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

A Preprint

Cheulyoung 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

A multimodal dataset, namely K-EmoCon, of continuous emotions collected in naturalistic conversations is presented here. K-EmonCon consists of data from 32 participants debating on a social issue for approximately 10 minutes in pairs. The data sources include brainwaves (electroencephalogram), skin conductance (electrodermal activity), body temperature, heart rate (photoplethysmogram and electrocardiogram), blood volume pulse (PPG), and tri-axial acceleration. Videos from the frontal face view and first-person camera view are also included, so to capture facial expressions and gestures above the waist. Participants rated their own emotions (1st-person perspective label) and their partners' emotions (2nd-person perspective label) watching the video recording of their corresponding debate. Collected emotions include arousal-valence, five emotions related to the subjective stress level, and 13 educationally relevant emotions. Statistical analysis results in aggregated distributions of labels and comparative plots of individual variance in emotions. Overall, the proposed K-EmoCon allows, for the first time, the investigation of mismatch in emotion perception during social interactions, with potential use in facilitating emotional communication and development of assistive tools for social-emotional skills.

1 Background & Summary

Automatic emotion recognition is an ongoing challenging problem, as an emotion is a transitory phenomenon without a scientifically agreed definition, subject to social masking, individual randomness, and cultural variability while embedded in a noisy environment [1]. A high-quality database of labeled emotions is an integral component of emotion recognition research, and researchers must decide the type of emotion label that suits their research goal. In particular: (1) *1st-person labels*: how an individual subjectively perceives his/her emotions, (2) *2nd-person labels*: how an observer with the contextual knowledge of the situation which produced emotions perceives an emotional display, and (3) *3rd-person labels*: how an observer without the contextual knowledge perceives an emotional display. 1st-person labels (which can be further broken down into felt-sense labels and self-report labels [2]) are commonly used in systems that provide adaptive responses based on the emotional state of the user, i.e., intelligent tutoring systems [3] and smart

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

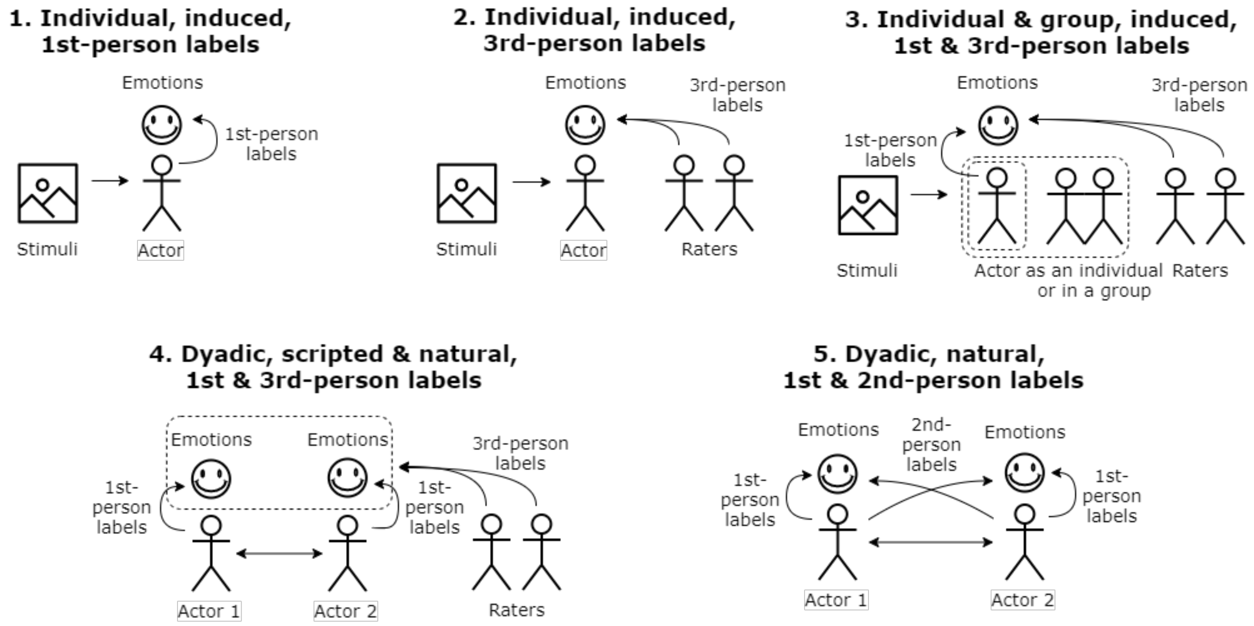

Figure 1: A conceptual diagram of K-EmoCon data sources showing five types of multimodal emotion recognition databases. Five database types are (1) databases of stimuli induced emotions from subjects in isolation with 1st-person labels, (2) databases of stimuli induced emotions from subjects in isolation, but with labels collected from external observers (3rd-person), (3) databases of induced emotions from a subject(s) as an individual or in a group, 1st-person labels with 3rd-person labels from external raters, (4) databases of naturalistic emotions collected in dyadic conversation with 1st-person labels and 3rd-person labels from external raters without the contextual knowledge, and (5) databases of naturalistic emotions collected in dyadic conversation with 1st-person labels and 2nd-person labels from conversation partners with the contextual knowledge of the interaction. K-EmoCon is in the fifth category and is unique as it is the only emotion recognition corpus of its kind to the best of our knowledge.

home devices [4]. 2nd and 3rd-person labels, on the other hand, are commonly used in systems that inform the user of the possible perception of his/her emotional display by others, i.e., an interview training system for job seekers [5] and assistive systems for individuals with an emotion perception deficiency, i.e., children with an autism spectrum disorder [6].

While there exist many emotion-related datasets, our comparative analysis of emotion recognition corpora (see Table 1) indicates no dataset, in our awareness, accommodates investigation of emotions in natural interaction, while comprising information from multiple modalities, including both explicit and implicit channels of human communication. Most datasets do not concern for the individual differences in people's capacity to convey and interpret emotions. The individual variability makes emotions susceptible to misrepresentation and misinterpretation and, further, influences the emotions of people involved in an interaction. For example, MELD, IEMOCAP, and SEMAINE are datasets gathered in a conversational context, but they do not fully capture the entire spectrum of information present in naturalistic conversations [7, 8, 9]. The majority of emotion recognition datasets also collected emotions from individuals when being isolated in a laboratory setting induced with artificial stimuli [10, 11, 12, 13, 14, 15]. This approach is convenient and allows constructing a dataset with a balanced distribution of emotions but neglects the social aspect of emotions [16, 17] and that human affect and associated emotions may undergo rapid changes, unlike mood, especially in the wild [18, 19]. For datasets based on physiological sensing [10, 11, 12, 13, 14, 15], using medical-grade equipment allows collecting data of fine granularity that may result in a more accurate emotion recognition model. Nonetheless, such elaborative data is not available outside specialized laboratories and hospitals as medical-grade equipment is immobile and unaffordable. Therefore, the use of a lightweight and commercially available sensing devices is pertinent in the scenario that requires real-time emotion recognition.

