K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations

Cheul Young Park*, **Narae Cha, Soowon Kang, Auk Kim, Uichin Lee***
Graduate School of Knowledge Service Engineering, KAIST
{cheulyop, nr.cha, sw.kang, kimauk, uclee}@kaist.ac.kr

Ahsan Habib Khandoker, Leontios Hadjileontiadis
Department of Biomedical Engineering, Khalifa University
{ahsan.khandoker, leontios.hadjileontiadis}@ku.ac.ae

Alice Oh
Department of Computer Science, KAIST
alice.oh@kaist.ac.kr

Yong Jeong
Department of Bio and Brain Engineering, KAIST
yong@kaist.ac.kr

Abstract

Recognizing emotions during social interactions has many potential applications with the popularization of low-cost mobile sensors, but a challenge remains with the lack of naturalistic affective interaction data. Most existing emotion datasets do not support studying idiosyncratic emotions arising in the wild as they were collected in constrained environments. Therefore, studying emotions in the context of social interactions requires a novel dataset, and K-EmoCon is such a multimodal dataset with comprehensive annotations of continuous emotions during naturalistic conversations. The dataset contains multimodal measurements, including audiovisual recordings, EEG, and peripheral physiological signals, acquired with off-the-shelf devices from 16 sessions of approximately 10-minute long paired debates on a social issue. Distinct from previous datasets, it includes emotion annotations from all three available perspectives: self, debate partner, and external observers. Raters annotated emotional displays at intervals of every 5 seconds while viewing the debate footage, in terms of arousal-valence and 18 additional categorical emotions. The resulting K-EmoCon is the first publicly available emotion dataset accommodating the multiperspective assessment of emotions during social interactions.

1 Background & Summary

Emotion recognition research seeks to enable computers to identify emotions. It is a foundation for creating machines capable of understanding emotions, and possibly, even expressing one. Such a set of skills to recognize, understand, and express emotions form emotional intelligence [1, 2]. It is suggested that emotional intelligence is necessary for the navigation of oneself within a society, as it allows one to reason what is desirable and what is not, and to regulate behaviors of self and others accordingly [3, 4].

Then why do machines need emotional skills? With advances in Machine Learning and Artificial Intelligence, the transition from human to machine is noticeable in all areas of the society, including those requiring expertise such as medical prognosis/diagnosis [5, 6] or automobile driving [7]. It seems inevitable that these narrow AI

*Correspondence and requests for materials should be addressed to C.P. (email: cheulyop@kaist.ac.kr) or U.L. (email: uclee@kaist.edu)

systems [8] supersede human experts in respective domains, as it has already been demonstrated with AlphaGo's superior performance in the game of Go over human champions [9, 10].

Not all AI will compete with humans, albeit their superhuman ability. Instead, many AI systems will work with us or for us. Emotional intelligence is critical for such human-computer interaction systems [11]. Imagine a smart speaker that delightfully greets users when they come home. How should a speaker greet when a user had a rough day? A speaker neglectful of the user's emotional states may aggravate the user, but a speaker aware of the user's temper could remain silent to avoid the trouble. Similarly, emotional intelligence is critical for AI systems designed for complex tasks. For example, on roads where autonomous and human-driven vehicles mix, accurate recognition of emotions of human drivers' by autonomous vehicles would lead to more safety as autonomous vehicles can better judge human drivers' intentions [12].

Now for machines to become emotionally intelligent, they must first learn to recognize emotions, and the prerequisite to learning is data. However, there lie several challenges in the acquisition of emotion data. While emotions are prevalent, their accurate measurement is difficult. Most commonly, emotions are viewed as psychological states expressed through faces, with distinct categories [13], but research evidence claims the contrary. Rather than distinct, facial expressions are compound [14], relative [15], and misleading [16]. A recent review of scientific evidence also presses against the common view, suggesting that facial expressions lack reliability, specificity, and generalizability [17], together with past studies on contextual dependency [18, 19, 20] and individual variability of emotions [21, 22].

Such inherent elusiveness of emotion renders many existing emotion datasets inapplicable for studying emotions in the wild. The majority of emotion datasets consist of emotions induced with selected stimuli in a static environment, i.e., a laboratory [23, 24, 25, 26, 27, 28, 29]. This method provides experimenters with full-control over data collection, allowing assessment of specific emotional behaviors [30, 31] and acquiring fine-grained data with advanced techniques like neuroimaging. Nevertheless, lab-generated data may generalize poorly to realistic scenarios as they frequently contain intense expressions of prototypical emotions, which are rarely observed in the real world [32, 33], acquired from only a subset of the population [34].

An alternative approach utilizes media contents [35, 36, 37, 38] and crowdsourcing [39], compensating for the shortcomings of the conventional method. The abundance of contents available online, such as TV-shows and movies, allows researchers to glean rich emotion data representative of various contexts efficiently. Crowdsourcing further supports inexpensive data annotation while serving as another data source [40, 41]. Datasets of this type have advantages in sample size and the diversity of subjects, but generalizability remains an issue. Datasets based on media contents often contain emotional displays produced by trained actors supposing fictitious situations. To what extent such emotional portrayals resemble spontaneous emotional expressions is debatable [42, 43, 44]. They also provide no access to physiological signals, which are known to carry information vital for the detection of less visible changes in emotional states [45, 46, 47, 48, 49, 50].

To amend this lack of a dataset for recognition of emotions in their natural forms, we introduce K-EmoCon, a multimodal dataset acquired from 32 subjects participating in 16 paired debates on a social issue. It consists of physiological sensor data collected with three off-the-shelf wearable devices, audiovisual footage of participants during the debate, and continuous emotion annotations. It contributes to the current literature of emotion recognition, as according to our knowledge, it is the first dataset with emotion annotations from all possible perspectives as the following: subject him/herself, debate partner, and external observers.

2 Methods

2.1 Dataset design

Intended usage Inspired by previous works that set out to investigate emotions during conversations [51, 52, 53, 38], K-EmoCon was designed in consideration of a social interaction scenario involving two people and wearable devices capable of unobtrusive tracking of physiological signals. The dataset aims to allow a multi-perspective analysis of emotions with the following objectives:

1. Extend the research on how having multiple perspectives on emotional expressions may improve their automatic recognition.
2. Provide a novel opportunity to investigate how emotions can be perceived differently from multiple perspectives, especially in the context of social interaction.

Previous research has shown that having multiple sources for emotion annotations can increase their recognition accuracy [54, 55]. However, no research in our awareness employs all three available perspectives in the annotation of emotions (i.e., subject him/herself, interacting partner, and external observers). Having multiple perspectives relates to the issue of establishing ground truth in emotion annotations. Emotions are inherently internal phenomena, and their mechanism is unavailable for external scrutiny, even for oneself who is experiencing emotions. As a result, there may not be a ground truth for emotions. Should we consider what is most agreed upon by external observers of emotions as the ground truth, or what the person who experiences emotions reports to have felt the ground truth [56]? Two views are likely to match if emotions are intense and pure, but as discussed, such emotions are rare. Instead, self-reported and observed emotions are likely to disagree for a variety of reasons. People often conceal their true emotions; sometimes, people are not fully mindful of their internal states; and some people may have difficulties in interpreting or articulating emotions [57, 58].

With K-EmoCon, we intend to enable the comprehensive examination of such cases where perceptions of emotions do not match, by bringing all three available perspectives into the annotation of emotions, in the context of a social interaction involving three parties of:

1. *The subject* – is the source who experiences emotions firsthand and produces *self annotations*, particularly the *"felt sense"* [55] of the emotions.
2. *The partner* – is the person who interacts with the subject, experiencing the subject's emotions secondhand; thus, he or she has a contextual knowledge of the interaction that induced the subject's emotions and produces *partner annotations* based on that.
3. *The external observers* – are people who observe the subject's emotions without the exact contextual knowledge of the interaction that induced the emotions, producing *external observer annotations*.

