2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

A Dyadic Conversation Dataset On Moral Emotions

Louise Heron*
University of Bath
Bath, UK
L.Heron@bath.ac.uk

Jaebok Kim*
Human Media Interaction group
University of Twente
Enschede, Netherlands
j.kim@utwente.nl

Minha Lee*
Human-Technology Interaction group
Technical University of Eindhoven
Eindhoven, Netherlands
M.Lee@tue.nl

Kevin El Haddad*
University of Mons
Mons, Belgium
kevin.elhaddad@umons.ac.be

Stéphane Dupont
University of Mons
Mons, Belgium
stephane.dupont@umons.ac.be

Thierry Dutoit
University of Mons
Mons, Belgium
thierry.dutoit@umons.ac.be

Khiet Truong
Human Media Interaction group
University of Twente
Enschede, Netherlands
k.p.truong@utwente.nl

***Abstract*—In this paper, we present a dyadic conversation dataset involving topics related to moral emotions which are ethically relevant. To the best of our knowledge, it is the first dataset where the main focus is moral emotions. This dataset also focuses on speaker-listener reactions during a dyadic conversation. Although some of the currently available datasets contain dyadic conversations, they were not conceived with the idea of focusing on the speaker-listener setup. Thus making it difficult to use them to study reactions related to speakers and listeners. Some preliminary analyses of the data are presented as well as our thoughts on future work related to this dataset.**

***Keywords*-Moral emotions; Dataset; Dyadic interaction; Non-verbal expressions; Multimodal data; Affective computing**

I. Introduction

Human-agent interactions (HAI) systems are emerging in our daily lives. It is important that these agents learn not only our verbal, but also nonverbal way of communicating. This would allow them to better "understand" underlying messages and make it more comfortable to interact with them since they would have more human-like behavior. In this paper, we present a corpus of spontaneous dyadic conversations with a rich amount of verbal and nonverbal expressions. Our contributions are two-fold:

1) focusing on listener-speaker reactions during a dyadic conversation
2) obtaining data related to moral emotions

Indeed, it is important for the agent to have the proper attitude as a speaker, but also as an attentive listener. The current dyadic interaction datasets such as: the Cardiff Conversation Database (CCDB), IFA Dialog Video corpus (IFADV) [1] and IEMOCAP [2] do not explicitly focus on the separation between speaker and listener. Although the roles of "speaker" and "listener" are implicitly present in a conversation, it is harder to draw a clear line between "speakers" and "listeners", since a "listener" could utter verbal or nonverbal sounds as feedback to the "speaker" and the speaker could be interrupted verbally by the "listener". This is why we decided to build our recording scenario by assigning clearly the "speaker" and "listener" roles to the subjects while keeping the interaction naturalistic.

The second purpose of the dataset was to look into moral emotions, which are emotions that are ethically relevant [3]. The emotional databases currently available contain categorical annotations of mainly the six basic emotions [4], or with codes largely based on the Circumplex Model of emotions [5]. To add diversity to emotion research, we focused on moral emotions. The eNTERFACE workshop [1] gave us the opportunity to collect our own dyadic conversation database [6]. To build this database, participants took turns as speakers and listeners, the latter asking questions to the former about memories of emotional states. Questions were on negative (i.e. guilt and shame) and positive (i.e. pride and compassion) emotions, Which are considered to be moral emotions [3]. In the following subsections, the data recording setup will be described in Section II. We will then go through the dataset content in Section III and our first analyses will be presented in Section IV. We will finally present conclusions in Section V.

II. Setup and Experience Description

A. Procedure

Participants were firstly asked to read the informed consent form, which stated that their participation was voluntary and unpaid. The consent form stated that moral emotions will be discussed by the participants. At the start of the experiment, participants were told that they will randomly be assigned to the role of speaker or listener, then switch roles.

*Authors with equal contribution

[1]eNTERFACE 2017 took place at the Centre of Digital Creativity (CCD), Escola das Artes, Universidade Catolica Portuguesa, in Porto, Portugal. More information is available at http://artes.ucp.pt/enterface17/.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00108

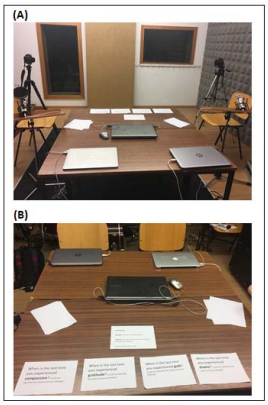

Figure 1. Pictorial depiction of the experimental setup from (A) side and (B) aerial views.