To address the limitations of existing emotion recognition corpora and provide to the emotion research community a new corpus collected in the context of dynamic social interactions, we constructed the K-EmoCon, a multimodal dataset based on physiological sensing for continuous emotion recognition in the context of naturalistic conversations. Over approximately three months (January to March 2019) in Daejeon, South Korea, 32 KAIST undergraduate and graduate

Table 1: A list of multimodal emotion recognition corpora, mainly conversational or based on physiological sensing or both.

Name (year)	Size (N)	Modalities	Spon. vs. posed	Natural vs. induced	Annotation method	Label type	Context
IEMOCAP (2008) [7]	10	Visual (face, gesture), Audio, Text	Both	Both	Turn-based	1st, 3rd	Dyadic
SEMAINE (2011) [8]	150	Visual (face), Audio, Text	Spon.	Induced	Trace-style continuous	3rd	Dyadic
MAHNOB-HCI (2011) [10]	30	Visual (face, eye gaze), Audio, Physiological	Spon.	Induced	Per stimuli	1st	Individual
DEAP (2012) [11]	32	Visual (face), Physiological	Spon.	Induced	Per stimuli	1st	Individual
DECAF (2015) [12]	30	Visual (face), Physiological	Spon.	Induced	Per stimuli	1st	Individual
ASCERTAIN (2016) [13]	58	Visual (face), Physiological	Spon.	Induced	Per stimuli	1st	Individual
DREAMER (2017) [14]	23	Physiological (EEG, ECG)	Spon.	Induced	Per stimuli	1st	Individual
AMIGOS (2018) [15]	40	Visual (face, body), Audio, Physiological	Spon.	Induced	Per stimuli	1st, 3rd	Individual, Group
MELD (2018) [9]	7	Visual (face), Audio, Text	Both	Both	Turn-based	3rd	Dyadic, Group
K-EmoCon (2019)	*32*	*Visual (face, gesture), Audio, Physiological*	*Spon.*	*Natural*	*Interval-based continuous*	*1st, 2nd*	*Dyadic*

students of mixed nationalities participated in 16 debate sessions, wherein each session a randomly selected pair of participants debated on a socially controversial issue. As shown in Figure 1, K-EmoCon is a new type of dataset that includes both 1st and 2nd-person perspective labels.

2 Methods

2.1 Dataset design

The intention in the design of K-EmoCon was to provide to the emotion recognition research community a corpus for the recognition of continuous emotions in naturalistic conversations that also provides physiological information of speakers. While there exist multimodal conversational datasets, K-EmoCon solely accommodates emotion recognition with physiological signals and captures full contextual information of dyadic conversations via a matching pair of 1st and 2nd-person perspective labels given an emotional display. In that regard, the choice of semi-structured debate on a societal issue as the context for data collection was critical, since it allows participants to engage in a genuine conversation, while providing a contained but not too constrained setting, sufficient to capture detailed data on conversations.

The use of portable, wearable, wireless, low-cost, and commercially available sensors to detect physiological signals was a choice to collect data that applies to everyday scenarios, unlike other emotion recognition databases, which used medical-grade equipment collecting data of high granularity, yet of low reusability. Therefore, K-EmoCon data can be used to test the viability of applying results from recent emotion recognition research to domestic use cases, such as wearable devices, and to support comfortable communication of emotions between close acquaintances like friends and family.

Collected data include video recordings from frontal face view and first-person camera view, which allow the analysis of facial expressions and gestures in the upper torso, including heads, hands, and arms. The K-EmoCon dataset is multimodal, comprised of visual information (facial expressions and gestures), audio information (utterances), EEG, peripheral physiological signals recorded during debate sessions, and thus providing a plethora of channels for recognizing emotions. Commonly used affective dimensions, such as arousal and valence, were collected as they are prevalently appearing measures of emotion in the literature [20]. Valence is positive or negative affectivity, whereas arousal measures how calming or exciting the information is. Furthermore, five emotion labels of cheerful, happy, angry, nervous, and sad were collected, which describe a subjective state of psychological stress [21], along with educationally-relevant emotion categories, which can estimate the attention level of a subject during the debate [34, 22].

The resulting K-EmoCon dataset consists of the following: (1) biosignal sensor data obtained from approximately 10 minutes long debate sessions for each participant, (2) 1st-person and 2nd-person perspective emotion labels participants recorded while watching video recordings of themselves and their respective debate partners, and (3) audio recordings of each debate session, and (4) body, hand, and facial keypoints extracted from video recordings of debate with CMU OpenPose [23]. All participants were required to speak in English during the debate.

2.2 Data collection setup

All data collection sessions were conducted in laboratory settings with controlled temperature and illumination. The setup of the data collection environment is shown in Figure 2.

