Personalized Re-Engagement Feedback for Adaptive Socially Assistive Robot Interventions

Author Names Omitted for Anonymous Review.

***Abstract*—Socially assistive robot companions have already shown great potential to augment therapeutic intervention and support behavior change for a broad variety of user populations, including users with special needs (e.d., children with autism spectrum disorders (ASD) and college students with attention-deficit/hyperactivity disorder (ADHD)). However, the ability to fluently perceive user disengagement and re-engage the user effectively in a personalized and closed-loop manner remain open challenges. To address these challenges, this work aims to develop and validate strategies that enable socially assistive robots to perceive disengagement using multimodal audio-visual signals in real time, leveraging pre-trained supervised machine learning models. We formulate *re-engagement* as a multi-arm bandit problem to personalize both the timing and content of the re-engagement feedback given to the user, enabling the agent to learn the user's preferences and to re-engage them more effectively over time. We describe a planned user study involving university students with self-reported attention difficulty engaging in a freestyle writing task, to validate the performance of our approach relative to two non-personalized baseline agents. This work paves the way toward developing personalized socially assistive robots and agents capable of delivering effective and adaptive behavioral interventions for positive behavior change.**[1]

I. Introduction

Socially assistive robot (SAR) companions have shown great potential in effectively promoting behavioral, cognitive, and socio-emotional outcomes [10, 2, 1, 15]. They may have a particularly important role to play in supporting behavior change for users with differences, including users with autism spectrum disorder (ASD) and ADHD. ASD affects 1 in 54 people in the United States alone [13] and negatively impacts their ability to communicate and understand social cues [6], among other symptoms. Behavioral therapists often use toys to create engaging interventions for children with ASD [12]; SAR is inherently engaging and has been shown to be effective for developing social skills in young users with ASD [20]. The US CDC estimates that 13 % of adolescents ages 12 to 17 years are diagnosed with ADHD [8]. Prior work has also shown the effectiveness of applying SAR to provide behavior change support for inattention and impulsive events [3]. Despite these promising results, Jain et al. [9] found that users with special attentional needs can still be easily distracted in the context of human-robot interaction (HRI), so recognizing dis-engagement and re-engagement–which varies across users–is crucial for the effectiveness of SAR interventions.

Past HRI research has focused on *post hoc* multimodal modeling of engagement, and much less on validating such models on real-time interactions. A large body of prior work has shown that audio-visual machine learning models can be applied to *post hoc* model engagement with satisfactory accuracy and has also been validated in recent constrained *post hoc* lab studies [19] and in-the-wild home environments [9]. These studies further highlight the importance of multimodal (audio and visual) fusion in engagement recognition. However, in the small number of past studies investigating closed-loop systems for engagement, only unimodal (visual) machine learning models [21, 14, 7] have been deployed. In this work, we aim to bridge that gap and develop an audio-visual model for real-time user engagement recognition.

An even more limited body of past literature has focused on how to re-engage users. Sun et al. [21] studied two robot re-engagement strategies (explicit and implicit) to remind users to re-focus on the interaction; they found that implicit cues were perceived as more polite and appropriate. In another study, Brown and Howard [4] compared different modes of re-engagement feedback in educational games: 1) verbal, 2) non-verbal, 3) mixture of verbal and non-verbal, and 4) no agent. The study found that verbal and mixture groups outperformed the rest in minimizing boredom. However, there has not yet been work exploring the potential of developing personalized strategies for re-engagement.

This paper aims to make the following contributions:

- Develop an audio-visual model for real-time engagement recognition in the context of education.
- Design and evaluate a re-engagement strategy using multi-arm bandit algorithms that personalizes the feedback timing and content based on user preferences.

II. Methods

A. Audio-Visual Real-Time Engagement Recognition

As shown in Figure 1, to enable the real-time engagement recognition from multimodal (audio-visual) affective signals (facial, body pose and audio features), we will first conduct a data collection to obtain training data. With IRB approval, we will recruit USC students to video-record themselves studying in their dorm rooms. Two trained annotators will label each video frame as engaged/disengaged; we will use Fleiss Kappa [17] to verify inter-rater reliability. We will also incorporate "study with me" (SwM) Youtube videos to increase the dataset. SwM videos are ideally suited because they are filmed in a variety of environments and conducted in "working" and "break" intervals, with the interval type labelled in each frame.

[1]Parts of this idea were introduced at the AAAI AI for Behavior Change Workshop; this is a more expanded and detailed version of the work.

