Engagement Recognition Using Audio Channel Only

DENIS DRESVYANSKIY*, Institute for Communications Engineering, Ulm University, Germany and ITMO University, Russia
INGO SIEGERT*, Mobile Dialog Systems, Institute for Information Technology and Communications, Germany
ALEXEI KARPOV, SPIIRAS, SPC RAS, Russia and ITMO University, Russia
WOLFGANG MINKER, Institute for Communications Engineering, Ulm University, Germany

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: human-computer interaction, paralinguistics, engagement recognition, audio processing

1 INTRODUCTION

Utilizing dialogue assistants endowed with weak artificial intelligence has become a common technology, which is widespread across many industrial spheres - from operating robots using voice to speaking with an intelligent bot by telephone. However, such systems are still far from being essentially intelligent systems, since they cannot fully mimicry or replace humans during human-computer interaction (HCI). Nowadays, paralinguistic analyses is becoming one of the most important parts of HCI, because current requirements to such systems have been increased due to sharped improvement of speech-recognition systems: now, the HCI system should not only recognize, ***what*** the user is talking about, but also ***how*** he/she is talking, and ***which intention/state*** does he/she have now. Those include analyzing and evaluating such high-level features of dialogue as stress, emotions, engagement, and many others.

Although there have been a lot of studies in paralinguistics devoted to recognizing high-level features (such as emotions[1] and stress[17, 25]) using audio cues, there are still almost no insights on how it could work for engagement.

2 WHAT DO WE KNOW ALREADY?

Engagement is a complex phenomenon, which is still not strictly defined in the scientific community. The most common definitions in the context of HCI are stated in the following. Sidner et al. in [24] stated engagement as a "process by which two (or more) participants establish, maintain and end their perceived connection", where "connection" can be expressed in various ways. Poggi [21] characterized engagement by "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". However, the engagement was considered also in terms of qualities of interfaces [22], user experience [20], in the context of social media [12], and many other areas.

From a theoretic point of view, researchers divide the engagement concept into cognitive, emotional, and behavioral components [7]. The cognitive part is mostly expressed by a person's attention to the interlocutor or task to do, while emotional (affective) engagement encompasses the person's emotions and attitudes, which are reflected by the enjoyment of the particular action. Apart from "brain"-related components of engagement, some researchers further point out a behavioral construct of engagement[5, 16], which is strictly conveyed by actions, giving a possibility to measure engagement more objectively.

*Both authors contributed equally to this research.

Whichever point researchers will lean ultimately, engagement recognition using audio-only is becoming a hot topic in HCI, especially in dialogue systems and human-robot interactions. The key point is that people are able to understand, whether an interlocutor is engaged in the conversation or not, yet it is still difficult for all kinds of HCI systems.

3 HOW DO WE STUDY THE PHENOMENON?

Depending on the usage, the present engagement recognition systems are based on either facial or multi-modal features. In e-learning, researchers [18, 26] use mostly facial features, since a participant is usually silent during lecturer's monologue. However, when the participant is involved in the conversation with other persons (or robots), utilization of facial features is not enough, since a complexity of signals interpretation increases: there are more visual occlusions due to high movements amount of other participants and visual activity of the target participant. However, at the same time, we have a richness of the acoustic signals - speech itself and social signals expressed via laughter, fillers, backchannels and many others. In that case, researchers deploy multi-modal systems, which fuse various cues to do a final prediction. Those include postures and gestures [6, 9], gaze activity [11], visual focus of attention [23], and audio features[14, 15]. It should be noted, however, that audio features in this case mean high-level features such as laughter, backchannels, and turn-taking and play just a complementary role for the generation of the final decision.

According to aforementioned use-cases, there are several databases to study engagement: (1) devoted to e-learning [10, 19], (2) acquired during human-robot interaction [3, 13] and (3) represent human-human dyadic conversations [4]. To the best of our knowledge, neither was exploited for engagement recognition using audio-only features.

4 WHAT WE WOULD LIKE TO KNOW?

One of the key problems in all presented databases and in the engagement recognition domain overall is inconsistency in label scale - it differs from paper to paper, including 2-point scale (disengaged, engaged) [15], 4-point scale (very low, low, high, very high engagement) [10, 14] and 5-point scale (disengaged, low engaged, neutral, engaged, highly engaged)[4, 6]. Sometimes researchers use even more fine-grained scales with more classes [2, 9], although it is rare. Utilization of the 5-level scale looks the most attractive since we can neatly adjust the system response to the user. However, it is not clear, whether human annotators are able to distinctly separate the engagement on 5 states. To prove it, a comprehensive perception study is needed. First of all, there is an important question to be answered: ***Is it theoretically possible to set apart engagement on 5 levels***?

The second key problem lies in the domination of video-based systems in engagement recognition. While researchers are actively implementing multi-modal and video systems to capture user engagement in domains related to e-learning and offline conversations, dialogue related technical systems, such as voice assistants endure a lack of analysis of such characteristics just because they are limited to speech analyses. Audio is mostly exploited as complementary information in multi-modal systems for final decision making, and therefore is not used as a standalone signal for engagement recognition. However, there are many use-cases, when a researcher has only audio, yet should predict the user's state such as engagement. There are almost no studies on audio features in this direction. Thus, it comes to the second important question: ***Can we build a system, which is able to recognize interlocutor's engagement using audio-only (speech and linguistic) cues***?