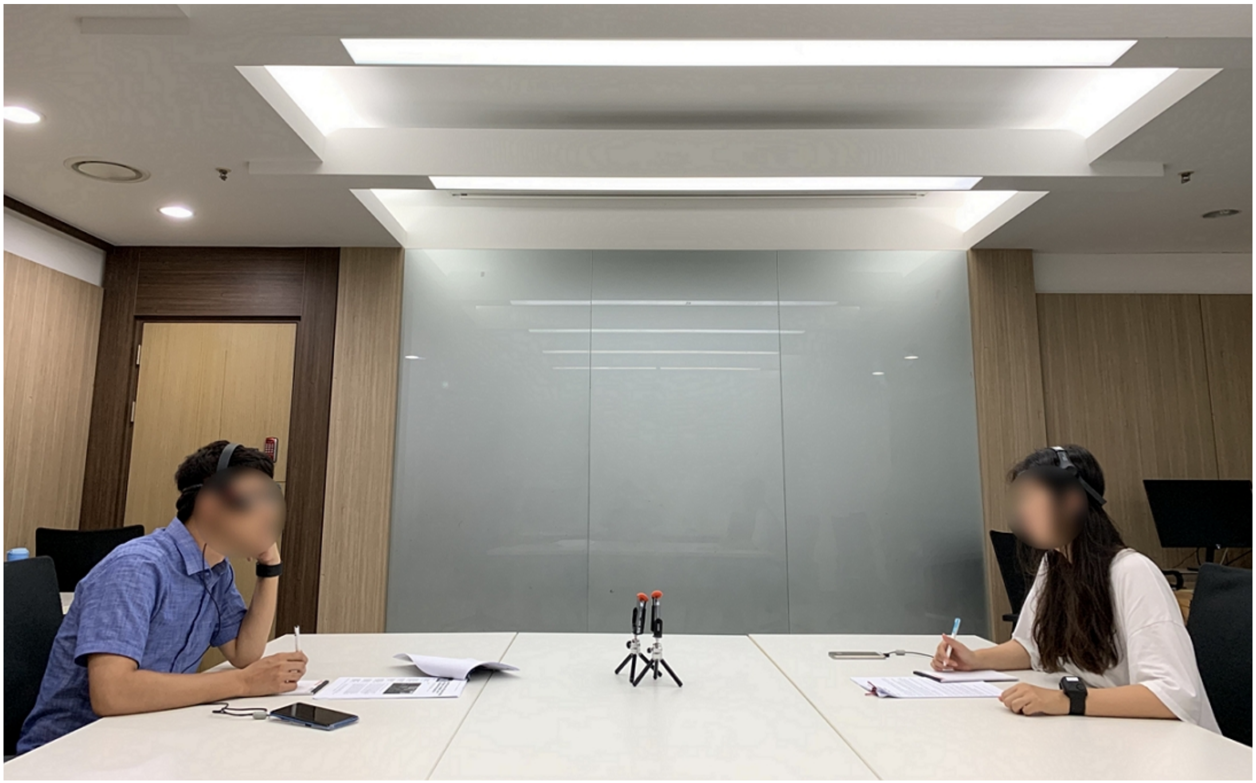

Figure 2: A pair of participants are sitting at a table facing each other with a sufficient distance in between. Two smartphones on tripods were placed in the middle of the table, facing each participant recording facial expressions and motion in the upper torso.

Two participants sat across a table facing each other with a sufficient distance in between for comfortable interaction (as shown in Figure 2). Samsung Galaxy S7 smartphones mounted on a tabletop tripod were placed on the table, facing each participant to capture facial expressions and movements in head, arms, and hands. Participants also wore a LookNTell head-mounted camera that recorded a video at a first-person camera view. Each participant wore a set of sensors shown in Figure 3. An Empatica E4 wristband (`https://empatica.app.box.com/v/E4-User-Manual`) included photoplethysmography (PPG), 3-axis acceleration, body temperature, and electrodermal activity (EDA) sensors. Heart rate and the inter-beat interval (IBI) was later derived from BVP measured by a PPG sensor. However, as PPG alone is vulnerable to motion, we supplemented the measurement of the heart rate with a Polar H7 heart rate monitor (`https://support.polar.com/en/support/H7_heart_rate_sensor`), which uses an electrocardiogram (ECG) sensor. A NeuroSky MindWave headset (`https://store.neurosky.com/pages/mindwave`) was used to collect Electroencephalogram (EEG) signals, and the headset uses two dry sensor electrodes: one on the forehead (fp1 channel-10/20 system at the frontal lobe) and one on the left earlobe (reference).

The aforelisted signals were collected with an Android application developed to collect and store data in an easily transportable format. The app was installed on smartphones and received sensor readings from brainwave headsets and

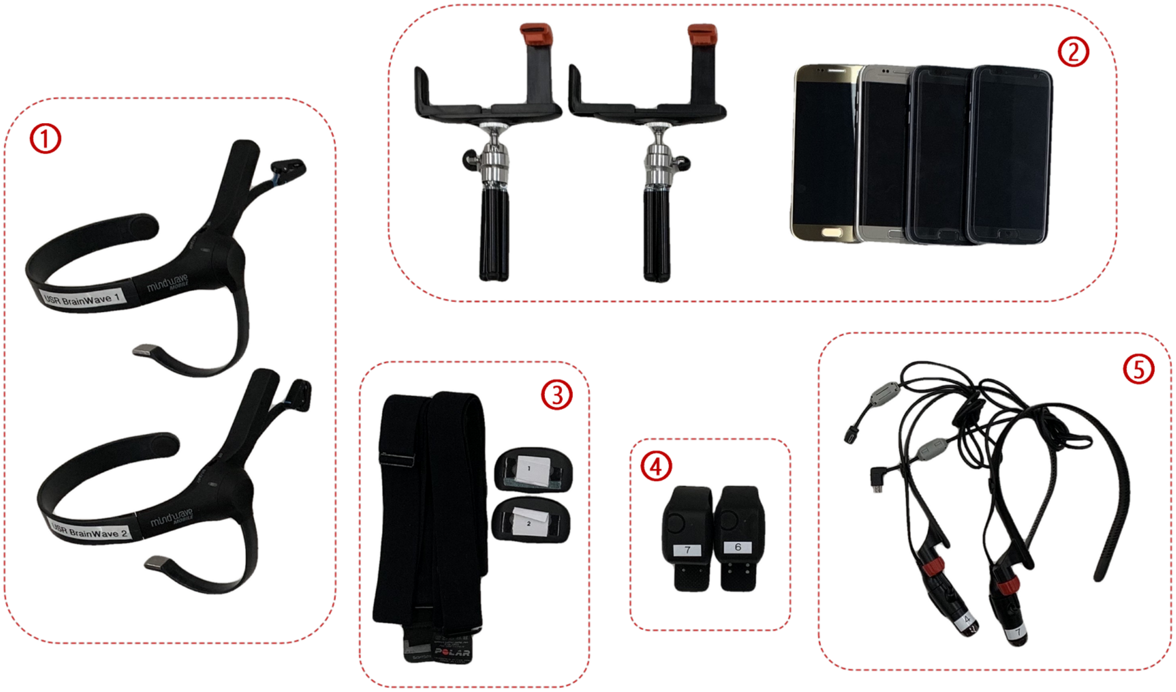

Figure 3: A set of devices used for data collection: 1) NeuroSky MindWave headset for EEG, 2) a pair of Samsung Galaxy S7 smartphones with tripods for recording frontal face view videos and another pair of smartphones for storing first-person camera view videos and sensor readings via BLE, 3) Polar H7 heart rate monitor with ECG sensors, 4) Empatica E4 wristbands (3-axial acceleration, PPG, body temperature, and EDA), and 5) LookNTell head-mounted camera for capturing first-person camera view videos.

Polar H7 sensors during a debate session. At the end of the debate, collected data were locally stored in SQLite format. Data from E4 wristbands were stored in a cloud server and later downloaded separately. Table 2 summarizes the signals collected during the debate; the frequency of data collection and the range of collected signals are also included in Table 2, if available, for each signal.

2.3 Participant recruitment

Participants were recruited between January and March of 2019, and a public announcement calling for participants was posted on an online forum accessible to all students at KAIST. The post specified that the data collection session is expected to last for approximately two hours, including the time to wear sensors, debate on Yemeni refugee issues in Jeju Island in South Korea [24] for approximately 10 minutes, and label emotions collected during the debate.

The inclusion criteria for participation were as follows: participants were required to be capable of debating competently in English, preferably at the level of native English speakers. As a verification for their ability to speak the language fluently, participants were required to have over three years of experience in living English-speaking countries, or have taken any one of English speaking tests and achieved a sufficiently high score: TOEIC speaking level 7, TOEFL speaking score 27, or IELTS speaking level 7 or higher. However, participants' actual proficiency in spoken English varied from intermediate to native speakers. The post also specified the minimum compensation of 40,000 KRW (which is approximately 33.60 USD), with the chance to acquire an extra 10,000 KRW (approx. 8.40 USD) for winning a debate. We proposed to pay an extra amount to the winner to incentivize participants to engage in the debate actively.