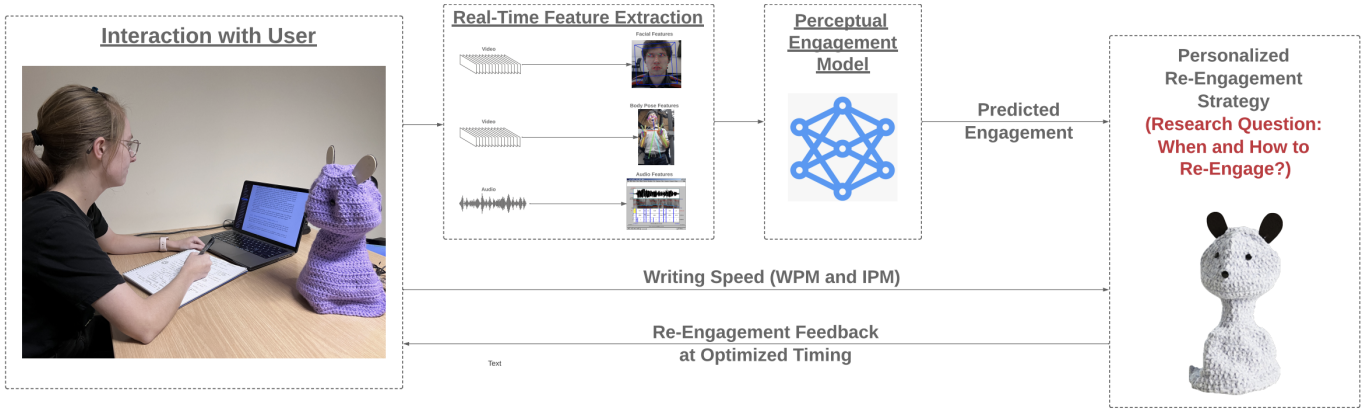

Fig. 1. **Our system setup**

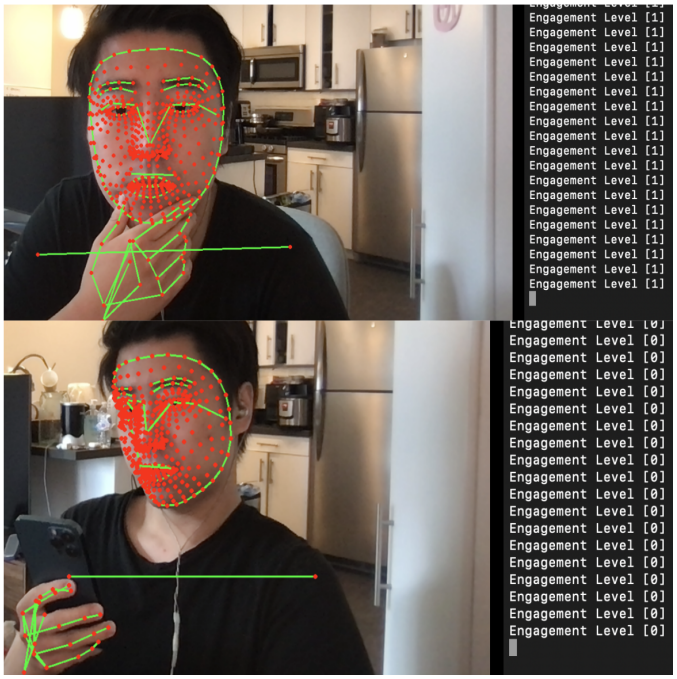

Fig. 2. **Our multimodal real-time engagement modeling pipeline**

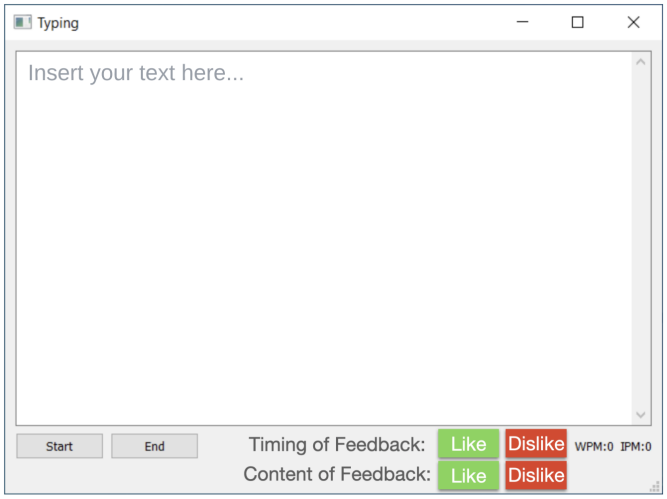

Fig. 3. **Writing interface**

As shown in Figure 2, after the training data are collected and annotated, we will implement the pipeline to enable engagement modeling in real time. We will use the open-source libraries to extract visual and audio features from the video frames, specifically MediaPipe [23] for body pose and facial features and PyAudio [18] for audio features. We will then perform fusion of features from different modalities, pre-process the multimodal data, and feed it into the selected machine learning models. We will perform model selection on different fusion strategies (e.g., early, mid, or late fusion) with different machine learning models (e.g., XGBoost [5], feed-forward neural networks). We will also explore pre-trained end-to-end models (e.g., 3D convolutional neural networks like R2+1D [22]) to learn feature extraction and classification together. We have already implemented a demonstration using MediaPipe [23], PyAudio [18], and XGBoost [5]; the Github codebase for the engagement recognition models will be released and updated in the paper after the double-blind review.

B. User Interface for the Study

We designed a writing user interface, shown in Figure 3, that collects the participants' written input and continually calculates the following information:

1) Words Per Min: $\text{WPM}(t) = \frac{alpha(t) - alpha(t-interval)}{5*(interval/60)}$, calculated to measure the participant's writing speed, where $alpha(t)$ represents all character input at time t second, and $interval$ is the time interval for calculating instantaneous speed (the default is 10 sec). The average English word length is 5 characters, so 5 was used to calculate the word count [16].
2) Input Per Min: $\text{IPM}(t) = \frac{alpha(t) - alpha(t-interval)}{(interval/60)}$, calculated to measure participants' keyboard input speed.
3) User Feedback: As shown in Figure 3, we designed a mechanism for the user to provide feedback, consisting of a panel with four buttons: like or dislike for timing and like or dislike for feedback content. The buttons

are in the lower right corner of the writing UI. After receiving feedback from the robot, users can provide feedback to the robot by selecting one of the buttons for timing and one for content.

C. Definition of Engagement level

Based on the predictions of our audio-visual engagement models, study participants' audio-visual states will be classified as either engaged or disengaged. Participants' writing speed will be measured in real time using WPM/IPM. By combining these two measurements, we can map the participants into one of the following engagement levels: 1) engaged with satisfactory writing speed; 2) engaged with unsatisfactory writing speed; 3) disengaged, but writing speed is satisfactory; 4) disengaged with unsatisfactory writing speed.

