Analysis of Engagement and User Experience with a Laughter Responsive Social Robot

Bekir Berker Türker, Zana Buçinca, Engin Erzin, Yücel Yemez, Metin Sezgin

Koç University, Turkey

bturker13,zbucinca16,eerzin,yyemez,mtsezgin@ku.edu.tr

Abstract

We explore the effect of laughter perception and response in terms of engagement in human-robot interaction. We designed two distinct experiments in which the robot has two modes: laughter responsive and laughter non-responsive. In responsive mode, the robot detects laughter using a multimodal real-time laughter detection module and invokes laughter as a backchannel to users accordingly. In non-responsive mode, robot has no utilization of detection, thus provides no feedback. In the experimental design, we use a straightforward question-answer based interaction scenario using a back-projected robot head. We evaluate the interactions with objective and subjective measurements of engagement and user experience.

Index Terms: laughter detection, human-computer interaction, laughter responsive, engagement.

1. Introduction

Engagement is a crucial component of user experience in human-computer interaction. Social agents need to build a bond with humans to retain their attention in conversations. Today, they still suffer to create and maintain the interest of individuals both in short and long time periods of interactions. Thus, understanding engagement and designing engaging agents is a step towards more naturalistic and sophisticated interactions.

With technological advancement, the robots' appearances have become more realistic, they possess more natural text-to-speech engines and can perform a plethora of complex tasks. These developments have contributed to abating the differences between human-robot and human-human interactions. However, they are still not sufficient for replacing human role in conversations with robots. Communication between two people consists of implicit and explicit channels for delivering essential signals to maintain the interaction as long as both parties desire. Agents should also be able to perceive, respond and make use of these signals. In this paper, we concentrate on exploring the effect of laughter and smile as backchannels in human-robot interaction.

We design an interaction scenario involving two people and a back-projected robot head. In this scenario, the robot plays a quiz game with the participants. We conduct two sets of experiments, where the only difference is the mode of the robot - laughter responsive or laughter non-responsive. In the laughter responsive mode, the robot utilizes our real-time multimodal laughter detection module, to perceive laughter, and respond to it with laughter, or smile (if speaking). Whereas, in the laughter non-responsive mode the robot does not respond to laughter by any means.

We evaluate the difference between the two kinds of interactions subjectively and objectively. For subjective evaluation, we use questionnaires to assess the participants experiences. For objective evaluations, we measure the level of engagement of the participants' in both experiments by using the four connection events - directed gaze, mutual facial gaze, adjacency pair and backchannel, as described by Rich et al. [1].

2. Related work

Many of the existing studies on engagement have concentrated on the notion of engagement [2, 3], while the others have primarily focused on its measurement, detection and improvement in HCI. A recent survey summarizes the issues regarding engagement in human-agent interactions and presents an application on engagement improvement in GRETA/VIB platform [4]. Being a thorough survey, this work emphasizes the importance of engagement in HCI and indicates the growing interest of researchers in the field.

Rich et al.'s work on engagement recognition is one of the pioneering studies in this area [1], where the authors propose an engagement model for collaborative interactions between human and computer. They conduct experiments on both human-human and human-robot interactions to have insight and evaluate their approach. Compared to their earlier work [5], they present a shorter list of dialog dynamics, which includes directed gaze, mutual facial gaze, adjacency pairs and backchannels. They refer to these four events as connection events (CE) between user and the robot, and use their timing statistics (min, mean, max of delays) to compare the engagement levels in two distinct scenarios, as well as an additional metric referred to as pace. Pace recapitulates the timing statistics, and it is inversely proportional to the mean time between connection events. The idea is that each CE refreshes the bond between human and robot, and increases the pace metric which is assumed to be proportional to the engagement level.