Once participants signed up, they were provided four supplementary news articles via email. The articles included two neutral articles [25, 26] providing unbiased views on the issue, one in favor of the refugees [28] reprimanding the anti-immigration movement, and one in opposition of the refugees [27] advocating people's nuanced concern on admitting immigrants. We advised the participants to read the articles before the debate to help themselves prepare the debate with a balanced view on the topic and decide their position at the time of the debate without difficulty. We

Table 2: Signals collected during the debate with respective frequencies and ranges.

Collection devices	Collected data	Collection frequency	Signal range [min, max]
Empatica E4 Wristband	3-axial acceleration	32Hz	[-2g, 2g]
	BVP (PPG)	64Hz	-
	EDA	4Hz	[0.01μS, 100μS]
	Heart rate (from BVP)	1Hz	-
	IBI (from BVP)	n/a	-
	Body temperature	4Hz	[$-40\,^{\circ}$C, $115\,^{\circ}$C]
NeuroSky MindWave Headset	Brainwave (fp1 channel EEG)	125Hz	-
	Attention	1Hz	-
Polar H7 HR Monitor	HR (ECG)	2Hz	-

addressed any likelihood of intimacy affecting participants' emotions by randomly assigning participants among the dates they marked available.

In total, we recruited 32 participants for 16 experiment sessions from January to March of 2019. 20 out of 32 participants were males, and their mean age was 23.31, with a standard deviation of 3.38. The oldest participant was aged 36, and the youngest was 19. All participants were either undergraduate or graduate students at KAIST of mixed nationalities and ethnicities.

2.4 Data collection protocol

All data collection sessions were monitored by two KAIST researchers, who later performed as judges for the debate. Upon arrival, we provided participants consent forms, approved by the KAIST Institutional Review Board (IRB), describing the purpose and the procedure of the data collection. In particular, participants were asked to provide written consent indicating that they agree to participate in data collection, allowing the use of personal information collected during the debate, which is the video recording of their faces and speech during the debate, either in full or partial disclosure. We also notified participants that their participation is voluntary, and they can terminate the session at any point during data collection, at their convenience. All data collection was conducted in accordance with the KAIST IRB.

After participants gave their consent and agreed to continue to participate, researchers provided brief instructions regarding the rules of the debate. Taking a loose form of Lincoln-Douglas debate, as who opens the debate decided randomly by a flip of the coin, during the ten-minute duration of the debate, participants took turns to argue in support of their position, with each turn lasting up to two minutes. Researchers intervened at the end of two minutes to ask a currently speaking participant to yield the turn to the opponent. Also, as the purpose of the debate was to collect spontaneous emotions in the context of a naturalistic conversation, participants were explicitly told that they are allowed to interject during an opponent's turn. Participants were asked if they have any preference regarding the side of the argument and were randomly assigned a position if they did not have any preference. The order of debate was also randomly assigned unless one or more participants had a preference in going first. In the case of a tie, the positions and the order were randomly assigned.

After deciding the position and the order of debate, participants were given 15 minutes to prepare their arguments. They were given pen, paper, and the prints of supplementary articles, which they previously received via email, and were allowed to access the Internet. After 15 minutes of preparation, researchers provided an overview of wearable sensors and assisted participants in wearing them (see Figure 4). Participants were instructed to wear E4 on their non-dominant hand, as excessive arm movements may deter the device from accurately recording physiological signals. Researchers assured that the wristband is tightly fastened and that the electrodes were in good contact with the participants' skin. Researchers also helped participants in wearing the NeuroSky EEG headset and the LookNTell head-mounted camera to make sure devices are well fitted on the participants' heads. Researchers manually adjusted the lens of the head-mounted camera, so the view of the camera simulates the participants' perspective. Participants wore the Polar H7 ECG sensor attached to a flexible band under their clothes and placed the sensor above their solar plexus, so the electrodes are in direct contact with the skin.

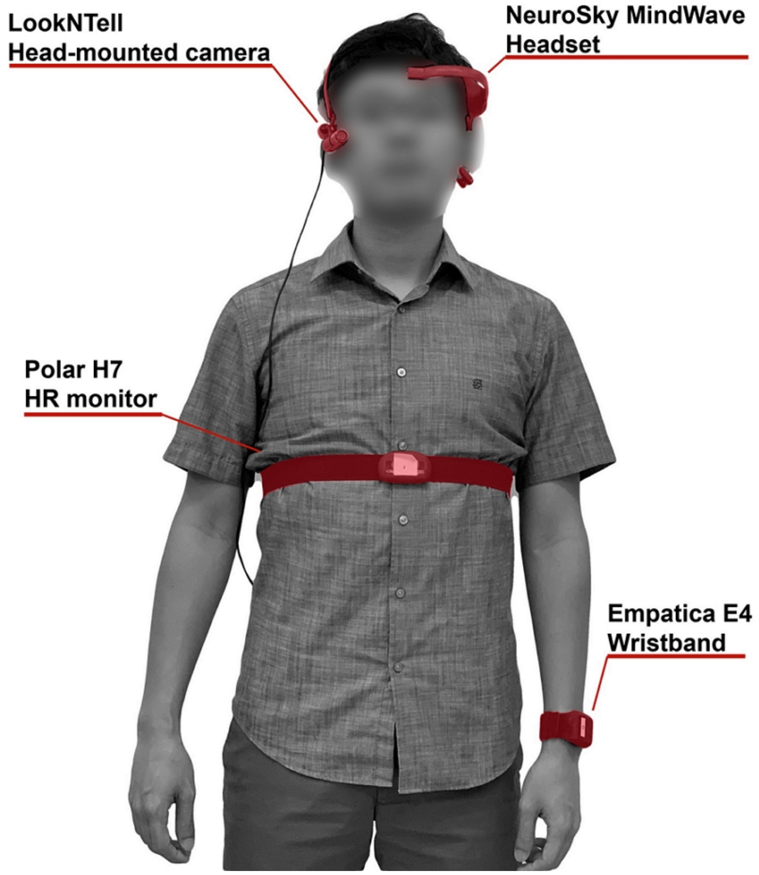

Figure 4: Frontal view of one study participant equipped with data collection devices. Researchers manually adjusted NeuroSky MindWave headset and LookNTell head-mounted camera, so both of them to tightly fit on the subject's head. The head-mounted camera was connected to a smartphone to store the first-person video recording in real-time. Empatica E4 wristband was worn on a non-dominant hand to minimize movement-induced noises, and Polar H7 was worn around the body, such that the electrodes are placed above the solar plexus and in direct contact with the skin. It should be noted that *although the Polar H7 heart rate monitor is shown above the clothes in the image for the presentative purpose, it was worn under clothes during the data collection as electrodes must be in contact with the skin to collect signals.*

At the beginning of each session and before data collection, each participant's physiological signals were acquired while they were watching a video to estimate the baseline level of physiological signals that constitute the neutral state in each participant, which may be subject to individual differences in physiology and psychology. Establishing a baseline for the neutral state is a commonly used approach in the construction of an emotion corpora to account for the individual bias and reduce the bias from emotional states, in the case of the repeated measures.