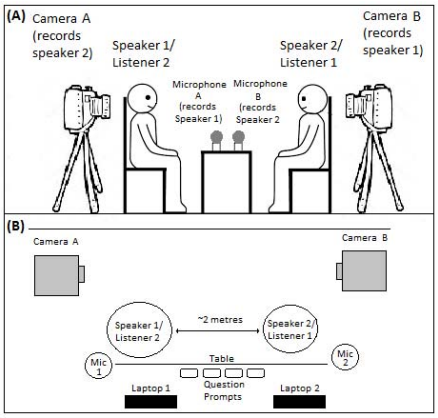

Figure 2. (A) A side on view of the experimental setup from a third party observer and (B) the setup from the participants perspective.

The speaker answered questions about moral emotions, whilst the listener listened to the speakers answers and asked any follow up questions as necessary. The instructions were purposively vague to ensure that the dyadic interaction was as natural as possible. Indeed, the listener was free to discuss and/or interrupt the speaker just like in any normal conversation. The order in which participants discussed the emotions was randomised: listeners chose one of the two moral emotion options (positive, negative) from question prompts on a table. Each interaction started with the listener asking a question that varied in the emotion category: When was the last time you experienced gratitude/compassion/guilt/shame? Can you describe the event and your feelings? The speaker responded to each question. The interaction lasted until the interlocutors both indicated to experimenters that the conversation was finished.

B. Experimental Setup: Video/Audio Acquisition

Video and audio were recorded in a soundproof room at the Catholic University of Porto, Portugal. Two Canon Cameras: EOS 550D and EOS 6D were used to record the interactions. Camera A (beside Speaker 1, recorded Listener 1/Speaker 2) and Camera B (beside Speaker 2, recorded Speaker 1/Listener 2). The camera angle and distance were tailored to each participant, ensuring that the head to torso area was captured. The distance between speaker and listener remained constant. Two Rode Podcaster USB microphones on pop shield shock mounts recorded speaker and listener audio. Laptops were attached to microphones for audio recordings and were also used for pre- and post- experimental questionnaires (see Figures 15 and 16 for diagrams of the experimental setup). Two experimenters were in the room and started and stopped video and audio recordings.

C. Data Post-processing

The participants were asked to clap just before each recording session. The claps were then used to synchronise the videos and corresponding audio files as well as the interlocutors' data with each other. So far, most of the data has been annotated and separated in topics (gratitude, compassion, guilt and shame) and with respect to the role of the subject visible in the video (speaker or listener). We are currently working on segmenting and annotating the dataset in more detail. Indeed, as mentioned earlier, although the roles were split into speakers and listeners, we did not prevent the listener from responding. Thus, segmenting the dataset into speakers and roles in each of the separated audio/video files will be our next goal. Other annotations will be added in the near future to help us study these interactions better. First, non-linguistic nonverbal conversation expressions such as laughs, smiles, eyebrow movements and head movements will be annotated. This will help us model the nonverbal interactions between the interlocutors. Then, the dialogs will be transcribed in order to obtain the semantic content related to the topics of moral emotions.

III. CONTENT

This database was designed to mimic, as much as possible, real world conversations by being unscripted and containing a mixture of nationalities and familiar/unfamiliar individuals. We also took care to note whether it was the participants' first encounter or not. This last point is an important point to note since it affects the interaction [7].

A. Database Demographics

As summarized in Figure 17, the database contains 42 participants (21 pairs), 32 males and 10 females. Participants were mainly students and professors (age range: 20 - 48). There were 14 male-male pairs, 3 female-female pairs and 4 male-female pairs, of which 11 pairs knew each other beforehand. The database comprises of 14 nationalities.

Figure 3. An example of captured video frame from the dataset

B. Questionnaire Data

Demographic information (e.g., age, sex) was collected. Further, to obtain a richer dataset, mood and personality traits were also measured via the questionnaire. To be specific, participants completed pre- and post- experiment mood evaluation by filling in the Positive and Negative Affect schedule [8]. Then after the experiment, they completed additional questions that measure stress and personality traits: Stress Response Scale [9] and the Big Five Inventory [10].

IV. ANALYSIS

Preliminary analyses were undertaken on the part of the dataset currently annotated into topics and roles as mentioned in SectionII-C which corresponds to 13 dialogues (26 participants). Since the database is not fully segmented and annotated yet, these results presented here are only preliminary and will serve as the basis of future analysis and modeling work. Two main modalities were analysed, the audio and visual.