The backchannel, defined as a connection event, is an important aspect of engagement. As one of the social signals, it is a type of multimodal feedback, defined by Yngve [6] as non-intrusive acoustic and visual signals provided by listener during speaker's turn. Humans, even unconsciously, respond to speaker using facial expressions, nodding, smiling back, using non-verbal vocalizations (mm, uh-huh), or verbal expressions (yes, right), which are all examples of backchannelling. There exist several studies which concentrate on backchannel timing prediction [7, 8, 9, 10], as well as various others addressing evaluation of backchannel timing such as [11, 12].

We specifically consider laughter as a backchannel signal in human-robot interaction. There exist other studies which integrate laughter in HCI and monitor its effect on the user. However, to the best of our knowledge, none of these studies evaluates the impact of a laughter responsive agent in terms of engagement. For example, Niewiadomski et al. experiment with a virtual agent in a simple interaction scenario with no verbal communication [13], where subjects watch funny videos together with a laughter-aware virtual agent which mimics the
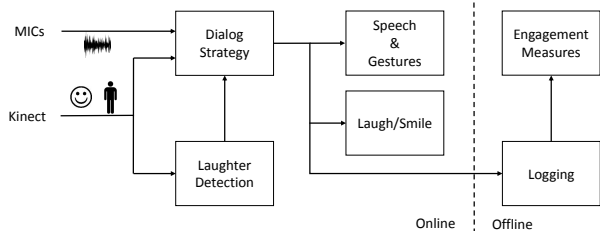
Figure 1: *System overview*

subjects laughter. The humor experience of the subject is evaluated focusing on the quality of the synthesized laughter of the agent, and its aptness in timing. Another similar work is that of [14], where the interaction scenario is also non-verbal. The subject and the virtual agent together listen to some funny music, and the agent mirrors the subject's laughter. The experience is then evaluated utilizing questionnaires. El Haddad et al. experiment with a virtual agent which can predict smile and laughter based on non-verbal expression observations from the speaker [15]. Nonetheless, their focus is on accuracy of the laughter prediction, and the naturalness of the synthesized laughter. They evaluate their system subjectively, using Mean Opinion Score (MOS).

3. System overview

Our main objective is to analyze the role of laughter to engage users in human-robot interaction. Hence the robot should be able to perceive users' laughter and smiles in real-time and to respond back using these non-verbal expressions in its dialog flow. We hypothesize that such an ability of a robot will contribute to the engagement of the user during interactions.

Figure 1 shows the overview of the system. We have a dialog management block [16], which takes user speech from microphones and positions from Kinect as inputs. It then creates a flow of dialog and gestures according to a rule-based strategy. The laughter detection module is involved with the dialog flow to trigger responses (laugh or smile) based on the detection results. All inputs and dialog flow components (produced gestures, speech etc.) are logged, and then processed so as to extract CEs and to compute engagement measures.

We employ a question-answer based scenario in which two subjects participate together and play a game of quiz [16] with the robot. Basically, the robot starts with a short introduction. It asks participants' names and whether they know each other. The robot then proceeds with the quiz and poses some questions which are hard to guess but likely to draw attention from participants. An example question is "What color are sunsets on planet Mars?" and the given options are "green, blue, pink, orange". In order to finalize the quiz and the interaction, the robot keeps score of the correct answers of each participant, and declares the winner as the first to reach 3 points.

We exploit the fact that laughter mimicking by listener is the most natural response to speaker's laughter. Therefore, during the interaction, when the robot is in its laughter responsive mode, it utilizes our laughter detection system for laughter detection. It thereby responds to laughter with laughs while listening, but with smiles while speaking (since it is hard to incorporate naturalistic laughter to speech). Whereas, in its laughter non-responsive mode, the robot does not perform laughter detection, and hence does not respond to laughs.

Figure 2 shows the experimental setup for the interaction scenario. The robot [17, 18] (Furhat in this study) sits on one side of the round table. Participants are placed facing towards the robot. Kinect is on a tripod and able to see both participants' upper-body and face. Individual microphones are attached to participants' collars. Also, one video camera records the whole scene.