Notice, that while our definition of perspectives involved in emotion annotation is similar to definitions previously used by other researchers (self-reported vs. perceived [55]/observed [59]), we further segment observer annotations based on whether the contextual information of the situation in which the emotion was generated is available to an observer, as we wish to consider the role of contextual knowledge in emotion perception and recognition.

Existing datasets of emotions in conversations provide a limited scope on this issue as they at most contain emotion annotations from subjects and external observers [51], leaving out annotations from other people who engaged in the conversation (whom we call partners). Or, they either only consider a particular type of annotations that is sufficient to serve their research goal [53] or their designs do not allow acquiring multi-perspective annotations [52, 38] (e.g., a dataset is constructed upon conversations from a TV-show, only allowing the collection of external observer annotations). Refer to Table 1 to see how K-EmoCon is distinguished from existing emotion datasets.

Context of data collection In this regard, we chose a semi-structured, turn-taking debate on a social issue with randomly assigned partners as the setting for data collection. This setting is appropriate for collecting emotions that may naturally arise in a day, as it is similar to a social interaction that one could engage in a workplace.

Also, the setting is particularly suitable for studying the misperception of emotions. It is sufficiently formal and spontaneous as it involves randomly assigned partners. We expect such formality and spontaneity of the setting compelled participants to regulate their emotions in a socially appropriate manner, allowing us to observe less pronounced emotions from participants, which were more likely to be misperceived by their partners [60].

Data collection apparatus Our choice of mobile, wearable, and low-cost devices to collect affective physiological signals together with audiovisual recordings, while primarily aims to make findings based on our data more reproducible and expandable, was also in consideration of our goal of investigating mismatches in perceptions of emotions in the wild. Research has shown that fusing implicit and explicit affective information can result in more accurate recognition of subtle emotional expressions from professional actors [61]. However, no work we are aware of has shown that a similar result can be achieved for subtle emotions collected from in-the-wild social interactions of individuals without professional training in acting. Therefore, our dataset provides an opportunity to examine if emotions of lower intensity, produced from non-actors during communication, can be recognized accurately.

It is also interesting to examine whether subtle emotions could signal instances where emotions are misperceived during communication if their accurate detection is possible. In the same vein, to what extent the intensity of emotions influences their decoding accuracy during a social interaction, where a broader array of contextual information is present, is also worth exploring. K-EmoCon could enable an in-depth investigation of such issues.

Further, we considered the use case of mobile and wearable technologies for facilitating emotional communication. Researchers are actively exploring the potential for using expressive biosignals collected via wearables to communicate one's emotional and psychological states with others [62, 63, 64, 65, 66]. Our dataset can contribute to the research of

Table 1: Comparison of the K-EmoCon dataset with the existing multimodal emotion recognition datasets. Posed emotions are when a subject is instructed to enact a particular emotion while Spon. = spontaneous. Similarly, induced emotions are when a set of selected stimuli is used for their elicitation. For annotation types, S = *self annotations*, P = *partner annotations*, and E = *external observer annotations*.

Name (year)	**Size**	**Modalities**	**Spon. vs. posed**	**Natural vs. induced**	**Annotation method**	**Annotation type**	**Context**
IEMOCAP (2008) [51]	10	Videos, face motion capture, gesture, speech (audio & transcribed)	Both	Both†	Per dialog turn	S, E	Dyadic
SEMAINE (2011) [52]	150	Videos, FAUs, speech (audio & transcribed)	Spon.	Induced	Trace-style continuous	E	Dyadic
MAHNOB-HCI (2011) [23]	27	Videos (face and body), eye gaze, audio, biosignals (EEG, GSR, ECG, respiration, skin temp.)	Spon.	Induced	Per stimuli	S	Individual
DEAP (2012) [24]	32	Face videos, biosignals (EEG, GSR, BVP, respiration, skin temp., EMG & EOG)	Spon.	Induced	Per stimuli	S	Individual
DECAF (2015) [25]	30	NIR face videos, biosignals (MEG, hEOG, ECG, tEMG)	Spon.	Induced	Per stimuli	S	Individual
ASCERTAIN (2016) [26]	58	Facial motion units (EMO), biosignals (ECG, GSR, EEG)	Spon.	Induced	Per stimuli	S	Individual
MSP-IMPROV (2016) [53]	12	Face videos, speech audio	Both	Both†	Per dialog turn	E	Dyadic
DREAMER (2017) [27]	23	Biosignals (EEG, ECG)	Spon.	Induced	Per stimuli	S	Individual
AMIGOS (2018) [28]	40	Vidoes (face & body), biosignals (EEG, ECG, GSR)	Spon.	Induced	Per stimuli	S, E	Individual, Group
MELD (2019) [38]	7	Videos, speech (audio & transcribed)	Both	Both†	Turn-based	E	Dyadic, Group
CASE (2019) [29]	30	Biosignals (ECG, respiration, BVP, GSR, skin temp., EMG)	Spon.	Induced	Trace-style continuous	S	Individual
CLAS (2020) [95]	64	Biosignals (ECG, PPG, EDA), accelerometer	Spon.	Induced	Per stimuli/task	Predefined‡	Individual
K-EmoCon (2020)	*32*	*Videos (face, gesture), speech audio, accelerometer, biosignals (EEG, ECG, BVP, EDA, skin temp.)*	*Spon.*	*Natural*	*Interval-based continuous*	*S, P, E*	*Dyadic*

† A dataset was considered to contain induced emotions if scripted interaction was involved in the data collection, even though no artificial stimuli (such as an emotion inducing video clip) was used.

‡ Predefined emotion categories of stimuli and success rates of participants in a set of purposefully selected cognitive tasks were used as ground-truth labels.

biosignal-based assistive technologies to enable affective communication by providing insights on when are apposite moments for communicating emotions.

2.2 Ethics statement

The construction of the K-EmoCon dataset was approved by the Korea Advanced Institute of Science and Technology (KAIST) Institutional Review Board. KAIST IRB also reviewed and approved the consent form, which contained information on the following: the purpose of data collection, data collection procedure, types of data to be collected from participants, compensation to be provided for participation, and the protocol for the protection of privacy-sensitive data.

Participants were given the same consent forms upon arriving at the data collection site and were asked to provide written consent after fully reading the form indicating that they are willing to participate in data collection. Since K-EmoCon is to be open to public access, a separate consent was obtained for the disclosure of the data that contains personally identifiable information (PII), which is the audiovisual footage of participants during debates, including their faces and voices. Participants were also notified that their participation is voluntary, and they can terminate the data collection at any. The resulting K-EmoCon dataset includes the audiovisual recordings of 21 participants, out of 32, who agreed to disclose their personal information, excluding the 11 who did not agree.

2.3 Participant recruitment and preparation

32 participants were recruited between January and March of 2019. An announcement calling for participation in an experiment on "emotion-sensing during a debate" was posted on an online bulletin board of a KAIST student community. The post stated that participants would have a debate on the issue of accepting Yemeni refugees on Jeju Island of South Korea for 10 minutes. It also stated that the debate must be in English, and participants should be capable of speaking competently in English, but not necessarily at the level of a native speaker. Specifically, participants were required to have at least three years of experience living in an English-speaking country, or have achieved a score above criteria in any one of standardized English speaking tests as listed here: TOEIC speaking level 7, TOEFL speaking score 27, or IELTS speaking level 7.

Once participants were assigned a date and time to participate in data collection, they were provided four news articles on the topic of the Jeju Yemeni refugee crisis via email. The articles included two articles with neutral views on the issue [67, 68], one in favor of refugees [69], and one in opposition to refugees [70]. We instructed the participants to read the articles beforehand to familiarize themselves with the debate topic.

All selected participants were students at KAIST, but their ages varied from 19 to 36 years old (mean = 23.8 years, stdev. = 3.3 years), as well as their gender and nationality. We randomly paired participants into 16 dyads based on their available times. See Table 2 for the breakdown of participants' gender, nationality, and age.