The exact methodology for the baseline measurement varies across researchers and is contingent upon the goal of the experiment [29]. In stimuli-based experiments involving physiological signals, researchers often ask subjects to watch a stimulus inducing a neutral emotional state and take measurements [10, 11], or record the resting state activity in-between stimuli [12] in the case of multiple measurements. For datasets that measure facial expressions, researchers ask the participants to display a neutral emotion and account for that as the neutral state [7]. For K-EmoCon, participants were requested to watch a *Color Bars* clip used in the work of Gross et al. [30] that is 1 minute 31 seconds long and was reported to induce neutral emotions. With the baseline measurement, researchers also ensured that none of the devices are malfunctioning.

Table 3: Detailed steps for a data collection session. Each session lasted approximately two hours.

Step	Time allocated	Description
Read instructions and sign consent forms	10 min.	Researchers provided participants a form explaining the experiment and collected their written consent for the collection of privacy-sensitive information.
Decide the position and order of the debate	5 min.	Participants were assigned to either the affirmative or negative side and decided the order of debate.
Prepare debate	15 min.	Each participant read provided supplementary materials to prepare respective arguments for the debate.
Wear sensors	10 min.	Researchers explained each sensor to participants and assisted them in wearing the sensors.
Establish baseline	2 min.	Researchers measured a baseline for each participant to establish an estimate for a neutral state.
Overview the debate	5 min.	Researchers explained the rules of the debate and instructed participants that they are allowed to interject and are expected to communicate naturally.
Debate	10 min.	Each participant was given up to 2 minutes per turn and was notified at 60 and 90 seconds before the end of the debate. One debate session lasted approximately 10 minutes.
Label emotions	60 min.	Participants labeled emotions for every 5 seconds interval while watching video recordings of themselves and their partners. Both 1st and 2nd-person perspective labels were collected.

Participants then began the debate at the sign of the researcher moderating the debate. During the debate, the moderator notified participants at 60 and 30 seconds before the end of each turn with a hand sign. When a participant exceeded the allocated time of two minutes during his/her turn, the moderator intervened to ask the participant to pass over the turn. The debate was terminated at the ten-minute mark with some flexibility to allow the last speaker to continue speaking until s/he finished the argument. After the debate, participants assessed their emotions during the debate and their debate partner's emotions. The section that follows explains the emotion labeling process in detail.

2.5 Emotion labeling process

After the debate, participants were each assigned a PC station and were instructed to label emotion during the debate, watching the video recording of themselves and their debate partners. Respectively, participants watched frontal face view recordings of themselves and their partners to label their emotions and partner's emotions. Participants labeled emotions using an excel file where each row corresponds to a non-overlapping 5s window with the first row as 00:00 in the debate, and columns correspond to labeling items. Participants labeled emotions only for the duration of the debate, and researchers manually set the start time and the end time of the debate in videos to obtained synchronized labels from two participants. The rationale to label emotions every 5 seconds is based on findings from linguistics research. It is suggested that the mean length of utterance (MLU) [31], which is the average number of morphemes per utterance, for young adults is about 12 [32]. Then the average speaking rate measured in words spoken per minute for non-native speakers of English, particularly the native speakers of Korean, is approximately 150 [33]. Combining these two numbers results in 4.8 seconds per utterance. Indeed, 5 seconds was neither too long to make the labeling process overly laborious nor too short to be incapable of capturing the dynamic variation of emotions within an utterance. Busso et al. also reported in their work that the average duration of a speaker turn in dialogs was 4.5s [7].

Before participants began labeling, researchers explained individual labeling items to ensure that participants correctly understood their meanings and labeling procedures specific to each item. Researchers also demonstrated labeling for the first 20 seconds of the recording. Overall, participants labeled the total of four label categories consisting of 22 labels (20 emotion labels, including two dummy labels indicating the absence of emotions). The average length of the debate was 676.44 seconds, with a standard deviation of 63.40 seconds. As the labeling was done with an interval of 5 seconds throughout the debate, 8,318 labels were collected for each emotion measured on the Likert scale (arousal, valence, cheerful, happy, angry, nervous, and sad); every emotional display in a 5-second window was labeled twice

Table 4: Collected emotion labels and their descriptions.

Label	Description	Measurement scale
Arousal / Valence	Two affective dimensions from Russell's circumplex model of affect [20]	1: very low / 2: low. 3: neutral / 4: high / 5: very high
Cheerful / Happy / Angry / Nervous / Sad	Emotion states describing a subjective stress state [21]	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 [22]	Choose one
Confrustion / Contempt / Dejection / Disgust / Eureka / Pride / Sorrow / None	Less commonly used BROMP educationally relevant affective categories [34]	Choose one

Table 5: The summary of data collection results and the dataset. M = mean and SD = standard deviation

Data collection summary	
Number of participants	32 (20 males and 12 females)
Participants age	19 - 36 (M = 23.31, SD = 3.38)
Session duration	M = 646.44s, SD = 63.40s
Collected labels and scales	**1 - 5:** Arousal, Valence **1 - 4:** Cheerful, Happy, Angry, Nervous, Sad **Choose one:** Common BROMP affective categories, and less common BROMP affective categories
Collected biosignals (sampling frequency)	3-axis Acc. (32Hz), BVP (64Hz), EDA (4Hz), HR (1Hz), IBI (n/a), Body temperature (4Hz), EEG (fp1, 32Hz), ECG (2Hz)
Dataset summary	
Audio recordings of all debates	344.77 minutes (from 32 participants)
Face, body, and hand keypoints	344.77 minutes (from 32 participants)
Biosignal readings	298.5 minutes (from 28 participants)
Emotion labels	8,318 labels per emotion measured in Likert scales (each label = 5s window, with each window rated twice: by self & partner)

(hence 4,159×2 = 8,318), once by participants themselves (1st-person perspective labels) and again by their respective debate partners (2nd-person perspective labels). Table 4 summarizes labeling categories and emotions labels in detail.

3 Data Records

3.1 Dataset summary

Although audiovisual recordings were collected during debate sessions, they are not included in the dataset as the recordings consist of participants' faces along with speech, which combined can allow the identities of individual participants to be inferred. Instead, the face, body, and hand keypoints estimated from the recordings with CMU OpenPose are included for all participants. All data was clipped and synchronized to include only the portion corresponding to the debate. The resulting dataset includes sensor data of 28 participants (excluding 4 participants due to E4 malfunction), face/body/hand keypoints for 32 participants, audio recordings of all debates, and emotions labels from 1st and 2nd-person perspectives from 32 participants. The following table shows the summary of data collection and the dataset contents.