IrisTK platform [16] is used with Furhat robot head, which provides functionalities, such as speech recognition and dialog management. On top of these modules, we build a real-time laughter detection module and engagement measurement methods, as we next explain in the sequel.
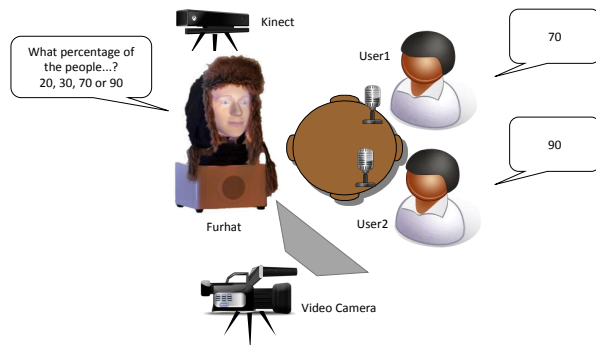

Figure 2: *Experimental setup of the interaction scenario*

4. Real-time laughter detection

We have a multimodal scheme for laughter detection in naturalistic interactions [19, 20]. Audio and facial features are used to feed the detector. We have developed a method of detection on continuous audiovisual streams. It basically creates temporally sliding window on the stream and classifies with SVM whether the window instance involves laughter or not.

The detection method is trained over a human-robot interaction dataset [21], which includes Kinect v2 data recordings. Kinect v2 provides whole body joints and facial landmark points along with high definition video and audio.

4.1. Audio and facial features

We compute 12-dimensional MFCC features using a 25 msec sliding Hamming window at intervals of 10 msec. We also include the log-energy and the first order time derivatives into the feature vector. The resulting 26-dimensional dynamic feature forms the audio features.

Kinect can provide 1347 vertices of face model. We capture only 4 of them, corresponding to lip corners and mid points of lips, which roughly represent the lip shape. We keep these points in 3D coordinates to create a facial feature vector.

4.2. Summarization and classification

Support vector machines (SVM) receive a temporal window of statistical summarization of the short-term features to perform binary classification for laughter and non-laughter classes. We also use probabilistic output of SVM classification for late fusion of modalities and setting different thresholds. The classification task is repeated for every 250 msec over overlapping temporal windows of length 750 msec.

4.3. Real-time implementation

The implementation of laughter detection module is coded in C++ by using Kinectv2 libraries on Microsoft Windows. The process is designed to have two worker threads to handle data acquisition and feature extraction of modalities (audio and facial) and a master thread which fuses the outputs of the worker threads and produces the decision output.

In audio worker thread, the audio stream (16kHz sampling rate) is acquired in chunks of 256 samples. Hence, a two stage sliding window operation is performed in hierarchy. First, MFCC features are extracted over the audio buffer. The extracted MFCC features are then buffered and a sliding classification window is applied.

The video worker thread is similar to the audio thread but with a simpler feature extraction process. Kinect provides each visual frame (body, face, video etc.) with at most 30 fps. However, frames have their special time stamps rather than having fixed sampling period with (1/30) sec. The video thread grabs lip vertices each time a new frame arrives. Feature vectors are buffered where a sliding window runs over in order to have statistical summarization and SVM classification.

5. Engagement measurement

We implement the methods proposed in [1] in order to measure engagement, which is applicable to face-to-face collaborative HCI scenarios.

Rich et al. have defined 4 types of 'connection events' (CEs) as engagement indicators:

- Directed gaze: Sharing the same location for both participants' gaze
- Mutual facial gaze: Face-to-face eye contact event
- Adjacency pair: The minimal overlap or gap between utterances (different speakers') during turn taking
- Backchannel: Backchanneling during other speaker's turn

We mostly follow the same methodology as [1] but with one small modification as we descibe in the following. In [1], there is only one participant interacting with the agent, and the directed gaze event is defined to happen when the agent and the participant look together at a nearby object related to the interaction. However, in our experiments, we have no objects of interest but an additional participant. Hence, when the robot, as a 'connection event initiator', changes its gaze direction from one participant to the other, this action initiates a 'mutual gaze' for one participant and a 'directed gaze' for the other.

