$	43.91± 4.58	**66.94± 8.28**	32.88± 3.90	84.58± 3.88	83.96± 1.46
$\mathrm{GRU}_{\text{user-robot}}$	44.68± 4.83	66.23± 7.83	33.92± 4.67	85.05± 3.53	84.51± 2.18

7. Conclusion

We presented a recurrent neural architecture for user engagement decrease detection in HRI that distinguishes itself from standard approaches by exploiting the robot's data. Our experiments on a real-world spontaneous HRI data set helped validate the underlying hypotheses to our model, namely (i) using robot's data improves engagement decrease detection since the robot's and user's behaviors are correlated and (ii) treating user's and robot's data separately in order to capture the user-robot interaction dynamics is more beneficial than fusing them in a unique feature vector. Future work directions include investigating longer HRI sequences and context vectors history.

8. Acknowledgments

This work is supported by the Data Science & Artificial Intelligence for Digitalized Industry & Services chair of Télécom Paris, the European project H2020 ANIMATAS (ITN 7659552), and the French National Research Agency's grant ANR-17-MAOI.

9. References

[1] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in Engagement for Humans and Robots," *Artificial Intelligence*, pp. 140–164, 2005.

[2] S. Dermouche and C. Pelachaud, "Engagement modeling in dyadic interaction," in *International Conference on Multimodal Interaction, (ICMI)*, 2019, pp. 440–445.

[3] V. Barrière, C. Clavel, and S. Essid, "Attitude Classification in Adjacency Pairs of a Human-Agent Interaction with Hidden Conditional Random Fields," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4949–4953, 2018.

[4] C. Langlet and C. Clavel, "Improving Social Relationships in Face-to-Face Human-Agent Interactions: When the Agent Wants to Know User's Likes and Dislikes," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 1064–1073.

[5] A. Ben-Youssef, G. Varni, S. Essid, and C. Clavel, "On-the-Fly Detection of User Engagement Decrease in Spontaneous Human-Robot Interaction Using Recurrent and Deep Neural Networks," *International Journal of Social Robotics*, 2019.

[6] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[7] T. Liu and A. Kappas, "Predicting engagement breakdown in HRI using thin-slices of facial expressions," in *Workshops of the AAAI Conference on Artificial Intelligence*, 2018, pp. 37–43.

[8] K. Inoue, D. Lala, K. Takanashi, and T. Kawahara, "Engagement Recognition in Spoken Dialogue via Neural Network by Aggregating Different Annotators' Models," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 616–620.

[9] W. Min, K. Park, J. Wiggins, B. Mott, E. Wiebe, K. E. Boyer, and J. Lester, "Predicting Dialogue Breakdown in Conversational Pedagogical Agents with Multimodal LSTMs," in *Artificial Intelligence in Education*, 2019, pp. 195–200.

[10] A. Ben Youssef, C. Clavel, and S. Essid, "Early Detection of User Engagement Breakdown in Spontaneous Human-Humanoid Interaction," *IEEE Transactions on Affective Computing*, 2019.

[11] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] W. Yun, D. Lee, C. Park, J. Kim, and J. Kim, "Automatic Recognition of Children Engagement from Facial Video using Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, 2018.

[13] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. F. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 6818–6825.

[14] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 2594–2604.

[15] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018, pp. 2122–2132.

[16] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[17] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019, pp. 5415–5421.

[18] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *ICLR*, 2017.

[19] N. Rollet and C. Clavel, ""Talk to you later": Doing Social Robotics with Conversation Analysis. Towards the Development of an Automatic System for the Prediction of Disengagement," *Interaction Studies*, pp. 269–293, 2020.

[20] 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 *Proceedings of the ACM International Conference on Multimodal Interaction*, 2017, pp. 464–472.

[21] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research," in *LREC*, 2006, pp. 1556–1559.

[22] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 59–66.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1459–1462.

[24] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification with Application to Musical Instrument Classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 174–186, 2009.

[25] D. P. Kingma and J. Ba, "ADAM: A Method for Stochastic Optimization," in *ICLR*, 2015.