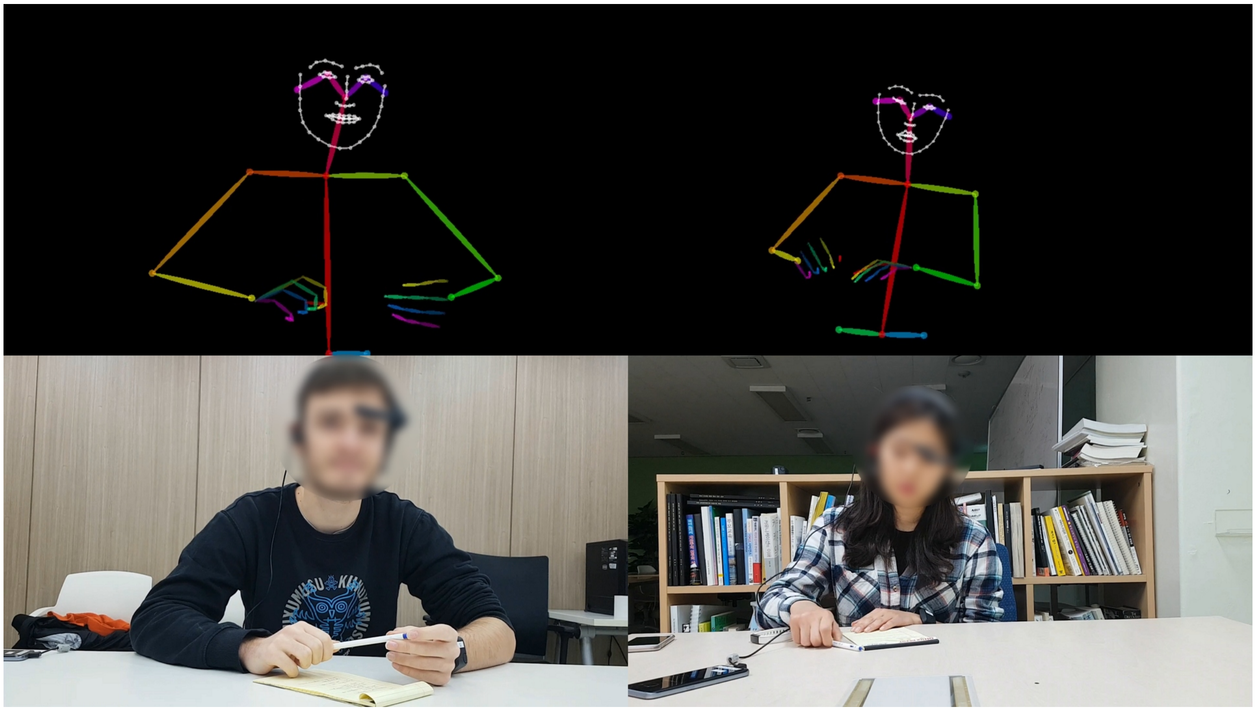

Figure 5: Examples of keypoints extracted with OpenPose overlaid on a black background (upper) from first-person camera view recordings (lower).

E4 data of 4 participants were excluded from the dataset due to an error in the collection device. Some of the physiological signals in the dataset are incomplete due to issues inherent to the device. Parts of IBI data are missing as the internal algorithm of E4 that derives IBI from BVP automatically discards an obtained value if its reliability is below a preset threshold. Also, parts of EDA readings were imprecisely registered for a few sessions, possibly due to poor contact between the device and a participant's skin.

3.2 Synchronization

Timestamps were converted from Korea Standard Time (UTC +9) to UTC +0 to allow the timewise synchronization across all data. Raw data was clipped, and only the part of data corresponding to the debate and the baseline measurement is included in the dataset to protect the privacy of participants.

Audio recordings and keypoints After extracting subclips corresponding to debates from original audiovisual recordings of data collection sessions, audio parts were separated from subclips. Face/body/hand keypoints in the dataset are also only for the duration of debates. Before estimating the keypoints, original recordings were preprocessed to 30 fps and resized to 960×540 from 1920×1080 for faster processing speed.

Biosignal data The data was clipped from the beginning of data collection to the end of the debate. The initial 1.5 to 2 minutes immediately after the beginning of data collection corresponds to baseline measurements for a neutral state. The part after the baseline measurement and before the debate corresponds to debate preparation. The rest of the data includes sensor readings during the debate.

Emotion labels Emotion labels were collected only for the duration of the debate. 20 emotion labels were collected for every 5 seconds window without overlapping and are indexed with the relative time difference from the beginning of the debate in seconds (e.g., 5 seconds, 10 seconds).

Dataset contents

The K-EmoCon dataset [35] is available on the *figshare* data repository. The following section details directories and files in the repository and their contents.

Metadata.tar: includes files with ancillary information about the dataset. Included files are as follows:

1. **subjects.csv**: lists the participant IDs (PID) and three timestamps in UTC +0 for each participant. Each timestamp respectively marks the beginning of the data collection (*initTime*), the start of the debate (*startTime*), and the end of the debate (*endTime*).
2. **data_availability.csv**: shows files available for each participant. For each participant (row), if a certain data file (column) is available, the corresponding cell is marked TRUE, otherwise FALSE.

Recordings.tar: consists of audio recordings of debates and OpenPose output clips of keypoints overlaid on a black background. Included subfolders are as follows:

1. **debate_audio**: contains audio recordings of debate sessions in WAV file format. The start and the end of recordings correspond to *startTime* and *endTime* values in the *subjects.csv* file.
2. **openpose_outputs**: contains clips overlaid with face/body/hand keypoints on a black background (without blending with original images). All clips were resized to 960×540 from original recordings of size 1920×1080 for a faster keypoint estimation, while the frame rate was conserved at 30 fps.

Data.tar: consists of biosignal data and OpenPose keypoints for each participant. The archive file includes the following files:

1. **e4_data.tar.gz**: contains sensor readings collected with an Empatica E4 wristband, including tri-axial acceleration (x, y, and z), blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), inter-beat interval (IBI), and skin temperature for all participants except P2, P3, P6, and P7 due to a malfunction in E4 during data collection.
2. **emotion_labels.tar.gz**: includes two subfolders *emotion_self* and *emotion_other*, each respectively containing 1st and 2nd-person perspective labels, which further contains individual folders for each participant. Emotion labels are stored in a table format (csv file) with a row of labels entered for every 5 seconds interval throughout the debate.
3. **neurosky_polar_data.tar.gz**: consists of brainwave readings collected with a NeuroSky MindWave headset (Attention.csv, BrainWave.csv, and Meditation.csv) and heart rates collected with a Polar H7 sensor (Polar_HR.csv).
4. **openpose_keypoints.tar.gz**: contains face/body/hand keypoints estimated with OpenPose, saved in the JSON format, for each participant excluding P12, P14, and P27 who did not consent to the release of privacy-sensitive data. Each JSON file corresponds to one frame in a clip and has arrays of values formatted as $[x1, y1, c1, x2, y2, c2, \ldots]$ where x and y are the coordinates of a point on a plane of area 960×540 while c is the confidence score between 0 and 1. Refer to `https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md#output-format` for a more detailed explanation of the output file format.