In our experiments, CEs are extracted through the logged dialog components and sensory data (from Kinect). The extracted CEs are then used to calculate a summarizing engagement metric called 'mean time between connection events' (MTBCE). MTBCE measures the frequency of successful connection events. Basically, MTBCE in a given time interval T is calculated by $T / (\text{\# of CEs in } T)$. As MTBCE is inversely proportional to engagement, similarly to [1], we use pace = 1/MTBCE to quantify the engagement between a participant and the robot. The pace measure is calculated over a range of different interaction durations such as the first 1 minute, the first 2 minutes and so on.

6. Experimental work and evaluation

In the experiments, we used Furhat [17, 18] as a conversational robot head. Furhat has the advantage of physical existence in

Table 1: *Interaction time statistics of the experiments*

	total #	Interaction Time (sec)		
		min	max	mean
Laughter Resp.	10	138.0	296.6	222.1
Laughter Non-Resp.	10	126.2	515.1	267.8

the scene as well as having ability of efficient facial animation production.

At the beginning of each experiment, participants are briefly informed about the experiment. They are told that they will simply play a quiz game with the robot. The operator explains the roles of the participants and the robot in the game without biasing them. Once participants are ready, they are left alone in an isolated experiment room.

In total, 20 experiments are performed in a randomly selected mode: laughter responsive or laughter non-responsive. A total of 10 experiments are conducted in each of the modes. Each experiment involves two people, therefore the engagement is evaluated over 40 subjects (28 male, 12 female, mean age: 25.9). The experiment ends when one of the participants reaches 3 points in the quiz (3 correct answers). The average time of an experiment is 4 minutes and 5 seconds. Table 1 indicates the statistics of interactions.
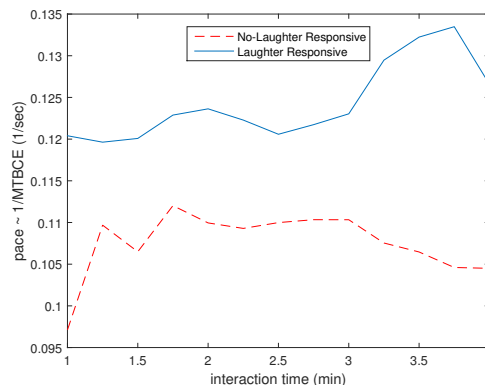

Figure 3: *Average pace values for laughter-responsive (blue) and laughter non-responsive modes (red) over increasing interaction durations.*

Figure 3 shows the average pace of the connection events over subjects in laughter responsive and laughter non-responsive mode of the robot. For example, the pace in the n-th minute is the average pace realized in the period from the beginning to the n-th minute of the interactions. We have calculated the pace for the first 4 minutes, as it is approximately the average duration of an interaction. We observe that the calculated pace values are significantly different between two modes for all interaction durations, indicating a considerable increase in engagement of a participant when interacting with a laughter responsive agent. We note that the pace samples belonging to the two modes exhibit statistically significant differences ($p << 6e - 10$) when 2-sample t-test is applied. Also, the difference between the pace curves starts increasing after the 3-rd minute. This may be due to the observation that, after a warm-up period, participants tend to lose or increase their engagement according to the experiment mode.

For subjective evaluation, the subjects were required to fill

Table 2: *Questionnaire items and mean scores over the laughter responsive and laughter non-responsive modes. Score scale: Strongly Agree (2), Agree (1), Undecided (0), Disagree (-1), Strongly Disagree (-2)*