5 WHAT DO WE WANT TO LEARN FROM DIFFERENT DISCIPLINES?

Trying to answer the second question, we need to define suitable audio features, which are able to characterize engagement. Since people are capable to understand the engagement of interlocutors, it can be assumed that there exist a set of features able to effectively express engagement through audio cues. Finding an "optimal" (in terms of the efficiency of the engagement recognition system) set of audio features, which will highly correlate with user engagement state, will allow training a machine learning model to automatize the process of engagement identification. However, just iterating over all available features is not efficient and hardly implementable, as to characterize acoustics a large number of different features can be used [8]. Thus, we need strong theory-based evidence, which can advise the direction of feature search.

The problem of engagement scales should be also solved starting from the theoretical background. Despite empirical studies on different scales and fact that 4-point and 5-point scales can be turned into 2-point or 3-point scales, the choice of scale should be firstly theoretically supported with psychological researches.

6 WHAT DO WE WANT TO TEACH OTHER DISCIPLINES?

The machine learning (ML) approach is widely known among many research areas. In truth, ML and deep learning (DL) have permeated more or less into all research disciplines. Today it is difficult to imagine the processing and analysis of some data acquired during any work without ML. On the other hand, DL is a fast developing domain of ML, earning a new "breath" with developing special frameworks capable to run DL scripts on GPUs. However, the task of an ML engineer is not only to develop algorithms, but also to work with data during the whole development loop: acquisition, analysis, pre-process, and post-process. All these stages require specific skills to be applied and failure in any of them can be cause of turning into "wrong" data, leading to the incorrect decision provided by the chosen algorithm.

In the case of engagement recognition, the implementation of an recognition system will be based on theoretically (and experimentally) proved valuable features. There are numerous techniques we can utilize for data analysis - processing the raw audios with 1D or 2D convolutional neural networks (CNN), engineering new features from raw signal to diverse the set of available features, using linear and non-linear transformations to reduce the dimension of selected features, evaluating the significance of the obtained features and many others.

7 CONCLUSION

It has been a long path to teach computer systems understanding of human speech. Today, such systems are able to some extend. However, still far from maintain the conversation in the way humans do: dialogue systems do not take into account paralinguistics, were we see engagement of the interlocutor (user) as one important aspect. Although the problem of engagement recognition is under consideration in the research community nowadays, the limitation to audio-only cues is still challenging. To eliminate this drawback, we the developments should be guarded by theoretical groundings to be able to define appropriate features and develop an automated engagement recognition system, which will be able to reliably predict user engagement and act appropriately when the user is about to fall into disengagement during a conversation with a speech-based technical system.

ACKNOWLEDGMENTS

The research was supported by the German Federal Ministry of Education and Research project “RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots".

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[2] Roman Bednarik, Shahram Eivazi, and Michal Hradis. 2012. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proc of the 4th Gaze-In'12.* 1–6.

[3] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[4] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350–359.

[5] Hamish Coates. 2007. A model of online and general campus-based student engagement. *Assessment & Evaluation in Higher Education* 32, 2 (2007), 121–141.

[6] Soumia Dermouche and Catherine Pelachaud. 2018. From analysis to modeling of engagement as sequences of multimodal behaviors. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[7] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: conception, theory and measurement. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–39.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the ACM MM-2010*.

[9] Joseph F Grafsgaard, Joseph B Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 42–49.

[10] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885* (2016).

[11] Ryo Ishii, Yukiko I Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–25.

[12] Alejandro Jaimes, Mounia Lalmas, and Yana Volkovich. 2011. First international workshop on social media engagement (SoME 2011). In *ACM SIGIR Forum*, Vol. 45. ACM New York, NY, USA, 56–62.

[13] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. 2013. The vernissage corpus: A conversational human-robot-interaction dataset. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 149–150.

[14] Jaebok Kim, Khiet P Truong, Vicky Charisi, Cristina Zaga, Vanessa Evers, and Mohamed Chetouani. 2016. Multimodal detection of engagement in groups of children using rank learning. In *International Workshop on Human Behavior Understanding*. Springer, 35–48.

[15] Divesh Lala, Koji Inoue, Pierrick Milhorat, and Tatsuya Kawahara. 2017. Detection of social signals for recognizing engagement in human-robot interaction. *arXiv preprint arXiv:1709.10257* (2017).

[16] Xuezhao Lan, Claire Cameron Ponitz, Kevin F Miller, Su Li, Kai Cortina, Michelle Perry, and Ge Fang. 2009. Keeping their attention: Classroom practices associated with behavioral engagement in first grade mathematics classes in China and the United States. *Early Childhood Research Quarterly* 24, 2 (2009), 198–211.

[17] Iulia Lefter, Gertjan J Burghouts, and Léon JM Rothkrantz. 2015. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Transactions on Affective Computing* 7, 2 (2015), 162–175.

[18] Jiacheng Liao, Yan Liang, and Jiahui Pan. 2021. Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence* (2021), 1–13.

[19] Love Mehta, Aamir Mustafa, et al. 2018. Prediction and localization of student engagement in the wild. In *Digital Image Computing: Techniques and Applications (DICTA), 2018 International Conference on, IEEE*.

[20] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.

[21] Isabella Poggi. 2007. *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler.

[22] W Quesenbery. 2003. Dimensions of usability. Content and complexity: Information design in technical communication.

[23] Hanan Salam and Mohamed Chetouani. 2015. Engagement detection based on mutli-party cues for human robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 341–347.

[24] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.

[25] Mariette Soury and Laurence Devillers. 2013. Stress detection from audio on multiple window analysis size in a public speaking task. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 529–533.

[26] Woo-Han Yun, Dongjin Lee, Chankyu Park, Jaehong Kim, and Junmo Kim. 2018. Automatic recognition of children engagement from facial video using convolutional neural networks. *IEEE Transactions on Affective Computing* 11, 4 (2018), 696–707.