D. Personalized Timing and Content of Re-Engagement Feedback

We plan to train one multi-arm bandit (MAB) model for personalizing re-engagement timing, and another MAB model for personalizing the content of the feedback. For the timing of re-engagement, we designed five tentative choices, involving re-engagement after 3 sec, 10 sec, 30 sec, 1 min, and 5 min. For the content of re-engagement feedback, we designed five tentative choices based on the literature [21, 4]: 1) explicit verbal feedback: reminding users about the goal and providing trainer-like comments; 2) implicit verbal feedback: providing encouragement and suggestions; 3) nonverbal audio feedback; 4) nonverbal movement feedback; and 5) no feedback. Each choice of timing and content will be considered as an arm in the MAB models. We also tentatively designed the reward function as:

$$R(t) = \lambda re_t + (1 - \lambda) rh_t$$

where $R(t)$ is the overall combined reward, re_t is environment reward (giving positive reward when user's engagement level improves after the re-engagement feedback), rh_t is user reward feedback for robot action recorded by the users' button pressing, and λ is the weight of two rewards (the default value is 0.5). We will explore different choices of timing and content with various values of λ in a pilot study before finalizing the values. The Github repository link for the code will be released and updated in the paper after the double-blind review. The Github codebase for user interface and re-engagement feedback algorithms will be released and updated in the paper after the double-blind review.

III. Experiment Setup

A. Research Hypotheses

We will investigate the following research hypotheses:

- H1: Users' disengagement behaviors can be robustly recognized by our method.
- H2: Our personalized re-engagement strategy will re-engage users more effectively than the non-personalized baseline strategy.
- H3: Our personalized re-engagement strategy will be preferred by users as more appropriate and a better study companion than the non-personalized baseline.

B. Study Design

We plan to use a within-subject study design to explore our hypotheses and compare our proposed personalized strategy with a non-personalized baseline that that will have constant choice of timing and feedback content based on the literature.

Participants will be invited to the lab twice to work on freestyle writing tasks similar in format to the writing section of the Graduate Record Examination (GRE). In each session participants will be randomly assigned to one of the two conditions (personalized/non-personalized) and asked to write an essay in 50 minutes responding to a given prompt from a pool taken from past GREs. The two sessions will be scheduled on different days to avoid fatigue as a potential confound. Participants will be given a single objective: to write as many words as possible with the goal of practicing their writing skills in this relatively intense task.