A. Visual Cue

The OpenFace tool [11] was used to extract Action Units (AUs) which served to analyse the facial expressions from each interlocutor by conversation topic. In order to eliminate outliers, we only considered the data corresponding to the frames for which the coordinates of the tip of the nose were within a margin corresponding to the majority of the subjects (these coordinates were also given by OpenFace). OpenFace extracts a set of features per video frame.

In some cases, facial expressions corresponding to the six basic emotions [4] can be related to combinations of AUs. In this work, we considered the ones representing happiness/joy (AU06 + AU12) and sadness (AU01+AU04+AU15). Tables I and II give the proportion of the number of frames containing the combinations of happiness and sadness respectively in the total number of frames. These tables also show the results by participant role (speaker, listener, or both) and conversation topic (shame, guilt, compassion and gratitude).

From the results in Table I, we can see that the proportion of happiness is higher in the "shame" and "gratitude" topics than in the "compassion" and "guilt" ones. In contrast, we can see that the proportion of sadness is higher for "compassion" and "guilt" compared to "shame" and "gratitude". To analyse them further, we first point out that the AUs constituting this combination correspond to smiling. We thus emit the hypothesis that the topic related to "shame" would be more likely to contain funny stories than "guilt" and "compassion". Additionally, "happy" stories would be more likely to occur in the "gratitude" topic than in the "guilt" and "compassion" one. Finally, "guilt" would be more likely to contain stories related to mistakes one made and the "compassion" topic stories related to pity. It would therefore be understandable that the ratio is higher for the former two topics than the latter two. These hypotheses were also based on an informal inspection on the semantic content of the database. However, in order to understand these phenomena, further work should be oriented towards the semantic analysis of the dialogues.

	Shame	**Guilt**	**Compassion**	**Gratitude**
Listener	23%	17%	21%	24%
Speaker	26%	24%	15%	21%
Both	24%	20%	18%	22%

Table I
PROPORTION OF THE NUMBER OF FRAMES CONTAINING AU COMBINATIONS FOR **HAPPINESS/JOY** IN THE TOTAL NUMBER OF FRAMES

	Shame	**Guilt**	**Compassion**	**Gratitude**
Listener	2.2%	3.1%	2.6%	3.5%
Speaker	2.0%	1.5%	1.7%	1.9%
Both	2.2%	2.5%	2.2%	2.1%

Table II
PROPORTION OF THE NUMBER OF FRAMES CONTAINING AU COMBINATIONS FOR **SADNESS** IN THE TOTAL NUMBER OF FRAMES

B. Audio Cue

Since we have not yet completed the speaker segmentation of our corpus, we focus on low level descriptors, pitch and energy that are associated with emotion in speech. We used the OpenSMILE toolkit for the feature extraction [12]. In addition, a Voice Activity Detection (VAD)[2] was used to keep only the speech parts of the conversations. Only these parts were considered for the analysis. Since pitch is voice specific, it was studied for the same interlocutors only. Indeed, we made the assumption that the files containing the "speakers" data contained in major parts, the speaker's voice. We thus compared, for each speaker, the median values of the pitch with respect to the topic. This showed different variations for each speaker and no consistent pattern could be found. The overall RMS energy was then compared per