Table 2: Participant pairs for debates.

Participants		Gender and ages	
P1	P2	M (25)	M (23)
P3	P4	M (36)	M (25)
P5	P6	M (22)	M (23)
P7	P8	M (22)	F (22)
P9	P10	M (21)	M (22)
P11	P12	M (22)	M (25)
P13	P14	M (22)	F (21)
P15	P16	M (30)	F (26)
P17	P18	M (21)	M (20)
P19	P20	M (21)	F (23)
P21	P22	M (25)	F (25)
P23	P24	M (22)	F (29)
P25	P26	F (26)	M (25)
P27	P28	F (24)	F (23)
P29	P30	F (23)	F (24)
P31	P32	M (24)	F (19)

2.4 Data collection setup

All data collection sessions were conducted in two rooms with controlled temperature and illumination. Two participants sat across a table facing each other with a distance in between for a comfortable communication (see Figure 1). Two Samsung Galaxy S7 smartphones mounted on tripods were placed in the middle of the table facing each participant, capturing facial expressions and movements in the upper body from the 3rd-person point of view (POV) along with the speech audio, via the camera app.

During a debate, participants wore a suite of wearable sensors, as shown in Figure 2, which includes:

1. *Empatica E4 Wristband* – captured photoplethysmography (PPG), 3-axis acceleration, body temperature, and electrodermal activity (EDA). Heart rate and the inter-beat interval (IBI) were derived from Blood Volume Pulse (BVP) measured by a PPG sensor.

Figure 1: Picture on the left shows a pair of participants sitting at a table preparing for a debate. Two smartphones on tripods in the middle of the table (highlighted in red) recorded participants' facial expressions and movements in their upper body, as shown on the right in the sample screenshot of footage.

Figure 2: Frontal view of a participant equipped with wearable sensors.

2. *Polar H7 Bluetooth Heart Rate Sensor* – detected heart rates using an electrocardiogram (ECG) sensor and was used to complement a PPG sensor in E4, which is susceptible to motion.
3. *NeuroSky MindWave Headset* – collected electroencephalogram (EEG) signals via two dry sensor electrodes, one on the forehead (fp1 channel-10/20 system at the frontal lobe) and one on the left earlobe (reference).
4. *LookNTell Head-Mounted Camera* – with a camera attached at one end of a plastic circlet, was worn on participants' heads to capture videos from a first-person POV.

Table 3: Data collected with each wearable device, with respective sampling rates and signal ranges.

Devices	Collected data	Sampling rate	Signal range [min, max]
Empatica E4 Wristband	3-axis acceleration	32Hz	[-2g, 2g]
	BVP (PPG)	64Hz	n/a
	EDA	4Hz	[0.01μS, 100μS]
	Heart rate (from BVP)	1Hz	n/a
	IBI (from BVP)	n/a	n/a
	Body temperature	4Hz	[$-40\,°C$, $115\,°C$]
NeuroSky MindWave Headset	Brainwave (fp1 channel EEG)	125Hz	n/a
	Attention & Meditation	1Hz	[0, 100]
Polar H7 Heart Rate Sensor	HR (ECG)	2Hz	n/a

All listed devices can operate in a mobile setting. Empatica E4 keeps the data on the device, and the collected data is later uploaded to a computer. Polar H7 sensor and MindWave headset can communicate with a mobile phone via Bluetooth Low Energy (BLE) to store data. Table 3 summarizes sampling rates and signal ranges of data collected from each device.

2.5 Data collection procedure

Administration All data collection sessions were conducted in four stages of 1) onboarding, 2) baseline measurement, 3) debate, and 4) emotion annotation. Two experimenters administered each session (see Table 4 for the overview of a data collection procedure). One experimenter served as a moderator during debates, notifying participants of the remaining time and intervening under any necessary circumstances, such as when a debate gets too heated, or a participant exceeds an allotted time of 2 minutes in his or her turn.

Onboarding Upon their arrival, participants were each provided a consent form asking for two written consents, first for the participation in data collection that was mandatory, and second for the disclosure of privacy-sensitive data collected during the session, which participants could opt-out without any disadvantage.

Once they agreed to participate in the research, participants decided whether they would argue for or against admitting the Yemeni refugees in Jeju. Participants could either briefly discuss with each other to settle on their preferred positions or toss a coin to decide at random. The same procedure was followed for deciding who goes first in the debate.

Next, participants were given up to 15 minutes to prepare their arguments. Each participant was given a pen, paper, and prints of the articles that they previously received via email. After they finished preparing, experimenters equipped participants with wearable devices. Participants wore E4 wristbands on their non-dominant hand, as arm movements may impede an accurate measurement of PPG. Experimenters assured that wristbands are tightly fastened, and electrodes are in good contact with participants' skin. Experimenters also assured the EEG headsets and head-mounted cameras are well fitted on participants' head, and manually adjusted head-mounted cameras' lens to make sure the captured views are similar to participants' subjective view. Participants wore Polar H7 sensors attached to flexible bands underneath their clothes, so the electrodes are in contact with their skin and placed the sensors above their solar plexus.

Baseline measurement With all devices equipped, sensor measurements were taken from participants while they watched a short clip. This step was to establish a baseline that constitutes a neutral state for each participant. Establishing a neutral baseline is commonly used in the construction of emotion datasets to account for individual biases and reduce the effect of previous emotional states, especially when repeated measurements are taken.

A procedure for a baseline measurement varies across researchers and is often dependent on the purpose of an experiment [71]. In stimuli-based experiments, researchers take measurements as their subjects watch a stimulus intended to induce a neutral emotional state [23, 24] or measure resting-state activities between stimuli if they are taking multiple consecutive measurements [25]. Similarly, for K-EmoCon, participants watched *Color Bars* clip, which was previously reported in the work of Gross et al. to induce a neutral emotion [72]. Experimenters also ensured that no devices were malfunctioning during the baseline measurement.

Table 4: Steps for a data collection session.

Step	Allocated time	Description
Read and sign consent forms	10 min	Experimenters provided consent forms to participants, and two written consents each for participation and the collection of privacy-sensitive data were obtained.
Choose sides and the order	5 min	Participants were assigned to either argue in favor of or against accepting refugees and decided on the first speaker.
Prepare debate	15 min	Participants were provided with supplementary materials to prepare their arguments.
Equip sensors	10 min	Experimenters explained wearable devices to participants and assisted them in wearing devices.
Measure baseline	2 min	A baseline corresponding to a neutral state was measured for each participant.
Overview debate	5 min	The moderator explained the debate rules and notified participants that they are allowed to intervene.
Debate	10 min	Participants could speak for two consecutive minutes during their turns and they were notified twice at 30 and 60 seconds before the end of the debate.
Annotate emotions	60 min	Participants annotated emotions at intervals of every 5 seconds, watching footage of themselves and their partners.

* One session lasted approximately two hours.

Debate A debate began at the sign of the moderator and lasted approximately 10 minutes. Participants' facial expressions, movements in their upper body, and speeches were recorded throughout a debate. Participants were allowed to speak consecutively up to two minutes during their turns, with turns alternating between two participants. However, participants were also notified that they could intervene during an opponent's turn, to allow a more natural communication. The moderator notified participants 30 and 60 seconds before the end of their turns and intervened if they exceeded two minutes. A debate stopped at the ten-minute mark with some flexibility to allow the last speaker to finish his or her argument.

Emotion annotation Participants took a 15-minute break upon finishing a debate. Participants then were each assigned to a PC and annotated their own emotions and their partner's emotions during the debate. Specifically, each participant watched two audiovisual recordings of him/herself and his/her partner from 3rd-person POV (including facial expressions, upper body movements, and speeches), to annotate emotions at intervals of every 5 seconds from the beginning to the end of a debate. We chose 5 seconds based on the report of Busso et al. that the average duration of the speaker turns in IEMOCAP was about 4.5s [51], and findings from linguistics research also support this number [73, 74, 75].