4 Technical Validation

The results of the preliminary analysis on the K-EmoCon dataset are presented here. These include descriptive statistics for the collected emotion labels and discussion on the mismatch between 1st-person perspective labels and 2nd-person perspective labels.

4.1 Descriptive statistics

By plotting the distributions of 1st-person and 2nd-person perspective labels aggregated across participants in parallel, we investigate the following question: *"Q1. Do labels from 1st-person and 2nd-person perspectives show a significant difference in their distributions and do the perception of emotional displays from two different perspectives exhibit bias in any way, and if it does, for which particular emotion?"* The resulting plot shows that distributions of aggregated emotion labels are without much difference across 1st and 2nd-person perspectives.

Figure 6 shows that the distribution of both 1st-person and 2nd-person labels are centered around a neutral value (3 for arousal and valence, and 1 for cheerful, happy, angry, nervous, and sad). Results of Chi-Square goodness of fit tests with 1st-person labels as expected frequencies and 2nd-person labels as observed frequencies indicate that there is no significant difference between the observed and the expected values for all emotion labels. This result is unsurprising, as

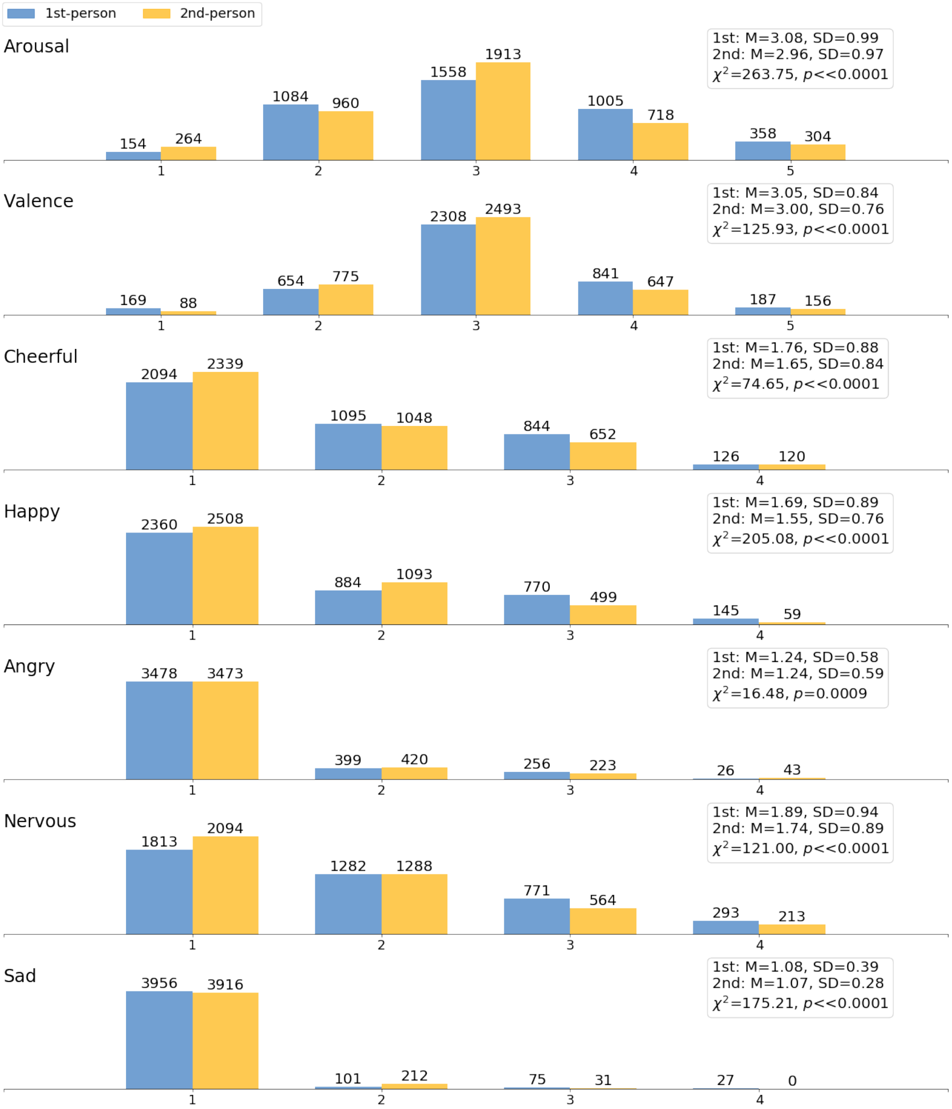

Figure 6: The plots of respective distributions of 1st and 2nd-person perspective labels aggregated across participants, with preliminary statistics (M=mean, SD=standard deviation) and results of Chi-Square goodness of fit test (χ^2=chi-squared test statistic, p=p-value). On the x-axis is the intensity of emotions in integer ranges of 1 - 5 (arousal and valence) and 1 - 4 (cheerful, happy, angry, nervous, and sad), and the height of each bar corresponds to an aggregated frequency for each emotion label.

participants possibly suppressed their emotional expressions given the context of the debate that conventionally requires debaters to base their arguments on logic and evidence, not emotions. Values higher than 1 (very low) for ‘negative’ emotions such as anger and sadness are reported less frequently compared to positive or neutral emotions, including cheerful, happy, and nervous, which again can be attributed to the formal context of the debate scenario. However, the similarity of two aggregated distributions is insufficient to conclude that how participants believe their partners perceive their emotional display (1st-person perspective labels) and how partners do perceive the display (2nd-person perspective labels) indeed match well, given that labels were aggregated across participants. It is necessary to conduct an individual-level analysis to draw a better conclusion on how the perception and interpretation of emotional displays might differ.