To determine the study size, we will conduct a pilot study and follow Kadam and Bhalerao [11] to utilize power analysis to derive the needed sample size. We will use university mailing lists to recruit students who self-report having difficulty with attention, as we have done in past work. Exclusion criteria will include not-fluency in English and auditory and/or visual impairments that would prohibit understanding the study stimuli.

C. Measures and Analysis

All of the study sessions will be recorded and coded by annotators for user engagement level, to test H1. Based on the annotated engagement level and number of words written in both conditions, we can also objectively compare our personalized strategy with the non-personalized baseline, to test H2. In addition, we plan to ask participants to fill out a set of qualitative and quantitative questionnaires so we can understand user-perceived effectiveness, appropriateness, and helpfulness of our proposed method compared to the baseline, to test H3.

IV. Conclusion

This work aims to make the following two main contributions: 1) develop an audio-visual real-time engagement model for educational setting; and 2) design and evaluate a re-engagement feedback strategy using multi-arm bandit algorithms to personalize the feedback timing and content based on user preferences. We hope this work would directly inform progress toward personalized socially assistive robots and agents capable of delivering more effective and adaptive behavioral interventions for positive behavior change.

Acknowledgments

This research was supported by the National Science Foundation National NSF ITE-2236320.

References

[1] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science robotics*, 3(21), 2018.

[2] Roger Bemelmans, Gert Jan Gelderblom, Pieter Jonker, and Luc De Witte. Socially assistive robots in elderly care: a systematic review into effects and effectiveness. *Journal of the American Medical Directors Association*, 13(2):114–120, 2012.

[3] Jonnathan Berrezueta-Guzman, Vladimir Robles-Bykbaev, Iván Pau, Fernando Pesántez-Avilés, and María-Luisa Martín-Ruiz. Robotic technologies in adhd care: Literature review. *IEEE Access*, 2021.

[4] LaVonda Brown and Ayanna M Howard. Engaging children in math education using a socially interactive humanoid robot. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 183–188. IEEE, 2013.

[5] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[6] Mauro Conti, Roberto Di Pietro, Luigi V. Mancini, and Alessandro Mei. (old) distributed data source verification in wireless sensor networks. *Inf. Fusion*, 10(4):342–353, 2009. ISSN 1566-2535. doi: http://dx.doi.org/10.1016/j.inffus.2009.01.002.

[7] Alessandro Di Nuovo, Daniela Conti, Grazia Trubia, Serafino Buono, and Santo Di Nuovo. Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. *Robotics*, 7(2):25, 2018.

[8] Centers for Disease Control, Prevention (CDC, et al. Mental health in the united states. prevalence of diagnosis and medication treatment for attention-deficit/hyperactivity disorder–united states, 2003. *MMWR. Morbidity and mortality weekly report*, 54(34): 842–847, 2005.

[9] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics*, 5(39), 2020.

[10] Katarzyna Kabacińska, Tony J Prescott, and Julie M Robillard. Socially assistive robots as mental health interventions for children: a scoping review. *International Journal of Social Robotics*, 13(5):919–935, 2021.

[11] Prashant Kadam and Supriya Bhalerao. Sample size calculation. *International journal of Ayurveda research*, 1(1):55, 2010.

[12] Connie Kasari, Alexandra Sturm, and Wendy Shih. Smarter approach to personalizing intervention for children with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research*, 61(11):2629–2640, 2018.

[13] Alison Knopf. Autism prevalence increases from 1 in 60 to 1 in 54: Cdc. *The Brown University Child and Adolescent Behavior Letter*, 36(6):4–4, 2020.

[14] Séverin Lemaignan, Fernando Garcia, Alexis Jacq, and Pierre Dillenbourg. From real-time attention assessment to "with-me-ness" in human-robot interaction. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 157–164. Ieee, 2016.

[15] Maja J Matarić and Brian Scassellati. Socially assistive robotics. *Springer handbook of robotics*, pages 1973–1994, 2016.

[16] Mark S Mayzner and Margaret Elizabeth Tresselt. Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic monograph supplements*, 1965.

[17] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[18] Hubert Pham. Pyaudio: Portaudio v19 python bindings. *URL: https://people. csail. mit. edu/hubert/pyaudio*, 2006.

[19] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19), 2018.

[20] Brian Scassellati, Henny Admoni, and Maja Matarić. Robots for use in autism research. *Annual review of biomedical engineering*, 14, 2012.

[21] Mingfei Sun, Zhenjie Zhao, and Xiaojuan Ma. Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 556–567, 2017.

[22] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[23] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.