This annotation method we employed, a *retrospective affect judgment protocol*, is widely used in affective computing to collect self-reports of emotions, especially in studies where an uninterrupted engagement of subjects during an emotion induction process is essential [76, 77, 78, 79]. Likewise, we opted for this method as participants' natural interaction was necessary for acquiring quality emotion data.

Note that we did not provide 1st-person POV recordings captured from head-mounted cameras to participants, and they only had 3rd-person POV recordings to annotate felt emotions. One may have a reasonable concern regarding this choice, that participants watching their faces likely caused them to occupy a perspective similar to an observer. Hence, this might have resulted in an unnatural measurement of felt emotions. Indeed, the headcam footage could have been a more naturalistic instrument, as we intuitively take an embodied perspective to recall how we felt at a specific moment in the past.

However, we found the extent of information captured by the headcam footage insufficient for accurate annotation of felt emotions. Experimenters manually adjusted headcam lenses, so the recordings resembled participants' subjective views, but the headcam footage was missing fine-grained information such as participants' gazes. Also, past research

Table 5: Collected emotion annotations.

Emotion annotation categories	Description	Measurement scale or method
Arousal / Valence	Two affective dimensions from Russell's circumplex model of affect [96]	1: very low - 2: low - 3: neutral - 4: high - 5: very high
Cheerful / Happy / Angry / Nervous / Sad	Emotion states describing a subjective stress state [97]	1: very low - 2: low - 3: high - 4: very high
Boredom / Confusion / Delight / Engaged concentration / Frustration / Surprise / None	Commonly used Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) educationally relevant affective categories [98]	Choose one
Confrustion / Contempt / Dejection / Disgust / Eureka / Pride / Sorrow / None	Less commonly used BROMP educationally relevant affective categories [98]	Choose one

on memories for emotions has shown that they are prone to biases and distortion [80, 81, 82]. In that regard, it seemed headcam videos, which contain limited information compared to frontal face recordings, would only result in an incorrect annotation of felt emotions, especially in retrospect. Further, we noted that it is not uncommon for people to infer emotions from their faces, as they frequently do when looking in a mirror or taking a selfie.

As a result, participants were given 3rd-person recordings of themselves for the retrospective annotation of felt emotions. In total, participants annotated emotions with 20 unique categories, as shown in Table 5. Experimenters assisted participants throughout the annotation procedure. Before participants began annotating, experimenters explained individual emotion categories to participants, so they correctly understood a meaning and a specific annotation procedure for each item. Experimenters also explicitly instructed participants to report felt emotions, not perceived emotions on their faces. Lastly, experimenters ensured that the start time and end time for two participants matched to obtain synchronized annotations.

External emotion annotation Additionally, we recruited five external raters to annotate participants' emotions during debates (see Table 6). We applied the same criteria we used for recruiting participants in data collection for the recruitment of the raters. The raters were provided the audiovisual debate footage of participants and annotated emotions following the same procedure our participants followed. External raters performed their tasks independently, and the experimenters communicated remotely with the raters. Once a rater finished annotating, an experimenter checked completed annotations for incorrect entries and requested a rater to review annotations if there were any missing values or misplaced entries.

Table 6: Gender and age of external raters.

Raters	Gender and age
R1	M (27)
R2	M (25)
R3	F (22)
R4	M (24)
R5	F (28)

3 Data Records

3.1 Dataset summary

The resulting K-EmoCon dataset contains multimodal data from 16 paired-debates on a social issue, which sum to 172.92 minutes of dyadic interaction. It includes physiological signals measured with three wearable devices, audiovisual recordings of debates, and continuous annotations of emotions from three distinct perspectives of the subject, the partner, and the external observers. Table 7 summarizes data collection results and dataset contents.

Preprocessing For the timewise synchronization across data, we converted all timestamps from Korea Standard Time (UTC +9) to UTC +0 and clipped raw data such that only parts of data corresponding to debates and baseline measurements are included. For debate audios and the footage, subclips corresponding to debates were extracted from the raw footage. Audio tracks containing participants' speeches were copied and saved separately as WAV files. Physiological signals were clipped from the respective beginnings of data collection sessions to the respective ends of debates, as the initial 1.5 to 2 minutes immediately after a session begins corresponds to a baseline measurement for a neutral state. Parts in between baseline measurements and debates correspond to debate preparations, which may be excluded from the analysis.

Table 7: Summary of data collection results and the dataset.

Data collection summary	
Number of participants	32 (20 males and 12 females)
Participants age	19 to 36 (mean = 23.8 years, stdev. = 3.3 years)
Session duration	Total 172.92 min, (mean = 10.8 min, stdev. = 1.04 min)
Emotion annotations categories	**1 - 5**: Arousal, Valence **1 - 4**: Cheerful, Happy, Angry, Nervous, Sad **Choose one**: Common BROMP affective categories + less common BROMP affective categories
Measured physiological signals	3-axis Acc. (32Hz), BVP (64Hz), EDA (4Hz), heart rate (1Hz), IBI (n/a), body temperature (4Hz), EEG (8 band, 32Hz), ECG (2Hz)
Dataset contents	
Debate audios	172.92 min (from 16 debate sessions)
Debate footage	223.35 min (from 21 participants)
Physiological signals	Refer to *Dataset contents* subsection
Emotion annotations (# of 5-second intervals annotated)	**Self**: 4,159 **Partner**: 4,159 **5 external observers**: 20,803

3.2 Dataset contents

The K-EmoCon dataset [83] is available upon request on *Zenodo*: `https://doi.org/10.5281/zenodo.3814370`. In the following, we describe directories and files in the dataset and their contents.

metadata.tar.gz: includes files with auxiliary information about the dataset. Included files are:

1. `subjects.csv` – each row contains a participant ID (`pid`) and three timestamps in UTC +0. Three timestamps respectively mark the beginning of a data collection (`initTime`), the start of a debate (`startTime`), and the end of a debate (`endTime`).
2. `data_availability.csv` – shows files available for each participant. For each participant (row), if a data file (column) is available, the corresponding cell is marked `TRUE`, otherwise `FALSE`.

data_quality_tables.tar.gz: includes seven CSV tables with information regarding the quality of physiological signals in the dataset. With participant IDs (`pid`) in rows and file types (`ACC`, `BVP`, `EDA`, `HR`, `IBI`, and `TEMP` for E4 data, and `Attention`, `BrainWave`, `Meditation`, and `Polar_HR` for NeuroSky + Polar H7 data) in columns, included files are as follows:

1. `e4_durations.csv` – contains the duration of each file in seconds, where `duration` = (last timestamp - first timestamp) / 1000.
2. `neuro_polar_durations.csv` – same as above.
3. `e4_zeros.csv` – contains the number of zero values in each file. `ACC` and `BVP` were excluded as zero crossings are to be expected during their measurement.
4. `neuro_polar_zeros.csv` – same as above. Note that zero values for NeuroSky data (`Attention`, `BrainWave`, `Mediation`) indicate the inability of a device at a given moment to obtain a sufficiently reliable measurement due to various reasons.
5. `e4_outliers.csv` – contains the number of outliers in each file. Chauvenet's criterion was used for outlier detection (refer to *Code Availability* section for its implementation in Python).

6. `e4_completeness.csv` – contains the completeness of each file as a ratio in the range of [0.0, 1.0]. 1.0 indicates a file without any missing value or an outlier. The completeness ratio was calculated as `completeness` = (total number of values - (number of outliers + number of zeros)) / total number of values.