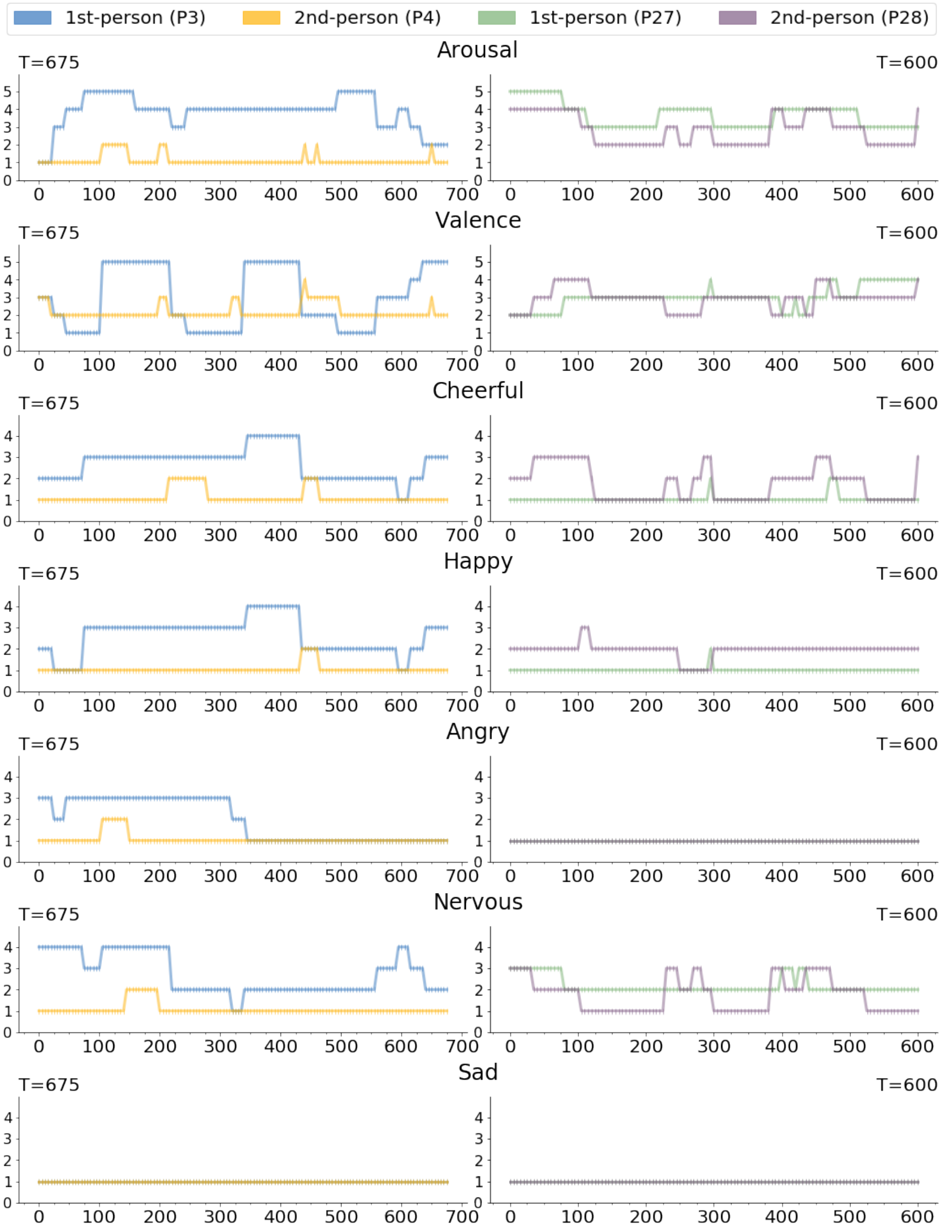

Figure 7: Temporal variance in emotions of two pairs of participants P3 - P4 (on the left column) and P27 - P28 (on the right column), over the duration of the debate.

4.2 Comparison of 1st and 2nd-person labels by individuals

As the 1st and 2nd-person labels are not significantly different when aggregated, we then investigate the following question: *"Q2. Does the level of mismatch between 1st and 2nd-person perspective labels vary across debate pairs, and are some emotions more prone to misinterpretation?"* Indeed, a comparison of 1st and 2nd-person perspective labels at the individual level shows that there exists a mismatch between the two, and its amount varies across pairs. Characteristic examples are shown in Figure 7 with line plots of 1st and 2nd-person emotion labels of two debate pairs: P3 - P4 (left column) and P27 - P28 (right column).

The disagreement between 1st-person and 2nd-person labels for P3 is apparent in Figure 7. Plots of P3's emotions on the time axis (left) show a large mismatch in 1st and 2nd-person labels. While the values of 1st-person (self-reported) labels show large variance over time, the variance in 2nd-person (reported) labels are subdued, possibly because the observer (partner) understated the intensity of perceived emotional displays. On the contrary, plots of emotion label values for P27 (right) show some agreement between two labels at several points; particularly for arousal, it is visible that 1st and 2nd-person label values follow a similar trend. Although minor discrepancy exists for the timing of a change in label values, the magnitude and direction of variance in two line-plots show a decent agreement. A similar observation holds for plots of valence and cheerfulness of P27.

Another interesting observation in the P27's plot is that the agreement in 1st and 2nd-person labels for nervousness is lower compared to the agreement in other emotions. Although not shown in the figure, a low agreement for nervousness was similarly found in the plots of the majority of participants. The same has been observed for sadness and anger as well, despite their rare occurrence. This result possibly indicates that it may be difficult for observers to detect negative emotions, such as anger, nervousness, and sadness, even with contextual knowledge, especially if the source of emotion was averse to candidly express emotions due to contextual factors specific to the setting (e.g., the general inclination to keep composure during the debate).

5 Usage Notes

5.1 Limitations

Contact-base EEG sensors are known to be susceptible to noise - for example, frowning, or eyes-movement might cause peaks in the recorded raw data. Signals collected with other devices may also be subjected to similar systematic errors in collection devices. The context of the debate may have caused participants to suppress their emotional expression, as the display of emotions is commonly regarded undesirable during the debate; hence, there exists more significant disagreement between the felt sense of emotions and perception of emotional displays by the debate partners. Many variables unaccounted during data collection, such as the rapport between debating pair of subjects, the competence in the language, and the familiarity with the debate topic, may have contributed to a significant variance in the level of mismatch between felt and perceived emotions.

References

[1] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43, 2015.

[2] 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. ACM, 2016.

[3] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 379–388. ACM, 2015.

[4] Sidney Fussell. Alexa wants to know how you're feeling today. *The Atlantic*, 2018. `https://www.theatlantic.com/technology/archive/2018/10/alexa-emotion-detection-ai-surveillance/572884/`.

[5] Michelle Fung, Yina Jin, RuJie Zhao, and Mohammed Ehsan Hoque. ROC Speak: Semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1167–1178. ACM, 2015.

[6] 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):112, 2017.

[7] 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.

[8] 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.

[9] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Gerben A Van Kleef. How emotions regulate social life: The emotions as social information (easi) model. *Current directions in psychological science*, 18(3):184–188, 2009.

[17] Gerben A van Kleef, Arik Cheshin, Agneta H Fischer, and Iris K Schneider. The social nature of emotions. *Frontiers in Psychology*, 7:896, 2016.

[18] Stacy C Marsella and Jonathan Gratch. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009.

[19] Panteleimon Ekkekakis. Affect, mood, and emotion. *Measurement in sport and exercise psychology*, 321, 2012.

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

[21] 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.

[22] 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.

[23] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[24] Hyun-ju Ock. Jeju refugee crisis and beyond: Yemeni asylum seekers build life in Korea. *The Korea Herald*, 2019. `http://www.koreaherald.com/view.php?ud=20190217000042`.

[25] 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`.

[26] 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`.

[27] 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`.

[28] 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/`.

[29] 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.

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

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

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

[33] 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.

[34] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 383–390. ACM, 2004.

[35] 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. *figshare*. `https://figshare.com/s/ad75b9ea7351a01cdba1`.