#	Questionnaire Items	Responsive		Non-Responsive	
		Mean	Std	Mean	Std
1	I liked the interaction with the robot.	1.50	0.60	1.40	0.60
2	The interaction was entertaining.	1.65	0.59	1.45	0.76
3	I felt boredom at times during the interaction.	-1.35	0.59	-1.00	1.03
4	The robot was responsive to my emotional mood.	**0.65**	0.75	**-0.35**	0.67
5	The interaction felt natural.	0.65	0.93	0.20	0.89

a questionnaire after their interaction. Table 2 shows the five questions of the survey. We use a 5-point likert scale: Strongly Agree (2), Agree (1), Undecided (0), Disagree (-1), Strongly Disagree (-2) for each of the questions. To keep the subjects unbiased, even in the questionnaire, the fourth question implicitly asks if the users were aware of the robot's laughter response. The Mann Whitney test for the fourth question, gives a statistically significant ($p = .0002$) difference between the laughter non-responsive and laughter responsive samples, which indicates that the users were aware of the laughter responsiveness of the robot during interaction. Question 5 also gives a statistically significant ($p = .05$) difference between the two samples, an evidence that laughter integration in HCI makes the interaction more naturalistic. The answers to other questions were not statistically different amongst the two samples. Nonetheless, this is expected because these questions are not intended to discern between the two modes of the robot, but rather to get feedback about the interaction scenario. Furthermore, since for most of the participants interacting with the robot was a first-time experience, they underwent the novelty effect. Simply put, even without laughter feedback from the robot, they enjoyed the interaction due to its novelty. Consequently, there is no statistically significant distinction for the first three enjoyment measuring questions.

Figure 4 plots the pace metric for each CE separately. All the CEs, except the mutual gaze (Figure 4b), yield higher pace curves for the robot's responsive mode. In the interactions, we observe the main reason behind the mutual gaze event loss: We discover that the majority of the participants, when amused, look whether the other participant is entertained, as well. In our scenario, this especially occurs when they are told their answer was wrong. In these occasions, the robot immediately shifts the attention from the current participant to require an answer from the other participant. Nonetheless, its initiation of mutual gaze event results in failure because the two participants are looking at each other.

Two strong tendencies of the responsive mode are observed in backchannel and directed gaze CEs. Figure 4d shows increase in the number of backchannel events, which are mostly laughter and smiles, in the second half of the interactions. Pace curves for the directed gaze events yield a decreasing trend for both modes, but responsive mode sustains higher pace values. This trend could be due to the participants' experience with Furhat. At the beginning of the experiment, participants are amazed when Furhat shifts his attention from one to the other participant by head and eye movement. Hence, participants tend to have successful directed gaze events by looking at the other participant when Furhat does so. However, participants get acquainted with these attention shifts with time, which might be the cause of the decreasing trend of the pace for directed gaze events.
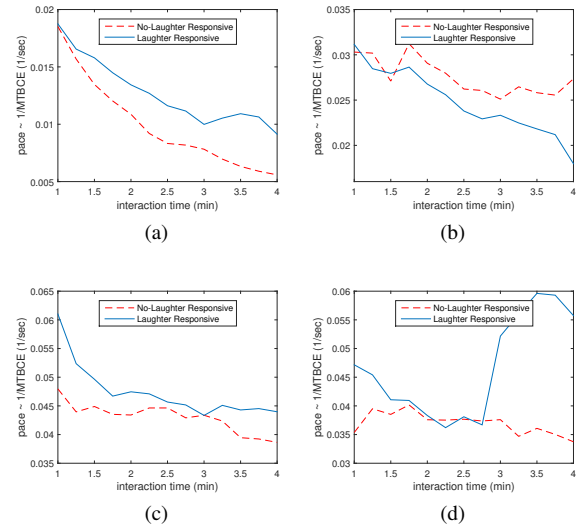

Figure 4: *Average pace values using individual CEs: (a) Directed Gaze, (b) Mutual Gaze, (c) Adjacency Pair, (d) Backchannel*

7. Conclusion

In this paper, we evaluated the effect of laughter in terms of engagement and user experience in human-robot interaction. In an interaction scenario with two people and a back-projected robot head, we experimented with two modes of the robot, laughter responsive and laughter non-responsive. In the laughter responsive mode, the robot responds to subjects' laughter by laughter or smile, whereas in laughter non-responsive mode the robot does not respond to any laughter at all. We measure the engagement of the participants in two sets of experiments, objectively by utilizing the four connection events directed gaze, mutual gaze, adjacency pair and backchannel. Our results indicate that the laughter responsiveness of the robot contributes to engagement of the participants. We also evaluate the user experience with a questionnaire, which likewise shows promising effects of laughter integration in an HCI system.