7. `neuro_polar_completeness.csv` – same as above, with completeness calculated as `completeness` = (total number of values - number of zeros) / total number of values.

debate_audios.tar.gz: contains 16 audio recordings of debates in the WAV file format. The name of each file follows the convention of p<X>.p<Y>.wav, where <X> and <Y> stand for IDs of two participants appearing in the audio. The start and the end of each recording correspond to `startTime` and `endTime` values in the subjects.csv file, respectively.

debate_recordings.tar.gz: contains video recordings of 21 participants during debates in the MP4 file format. The name of a file p<X>_<T>.mp4 indicates that the file is the recording of participant <X> that is <T> seconds long.

neurosky_polar_data.tar.gz: includes subdirectories for each participant, from P1 to P32, which may contain up to four files as the following:

1. `Attention.csv` – contains *eSense Attention* ranging from 1 to 100, representing how attentive a user was at a given moment. Attention values can be interpreted as the following: 1 to 20 – "strongly lowered", 20 to 40 – "reduced", 40 to 60 – "neutral", 60 to 80 – "slightly elevated", and 80 to 100 – "elevated". 0 indicates that the device was unable to calculate a sufficiently reliable value, possibly due to a signal contamination with noises.

2. `BrainWave.csv` – records the relative power of brainwave in the following 8 bands of EEG: delta (0.5 - 2.75Hz), theta (3.5 - 6.75Hz), low-alpha (7.5 - 9.25Hz), high-alpha (10 - 11.75Hz), low-beta (13 - 16.75Hz), high-beta (18 - 29.75Hz), low-gamma (31 - 39.75Hz), and middle-gamma (41 - 49.75Hz). The values are without a unit and are only meant for inferring the fluctuation in the power of a certain band or comparing the relative strengths of bands with each other.

3. `Meditation.csv` – contains *eSense Meditation* ranging from 0 to 100, measuring the relaxedness of a user. For their interpretation, use the same ranges and the meanings as those for the attention values.

4. `Polar_HR.csv` - contains heart rates measured with ECG sensors during debates.

e4_data.tar.gz: contains subdirectories for each participant (except P2, P3, P6, and P7), which may contain up to six files as the following:

1. `E4_ACC.csv` – measurements from a 3-axis accelerometer sampled at 32Hz in the range [-2g, 2g] under columns x, y, and z. Multiply raw numbers by 1/64 to convert them into units of g (i.e., a raw value of 64 is equivalent to 1g).

2. `E4_BVP.csv` – PPG measurements sampled at 64Hz.

3. `E4_EDA.csv` – EDA sensor readings in units of μS, sampled at 4Hz.

4. `E4_HR.csv` – the average heart rates calculated in 10-second windows. The values are derived from the BVP measurements, and the values are entered at the frequency of 1Hz. The first 10 seconds of data after the beginning of a recording is not included as the derivation algorithm requires the initial 10 seconds of data to produce the first value.

5. `E4_IBI.csv` – IBI measurements in milliseconds computed from the BVP. From a second row onwards, one row is separated from the previous row with an amount equal to a distance between two peaks (i.e., $t_{i+1} - t_i = IBI_i$). Note that HR in terms of BPM can be derived from IBI by taking $60/(IBI * 1000)$.

6. `E4_TEMP.csv` – a body temperature measured in the Celsius scale at the frequency of 4Hz.

Note that E4 data entries for P29, P30, P31, and P32 are entered with each row designated with either one of two unique `device_serial` values. It is necessary that users of this dataset only use rows corresponding to a single `device_serial`. We further recommend using rows with the following `device_serial` values:

- P29, P31 – `A013E1` for all files, except `A01525` for IBI.
- P30, P32 – `A01A3A` for all files.

emotion_annotations.tar.gz: includes four subdirectories as listed below, which each contain annotations for participant emotions during debates at intervals of every 5 seconds, acquired from three distinct perspectives:

1. `self_annotations` – annotations of participant emotions by participants themselves.
2. `partner_annotations` – annotations of participant emotions by respective debate partners.
3. `external_annotations` – annotations of participant emotions by five external raters. Files follow the naming convention of `P<X>.R<Z>.csv`, where `<X>` is a participant ID, and `<Z>` is a rater number.
4. `aggregated_external_annotations` – contains external rater annotations aggregated across five raters via majority voting. Refer to *Code Availability* section for the Python code implementing the majority vote aggregation.

The first row in a valid file has annotations for the first five seconds, and rows coming afterward contain annotations for the next consecutive five-second intervals, non-overlapping with each other. Also, each row in a valid file contains 10 non-empty values (eight numeric values, including `seconds` column, and two `x`'s). Note that annotation files for a participant may not have an equal number of rows (e.g., there may be more self-annotations than partner/external annotations for some participants). In that case, longer files should be truncated from the start such that they have the same number of rows as shorter files since the extra annotations at the beginning are possibly from participants mistakenly annotating emotions during baseline measurements.

4 Technical Validation

4.1 Emotion annotations

Distribution and frequency of emotions The distributions and the frequencies of emotion annotations are as shown in Figure 3. Overall, annotations for emotions measured on Likert scales (arousal, valence, cheerful, happy, angry, nervous, and sad) are biased towards a neutral with only a minuscule fraction of annotations for non-neutral states. Categorical emotion annotations (common and less common BROMP affective categories) are similarly biased, with a predominant portion of annotations falling under only two categories of concentration and none. This imbalance in annotations is as expected as emotion data is commonly imbalanced by its nature in the wild (i.e., people are more often neutral than angry or sad) [84, 85, 86].

Inter-rater reliability As individual-level information is missing in aggregated data, we used Krippendorff's alpha [87], which is a generalized statistic of agreement applicable to any number of raters, to measure the inter-rater reliability (IRR) of emotion annotations from different perspectives for each participant. Figure 4 shows heatmaps of alpha coefficients computed for seven emotions measured on ordinal scales (arousal, valence, cheerful, happy, angry, nervous, and sad).

All annotation values were interpreted as rank-ordered (ordinal scaled) for the IRR computation. Likert scales we used are not intervals or ratios with meaningful distances in-between. While participants and raters were provided numeric scales labeled with semantic meanings (see Table 5), the individual interpretations of scales were likely disparate.

Given that, before the computation, annotation values were scaled relative to a neutral, by estimating modes of columns as neutrals and deducting them from respective column values (i.e., if the mode of a *cheerful* column for a particular participant was one, then one was subtracted from all values in that *cheerful* column). This *mode-subtraction* step was necessary to prevent the underestimation of IRRs.

Annotations in our dataset for scaled emotions are highly biased as shown in Fig. 3). However, while arousal and valence are explicitly centered at zero (which corresponds to 3 = neutral), five emotions measured in the scale of 1 = *very low* to 4 = *very high* (cheerful, happy, angry, nervous, and sad) are systematically biased without a zero neutral. All of their values indicate that some emotion is present, and this absence of zero results in a widely varying interpretation of scale values by our participants and raters.

Consider the following scenario further elaborating this issue: a subject rates that she was cheerful as much as 1 for the first third of a debate, then 2 for the rest, but her debate partner rates that she was cheerful as much as 3 for the first third then 4 for the rest. In this example, self and partner annotations both imply that the subject was less cheerful for the first third of the debate. However, an IRR of two sets of annotations is close to zero without subtracting modes. Indeed, it is possible that the partner perceived the subject as more cheerful overall, compared to the subject herself. In that case, a low IRR correctly measures the difference between emotion perceptions of the subject and partner. Nevertheless, this assumption cannot be confirmed, as there is no neutral baseline.