[2]https://github.com/wiseman/py-webrtcvad

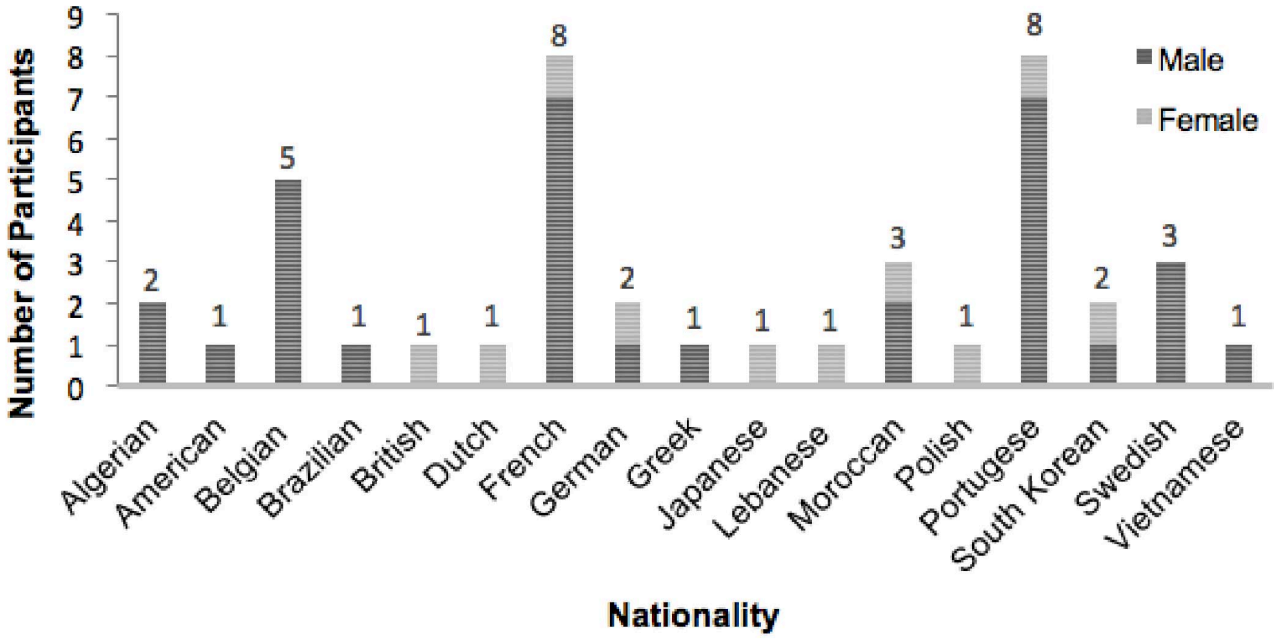

Figure 4. Participant age and nationality distribution in the database

topic and per speaker. No significant difference was found but, for the same speaker, the mean value of speech energy of the positive topics (compassion and gratitude) was roughly equal to the negative topics (shame and guilt).

In future work, after the dataset is segmented by speaker, a more detailed analysis will be undertaken. Additionally, we will analyse nonverbal audio expressions such as laughs and backchannels and speech rate.

C. Future Work on the Linguistic Cue

Initial investigations were carried out on the linguistic cues and therefore on the semantic content of the data. As we have not manually transcribed our data yet, an automatic speech recognition system (ASR) was tested on it. To do so, each speech segment detected by the VAD was sent to the Google Speech Recognition system of the SpeechRecognition python library[3]. The text output was meant to be sent to a sentiment and emotion classification system. Unfortunately, a large part of the data could not be properly recognised by the ASR mainly due to speaker pronunciation (most of the speakers are non-native English speakers). In future work, a manual transcription will be made to help us provide a semantic analysis content of the data.

[3]https://pypi.python.org/pypi/SpeechRecognition/

V. CONCLUSION AND FUTURE WORK

In conclusion, we present the dyadic conversation database on moral emotions. This could be of potential interest to researchers in social psychology, computer vision, affective computing, and computational linguistics, to name a few fields. A more thorough look at dyadic conversations in naturalistic settings, as well as on moral emotions, could be done in the future. Another suggestion is a more extensive dataset that builds on the current version. Dissecting various elements of dyadic conversations on a subject matter of emotions will require a multi-disciplinary approach, a step we hope to continue in.

REFERENCES

[1] R. van Son, W. Wesseling, E. Sanders, H. van den Heuvel *et al.*, "The ifadv corpus: a free dialog video corpus." in *LREC*, 2008, pp. 501–508.

[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[3] J. Haidt, "The moral emotions. handbook of affective sciences," 2003.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[6] K. E. Haddad, Y. Rizk, L. Heron, N. Hajj, Y. Zhao, J. Kim, T. N. Trong, M. Lee, M. Doumit, P. Lin, Y. Kim, and H. Cakmak, *End-to-End Listening Agent for Audiovisual Emotional and Naturalistic Interactions*, L. T. Jorge Cardoso, Andre Perrotta, Ed. Porto, Portugale: Journal of Science and Technology of the Arts (CITAR), no. 17-02.

[7] P. Paggio, J. Allwood, E. Ahlsén, and K. Jokinen, "The nomco multimodal nordic resource–goals and characteristics," 2010.

[8] D. Watson and L. A. Clark, "The panas-x: Manual for the positive and negative affect schedule-expanded form," 1999.

[9] S. Suzuki, H. Shimada, Y. Sakano, I. Fukui, and M. Hasegawa, "Stress response scale-18," *Kokoronet, Jul*, 2007.

[10] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[11] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[12] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224