8. Acknowledgements

9. References

[1] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2010, pp. 375–382.

[2] N. Glas and C. Pelachaud, "Definitions of engagement in human-agent interaction," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, Sept 2015, pp. 944–949.

[3] C. Peters, G. Castellano, and S. de Freitas, "An exploration of user engagement in hci," in *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, ser. AFFINE '09. New York, NY, USA: ACM, 2009, pp. 9:1–9:3. [Online]. Available: http://doi.acm.org/10.1145/1655260.1655269

[4] C. Clavel, A. Cafaro, S. Campano, and C. Pelachaud, *Fostering User Engagement in Face-to-Face Human-Agent Interactions: A Survey*. Cham: Springer International Publishing, 2016, pp. 93–120.

[5] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artif. Intell.*, vol. 166, no. 1-2, pp. 140–164, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2005.03.005

[6] V. H. Yngve, "On getting a word in edgewise," in *Chicago Linguistics Society, 6th Meeting*, 1970, pp. 567–578.

[7] K. P. Truong, R. Poppe, and D. Heylen, "A rule-based backchannel prediction model using pitch and pause information," 2010.

[8] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10458-009-9092-y

[9] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic *et al.*, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.

[10] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Computer Speech & Language*, vol. 28, no. 4, pp. 903–922, 2014.

[11] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth, "Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 181–188. [Online]. Available: http://doi.acm.org/10.1145/2522848.2522890

[12] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, *Creating Rapport with Virtual Agents*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 125–138.

[13] R. Niewiadomski, J. Hofmann, J. Urbain, T. Platt, J. Wagner, B. Piot, H. Cakmak, S. Pammi, T. Baur, S. Dupont, M. Geist, F. Lingenfelser, G. McKeown, O. Pietquin, and W. Ruch, "Laugh-aware virtual agent and its impact on user amusement," in *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, ser. AAMAS '13. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 619–626. [Online]. Available: http://dl.acm.org/citation.cfm?id=2484920.2485018

[14] F. Pecune, M. Mancini, B. Biancardi, G. Varni, Y. Ding, C. Pelachaud, G. Volpe, and A. Camurri, "Laughing with a virtual agent," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '15. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 1817–1818. [Online]. Available: http://dl.acm.org/citation.cfm?id=2772879.2773452

[15] K. El Haddad, H. Çakmak, E. Gilmartin, S. Dupont, and T. Dutoit, "Towards a listening agent: a system generating audiovisual laughs and smiles to show interest," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 248–255.

[16] G. Skantze and S. Al Moubayed, "Iristk: A statechart-based toolkit for multi-party face-to-face interaction," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 69–76. [Online]. Available: http://doi.acm.org/10.1145/2388676.2388698

[17] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, *Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 114–130.

[18] S. Al Moubayed, J. Beskow, and G. Skantze, "The furhat social companion talking head," in *Interspeech 2013, 14th Annual Conference of the International Speech Communication Association, August 25-29, 2013, Lyon, France*, 2013, pp. 747–749.

[19] B. B. Turker, S. Marzban, M. T. Sezgin, Y. Yemez, and E. Erzin, "Affect burst detection using multi-modal cues," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, May 2015, pp. 1006–1009.

[20] B. B. Turker, Z. Bucinca, E. Erzin, Y. Yemez, and M. T. Sezgin, "Real-time audiovisual laughter detection," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, May 2017.

[21] L. Devillers, S. Rosset, G. D. Duplessis, M. A. Sehili, L. Bchade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. Turker, M. Sezgin, K. E. Haddad, S. Dupont, D. Luzzati, Y. Esteve, E. Gilmartin, and N. Campbell, "Multimodal data collection of human-robot humorous interactions in the joker project," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, Sept 2015, pp. 348–354.