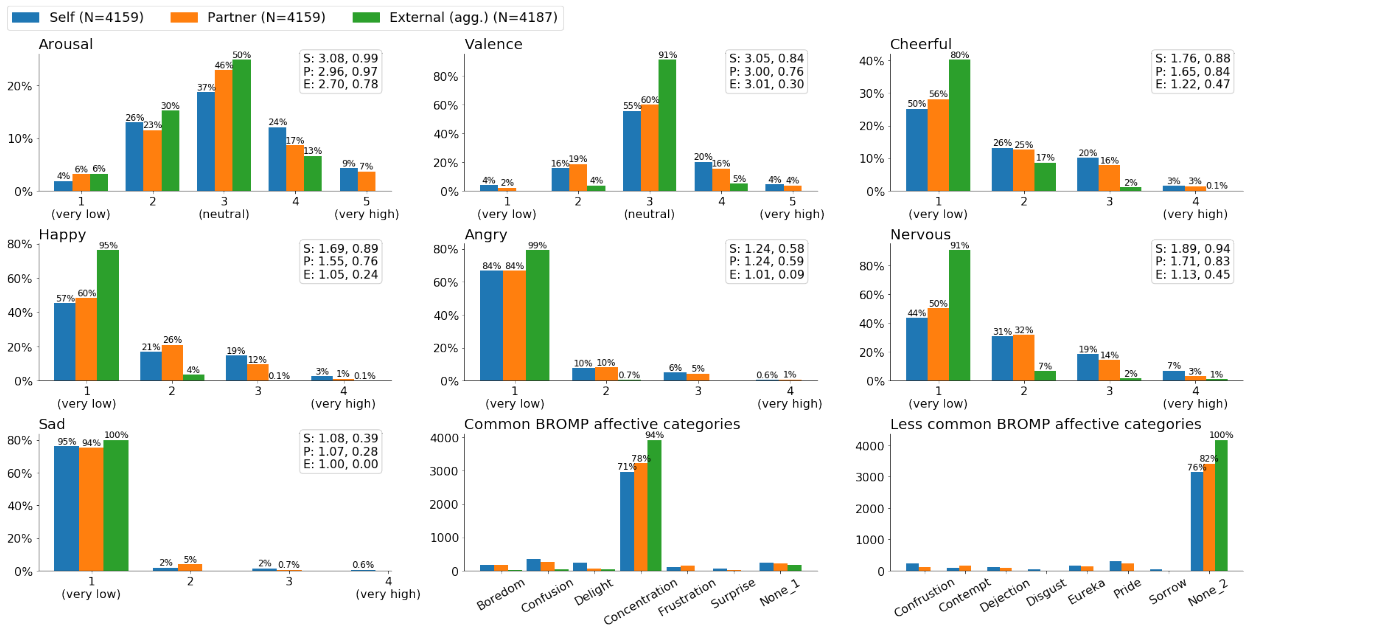

Figure 3: Distributions and frequencies of emotion annotations from three perspectives of self (S), partner (P), and external raters (E), with external annotations aggregated by majority voting. Annotations were summed across 32 subjects for each emotion and affective categories. Means and standard deviations measured respectively from three perspectives are shown on the upper right corner of figures if available.

Therefore, we applied the proposed mode-subtraction to emotion annotations such that alpha coefficients measure raters' agreement on relative changes in emotions rather than their absolute agreement with each other. This adjustment mitigates spuriously low alpha coefficient values obtained from raw annotations (refer to *Code Availability* section for the code implementing the mode-subtraction and plotting of heatmaps).

These fixed alpha coefficients are low in general. In particular, a noticeable pattern emerges when comparing alpha coefficients of self-partner (*SP*) annotations and self-external (*SE*) annotations. As shown in the last rows of heatmaps *(Diff. [SE - SP])* in Fig. 4, the differences between the IRRs of SE annotations and SP annotations tend to be above zero (for 20 out of 32 participants for arousal: mean = 0.143, stdev. = 0.322). This pattern possibly indicates that there exists a meaningful difference in the perception of emotions from different perspectives, while further study is required to validate its significance.

4.2 Physiological signals

Data quality The quality of physiological signal measurements in the dataset has been thoroughly examined. The examination results are included as a part of the dataset in the `data_quality_tables.tar.gz` archive file.

Missing data E4 data of 4 participants (P2, P3, P6, and P7) were excluded due to a device malfunction during data collection. While physiological signals in the dataset are mostly error-free with most of the files complete above 95%, a portion of data is missing due to issues inherent to devices or a human error:

- IBI – data from P26 is missing as the internal algorithm of E4 that derives IBI from BVP automatically discards an obtained value if its reliability is below a certain threshold.
- EDA – data from P17 and P20 is missing, possibly due to poor contact between the device and a participant's skin.
- NeuroSky (Attention, Meditation) – measurements from P1 and P20 are missing due to a poorly equipped device. A portion of data is missing for P19 (32%), P22 (59%) and P23 (36%) for the same reason. No BrainWave data was lost.
- Polar HR – data from seven participants (P3, P12, P18, P20, P21, P29, and P30) are missing due to a device error during data collection. Parts of data are missing from P4 (38%) and P22 (38%) due to poor contact.

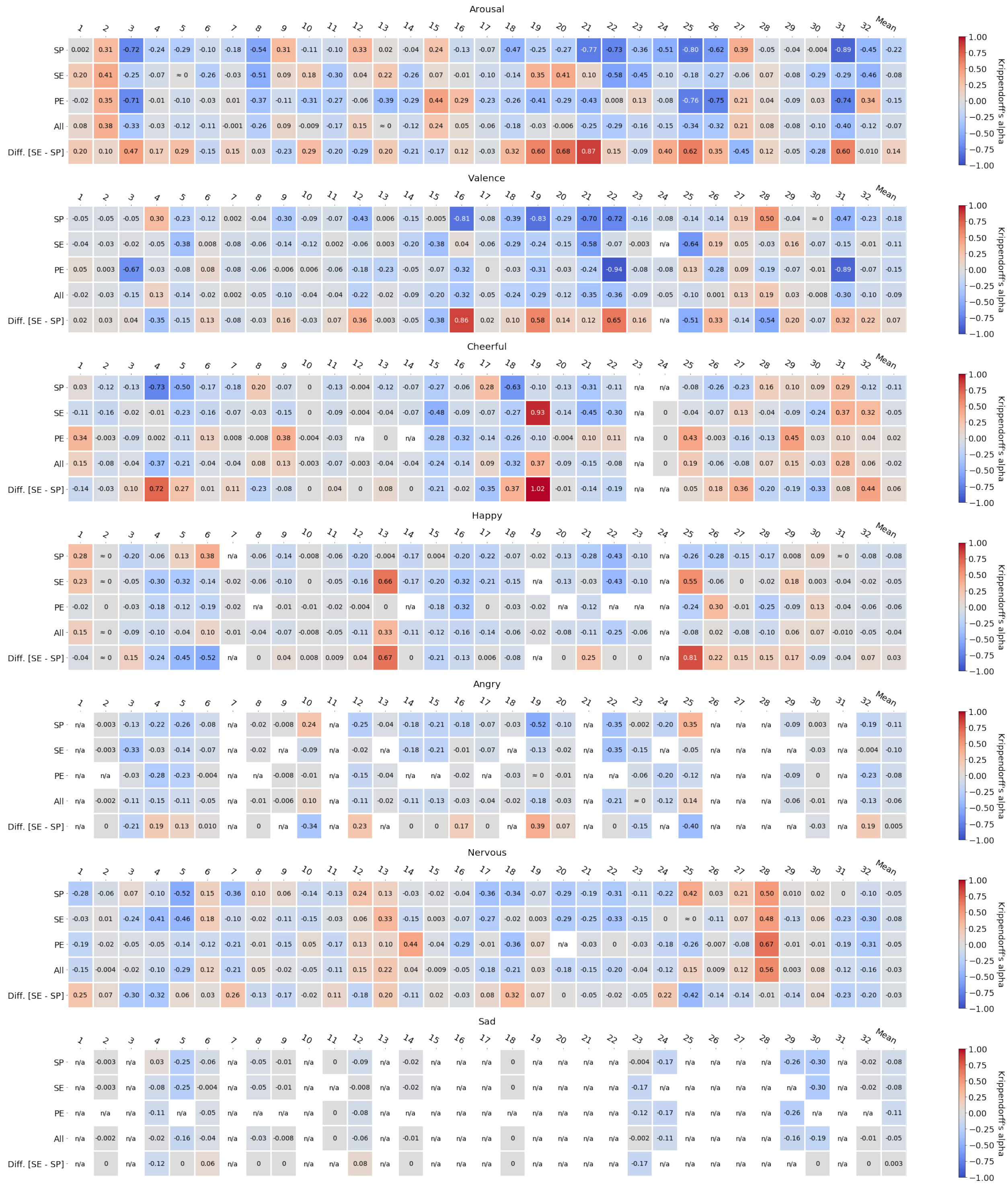

Figure 4: Heatmaps of inter-rater reliabilities measured with Krippendorff's alpha. External annotations were aggregated by majority voting. The first 4 rows of each heatmap show alpha coefficients across four different combinations of annotation perspectives: (1) *SP* = self vs. partner, (2) *SE* = self vs. external, (3) *PE* = partner vs. external, and (4) *All* = self vs. partner vs. external, while the last row (*Diff [SE - SP]*) shows the difference between self vs. external agreement and self vs. partner agreement. The columns show those for each participant.

5 Usage Notes

5.1 Potential applications

In addition to the intended usage of the dataset discussed above, there are uncertainties as to how physiological markers of an individual's capacity for flexible physiological reactivity relate to experiences of positive and negative emotions. Our dataset could potentially be useful to examine the role of physiological signal based markers in assessing an individual's use of emotion regulation strategies, such as cognitive appraisal.

Additionally, various data mining and machine learning techniques could be applied to set up the models for an individual's emotional profiles based on sensor-based physiological and behavioral recordings. This could further be transferred to other use-cases, such as helping children with autism in their social communication [88, 89], helping people who are blind to read facial expressions and get the emotion information of their peers [90], helping robots interact more intelligently with people [91, 92], and monitoring signs of frustration and emotional saturation that affect attention while driving, to enhance driver safety [93, 94].

5.2 Limitations

Data collection apparatus Contact-base EEG sensors are known to be susceptible to noises, for example, frowning or eyes-movement might have caused peaks in the data. Other devices may also have been subject to similar systematic errors.

Data collection context The context of the turn-taking debate may have caused participants to regulate or even suppress their emotional expressions, as an unrestrained display of emotions is often regarded undesirable during a debate. This may have contributed to a deflated level of agreement between self-reports and partner/external perceptions of emotions, which may not be a case for more natural interactions in the wild.

Demographics The participant demographics likely have introduced bias in the data. All of our participants and raters are young (their ages were between 19 to 36) and highly-educated, and the majority of them are individuals of Asian ethnicity. Therefore, our data may not generalize well to individuals of different ethnic groups or of younger or older age groups.

Unaccounted variables Many variables unaccounted during data collection, such as the level of rapport between debating pairs, a participant's competence in spoken English, and a participant's familiarity with the debate topic, may also have contributed to a variance in the level of mismatch between the perceptions of emotions across different perspectives.

6 Code Availability

Python codes implementing outlier detection using Chauvenet's criterion, majority voting, mode-subtraction, and other utility functions, including the generation of heatmap plots, are available on `https://github.com/Kaist-ICLab/K-EmoCon_SupplementaryCodes`. The *Krippendorff* Python package (`https://github.com/pln-fing-udelar/fast-krippendorff`) was used for the computation of Krippendorff's alpha. The Python of version 3.6.9 was used throughout.

Codes for preprocessing the raw log-level data in SQL databases to CSV files were implemented in Python with the *SQLAlchemy* package. However, these codes and the raw log-level data are not made available as they include privacy-sensitive information outside the agreed boundary for public sharing of the dataset, which was collected only for logistic reasons. Nevertheless, we welcome users of the dataset to contact the corresponding authors if they need any further assistance or information regarding the raw data, and it's preprocessing.

References

[1] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.

[2] John D Mayer, David R Caruso, and Peter Salovey. Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27(4):267–298, 1999.

[3] Peter Ed Salovey and David J Sluyter. *Emotional development and emotional intelligence: Educational implications.* Basic Books, 1997.

[4] Paulo N Lopes, Marc A Brackett, John B Nezlek, Astrid Schütz, Ina Sellin, and Peter Salovey. Emotional intelligence and social interaction. *Personality and social psychology bulletin*, 30(8):1018–1034, 2004.

[5] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[6] Rafail-Evangelos Mastoras, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Seada Kassie, Taoufik Alsaadi, Ahsan Khandoker, and Leontios J Hadjileontiadis. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Scientific reports*, 9(1):1–12, 2019.

[7] CB Insights. 40+ corporations working on autonomous vehicles. 2020. `https://www.cbinsights.com/research/autonomous-driverless-vehicles-corporations-list/`.

[8] Cassio Pennachin and Ben Goertzel. Contemporary approaches to artificial general intelligence. In *Artificial general intelligence*, pages 1–30. Springer, 2007.

[9] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[10] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[11] Byron Reeves and Clifford Ivar Nass. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.

[12] Aaron Turpen. Mit wants self-driving cars to traffic in human emotion. *New Atlas*, 2019. `https://newatlas.com/automotive/mit-self-driving-cars-human-emotion/`.

[13] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain.* Houghton Mifflin Harcourt, 2017.

[14] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[15] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255. IEEE, 2017.

[16] Mark G Frank and Elena Svetieva. Microexpressions and deception. In *Understanding facial expressions in communication*, pages 227–242. Springer, 2015.

[17] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[18] James M Carroll and James A Russell. Do facial expressions signal specific emotions? judging emotion from the face in context. *Journal of personality and social psychology*, 70(2):205, 1996.

[19] Richard T Cauldwell. Where did the anger go? the role of context in interpreting emotion in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[20] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

[21] Randy J Larsen and Ed Diener. Affect intensity as an individual difference characteristic: A review. *Journal of Research in personality*, 21(1):1–39, 1987.

[22] James J Gross and Oliver P John. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348, 2003.

[23] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

[24] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[25] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015.

[26] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016.

[27] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017.

[28] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 2018.

[29] Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data*, 6(1):1–13, 2019.

[30] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–7. IEEE, 2013.

[31] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 400–408, 2018.

[32] David Watson. *Mood and temperament*. Guilford Press, 2000.

[33] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth. How to find trouble in communication. *Speech communication*, 40(1-2):117–143, 2003.

[34] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[35] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, (3):34–41, 2012.

[36] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[37] Daniel McDuff, May Amr, and Rana El Kaliouby. Am-fed+: An extended dataset of naturalistic facial expressions collected in everyday settings. *IEEE Transactions on Affective Computing*, 10(1):7–17, 2018.

[38] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

[39] Daniel McDuff, Rana El Kaliouby, and Rosalind W Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012.

[40] R Morris, Daniel McDuff, and R Calvo. Crowdsourcing techniques for affective computing. In *The Oxford handbook of affective computing*, pages 384–394. Oxford Univ. Press, 2014.

[41] Olga Korovina, Marcos Baez, and Fabio Casati. Reliability of crowdsourcing as a method for collecting emotions labels on pictures. *BMC research notes*, 12(1):1–6, 2019.

[42] Michael T Motley and Carl T Camden. Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Communication (includes Communication Reports)*, 52(1):1–22, 1988.

[43] Rebecca Jürgens, Annika Grass, Matthis Drolet, and Julia Fischer. Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of nonverbal behavior*, 39(3):195–214, 2015.

[44] Patrik N Juslin, Petri Laukka, and Tanja Bänziger. The mirror to our soul? comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior*, 42(1):1–40, 2018.

[45] John T Cacioppo, Gary G Berntson, Jeff T Larsen, Kirsten M Poehlmann, Tiffany A Ito, et al. The psychophysiology of emotion. *Handbook of emotions*, 2:173–191, 2000.

[46] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[47] Christine Lætitia Lisetti and Fatma Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Advances in Signal Processing*, 2004(11):929414, 2004.

[48] Pierre Rainville, Antoine Bechara, Nasir Naqvi, and Antonio R Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International journal of psychophysiology*, 61(1):5–18, 2006.

[49] Lauri Nummenmaa, Enrico Glerean, Riitta Hari, and Jari K Hietanen. Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, 111(2):646–651, 2014.

[50] Edward F Pace-Schott, Marlissa C Amole, Tatjana Aue, Michela Balconi, Lauren M Bylsma, Hugo Critchley, Heath A Demaree, Bruce H Friedman, Anne Elizabeth Kotynski Gooding, Olivia Gosseries, et al. Physiological feelings. *Neuroscience & Biobehavioral Reviews*, 2019.

[51] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[52] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

[53] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016.

[54] Jennifer Healey. Recording affect in the field: towards methods and metrics for improving ground truth labels. In *International conference on affective computing and intelligent interaction*, pages 107–116. Springer, 2011.

[55] Biqiao Zhang, Georg Essl, and Emily Mower Provost. Automatic recognition of self-reported and perceived emotion: Does joint modeling help? In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 217–224, 2016.

[56] Khiet P Truong, David A van Leeuwen, and Mark A Neerincx. Unobtrusive multimodal emotion detection in adaptive interfaces: speech and facial expressions. In *International Conference on Foundations of Augmented Cognition*, pages 354–363. Springer, 2007.

[57] James B Grossman, Ami Klin, Alice S Carter, and Fred R Volkmar. Verbal bias in recognition of facial emotions in children with asperger syndrome. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41(3):369–379, 2000.

[58] Hannah Dickson, Monica E Calkins, Christian G Kohler, Sheilagh Hodgins, and Kristin R Laurens. Misperceptions of facial emotions among youth aged 9–14 years who present multiple antecedents of schizophrenia. *Schizophrenia bulletin*, 40(2):460–468, 2014.

[59] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech communication*, 54(9):1049–1063, 2012.

[60] Ursula Hess, Sylvie Blairy, and Robert E Kleck. The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4):241–257, 1997.

[61] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[62] Hyeryung Christine Min and Tek-Jin Nam. Biosignal sharing for affective connectedness. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 2191–2196. 2014.

[63] Mariam Hassib, Daniel Buschek, Paweł W Wozniak, and Florian Alt. Heartchat: Heart rate augmented mobile chat to support empathy and awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2239–2251, 2017.

[64] Fannie Liu, Laura Dabbish, and Geoff Kaufman. Supporting social interactions with an expressive heart rate sharing application. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–26, 2017.

[65] Fannie Liu, Mario Esparza, Maria Pavlovskaia, Geoff Kaufman, Laura Dabbish, and Andrés Monroy-Hernández. Animo: Sharing biosignals on a smartwatch for lightweight social connection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–19, 2019.

[66] Fannie Liu, Geoff Kaufman, and Laura Dabbish. The effect of expressive biosignals on empathy and closeness for a stigmatized group member. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–17, 2019.

[67] Soo Kim. South korea's refugee debate eclipses a deeper, more fundamental question. *The Hill*, 2018. https://thehill.com/opinion/international/395977-south-koreas-refugee-debate-eclipses-a-deeper-more-fundamental-question.

[68] Jin-kyu Kang. Yemeni refugees become a major issue on jeju. *Korea JoongAng Daily*, 2018. http://koreajoongangdaily.joins.com/news/article/article.aspx?aid=3049562.

[69] Nathan Park. South korea is going crazy over a handful of refugees. *Foreign Policy*, 2018. https://foreignpolicy.com/2018/08/06/south-korea-is-going-crazy-over-a-handful-of-refugees/.

[70] Bo Seo. In south korea, opposition to yemeni refugees is a cry for help. *CNN*, 2018. https://edition.cnn.com/2018/09/13/opinions/south-korea-jeju-yemenis-intl/index.html.

[71] Kersten Diers, Fanny Weber, Burkhard Brocke, Alexander Strobel, and Sabine Schönfeld. Instructions matter: a comparison of baseline conditions for cognitive emotion regulation paradigms. *Frontiers in psychology*, 5:347, 2014.

[72] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995.

[73] Susan Kemper and Aaron Sumner. The structure of verbal abilities in young and older adults. *Psychology and aging*, 16(2):312, 2001.

[74] Jiahong Yuan, Mark Liberman, and Christopher Cieri. Towards an integrated understanding of speaking rate in conversation. In *Ninth International Conference on Spoken Language Processing*, 2006.

[75] Cheryl Smith Gabig. Mean length of utterance (mlu). *Encyclopedia of autism spectrum disorders*, pages 1813–1814, 2013.

[76] AC Graesser, Bethany McDaniel, Patrick Chipman, Amy Witherspoon, Sidney D'Mello, and Barry Gholson. Detection of emotions during learning with autotutor. In *Proceedings of the 28th annual meetings of the cognitive science society*, pages 285–290. Citeseer, 2006.

[77] Shazia Afzal and Peter Robinson. Natural affect data—collection & annotation in a learning context. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[78] Sidney K D'Mello, Blair Lehman, and Natalie Person. Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education*, 20(4):361–389, 2010.

[79] Sidney K D'Mello. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*, 7(2):136–149, 2015.

[80] Linda J Levine and Martin A Safer. Sources of bias in memory for emotions. *Current Directions in Psychological Science*, 11(5):169–173, 2002.

[81] Martin A Safer, Linda J Levine, and Amy L Drapalski. Distortion in memory for emotions: The contributions of personality and post-event knowledge. *Personality and Social Psychology Bulletin*, 28(11):1495–1507, 2002.

[82] Heather C Lench and Linda J Levine. Motivational biases in memory for emotions. *Cognition and emotion*, 24(3):401–418, 2010.

[83] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations, April 2020.

[84] Ricardo A Calix, Sri Abhishikth Mallepudi, Bin Chen, and Gerald M Knapp. Emotion recognition in text for 3-d facial expression rendering. *IEEE Transactions on Multimedia*, 12(6):544–551, 2010.

[85] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE, 2012.

[86] Ruifeng Xu, Tao Chen, Yunqing Xia, Qin Lu, Bin Liu, and Xuan Wang. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, 7(2):226–240, 2015.

[87] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[88] Rosalind W Picard. Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3575–3584, 2009.

[89] Peter Washington, Catalin Voss, Aaron Kline, Nick Haber, Jena Daniels, Azar Fazel, Titas De, Carl Feinstein, Terry Winograd, and Dennis Wall. Superpowerglass: a wearable aid for the at-home therapy of children with autism. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(3):1–22, 2017.

[90] Hendrik P Buimer, Marian Bittner, Tjerk Kostelijk, Thea M Van Der Geest, Abdellatif Nemri, Richard JA Van Wezel, and Yan Zhao. Conveying facial expressions to blind and visually impaired persons through a wearable vibrotactile device. *PloS one*, 13(3), 2018.

[91] Cynthia Breazeal. Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155, 2003.

[92] Dong-Soo Kwon, Yoon Keun Kwak, Jong C Park, Myung Jin Chung, Eun-Sook Jee, Kyung-Sook Park, Hyoung-Rock Kim, Young-Min Kim, Jong-Chan Park, Eun Ho Kim, et al. Emotion interaction system for a service robot. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 351–356. IEEE, 2007.

[93] Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. Improving automotive safety by pairing driver emotion and car voice emotion. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1973–1976, 2005.

[94] Florian Eyben, Martin Wöllmer, Tony Poitschke, Björn Schuller, Christoph Blaschke, Berthold Färber, and Nhu Nguyen-Thien. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in human-computer interaction*, 2010, 2010.

[95] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

[96] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[97] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks*, pages 97–108. IEEE, 2011.

[98] Jaclyn Ocumpaugh. Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*, 60, 2015.