scientific data

OPEN

DATA DESCRIPTOR

EAV: EEG-Audio-Video Dataset for Emotion Recognition in Conversational Contexts

Min-Ho Lee[1], Adai Shomanov[1], Balgyn Begim[1], Zhuldyz Kabidenova[1], Aruna Nyssanbay[1], Adnan Yazici[1] & Seong-Whan Lee[2] ✉

Understanding emotional states is pivotal for the development of next-generation human-machine interfaces. Human behaviors in social interactions have resulted in psycho-physiological processes influenced by perceptual inputs. Therefore, efforts to comprehend brain functions and human behavior could potentially catalyze the development of AI models with human-like attributes. In this study, we introduce a multimodal emotion dataset comprising data from 30-channel electroencephalography (EEG), audio, and video recordings from 42 participants. Each participant engaged in a cue-based conversation scenario, eliciting five distinct emotions: neutral, anger, happiness, sadness, and calmness. Throughout the experiment, each participant contributed 200 interactions, which encompassed both listening and speaking. This resulted in a cumulative total of 8,400 interactions across all participants. We evaluated the baseline performance of emotion recognition for each modality using established deep neural network (DNN) methods. The Emotion in EEG-Audio-Visual (EAV) dataset represents the first public dataset to incorporate three primary modalities for emotion recognition within a conversational context. We anticipate that this dataset will make significant contributions to the modeling of the human emotional process, encompassing both fundamental neuroscience and machine learning viewpoints.

Background & Summary

Emotion Recognition in Conversation (ERC) is an emerging field focused on empowering machines to comprehend human emotions. Successful deployment of ERC technologies has the potential to lead to more engaging, personalized, and human-like interactions with artificial intelligence systems[1,2].

It is important to acknowledge the complexity of emotions and the considerable variation in their manifestation across individuals[3]. Moreover, emotion is a psycho-physiological process ignited by both conscious and unconscious perceptions, and the ability to perceive emotional states significantly impacts human interactions, particularly in conversational scenarios[4].

Numerous studies have delved into the intricacies of emotion recognition, focusing on different modalities such as acoustic, visual, or physiological information derived from individuals. Each modality presents its distinct challenges and benefits, thus leading to a multifaceted approach in the pursuit of advanced emotion recognition systems.

Indeed, facial expressions and speech signals are commonly utilized modalities in emotion recognition due to their relatively straightforward capture methods. They provide intuitive and abundant insights into an individual's emotional state, rendering them a valuable resource in the interpretation and classification of emotions.

Conversely, the EEG modality can intricately capture the outcomes of complex associations within the brain's interconnected activities, providing a more direct measurement of emotions when compared to audiovisual data. This potential for accuracy enhances its performance in emotion classification, particularly among participant groups who may have difficulty accurately expressing their emotions. This approach directly assesses brain activity, bypassing external behavioral displays and, as a result, can yield more genuine insights into an individual's emotional state, irrespective of their capacity or inclination to externally manifest their emotions.

The experimental setup for behavioral emotion environment is more challenging as it requires active participation and engagement from multiple individuals[5]. In EEG-based studies, passive tasks have been

[1]Nazarbayev University, Department of Computer Science, Astana, 010000, Republic of Kazakhstan. [2]Korea University, Department of Artificial Intelligence, Seoul, 02841, Republic of Korea. ✉e-mail: sw.lee@korea.ac.kr

predominantly employed to evoke emotions, such as watching videos or images, listening to music, or recalling memories[6–12]. Only a limited number of studies have explored EEG measurements in a conversational context. Saffaryazdi *et al*. introduced the PEGCONV[13] in which they recorded 16 channels of EEG along with physiological activities (PPG, GSP) during spontaneous conversations. Park *et al*. presented the K-EmoCon dataset[14], which includes single-channel EEG data, peripheral physiological signals, and audiovisual recordings.

In the domain of audiovisual measurements, a comprehensive collection of datasets has been established, featuring larger-sized conversational settings across various scenarios[15–25]. These dialogues were skillfully performed by professional actors[15,17,18,20,22], or they involved participants engaging with examiners[16,19] in predefined interaction scenarios designed to elicit specific target emotions. Some portions of this dataset were generated by extracting conversation segments from films[23], TV series[24], or video clips[25].

Emotion annotation presents one of the most demanding challenges in this context, and the strategies employed may vary based on the experimental conditions. Self-report annotation can be the most reliable means of labeling; however, individuals often mask their true emotions, or they may struggle to accurately identify their own emotions. This challenge intensifies in spontaneous and natural settings, where emotions can fluctuate unpredictably during conversations. In response to this challenge, Park *et al*.[14] incorporated three types of annotations: those from the participant, their conversation partner, and observers, to ensure comprehensive assessments from different viewpoints.

Emotion recognition systems have predominantly relied on a single primary modality from the EEG/Audio/Video spectrum. For instance, a study reported by Soleymani *et al*.[6] focused on EEG accuracy alone, despite also recording audio and video data. Although they measured EEG and audiovisual data simultaneously, their analysis predominantly highlighted the performance of one modality[13,14]. The study by Koelstra *et al*.[12] also measured EEG and visual data; however, it was not appropriate to heavily utilize the visual data since the paradigm was designed for a passive task like watching video clips, where visual cues do not dynamically engage the participant. Table 1 provides a summary of relevant papers addressing emotion recognition using multiple modalities, which are pertinent to our study. It includes key information such as the database title, primary modalities, language of the audio/video stimuli, subjects' elicitation method, and types of stimuli. Interestingly, no prior research has ventured into the concurrent use of EEG alongside facial or audio signals for emotion recognition within conversational contexts.

The aim of this study is to introduce a multimodal dataset designed to facilitate a more comprehensive understanding of human emotional behavior. This is achieved by analyzing both direct neuronal signals through electroencephalography (EEG) and indirect cues obtained from audiovisual sources. To create this dataset, we established a conversational setting in which participants voluntarily selected scripts that were significantly associated with the target emotions during a screening session. Subsequently, these participants engaged in interactions with a cue-based conversation system, expressing their emotions through facial cues and voice modulations. In total, 42 participants took part in this study, and their responses were captured using a 30-channel EEG and audiovisual recordings. The trials for each emotion were thoughtfully balanced to ensure a precise representation across a wide spectrum of emotions.

To enhance its suitability for classification tasks, we pursued the following objectives: 1) Attaining balanced class distributions. 2) Structuring data in suitable formats. 3) Maintaining consistent conversation duration in each trial. Given the use of scripted dialogues, the conversations were not entirely spontaneous but were guided by cues to ensure uniformity.

Our multimodal approach and the dataset we have developed hold the promise of making significant contributions to several research directions: 1) Techniques for integrating data from multiple sources[26–28]; 2) Analysis of disparities in emotions between observable human behavior and underlying brain functions[29,30]; 3) Examination of time-variant or contextual emotion analysis[31]; 4) Synthesis of the facial or vocal outputs[32,33]; and 5) Exploration of emerging machine learning topics, such as multimodal contrastive learning[34,35], generative models[36], or knowledge transfer learning[31].

Methods

The following section outlines the methodological framework employed in conducting a multimodal experiment for emotion recognition. The experiment entailed exposing participants to emotional stimuli through three distinct modalities: audio, visual (via camera recordings), and neurophysiological data (collected via EEG). This study focused on the recognition of five primary emotions: anger, happiness, sadness, calmness, and a neutral state.

Each dataset within these modalities adheres to a consistent format, and class distributions are balanced. In the controlled cue-based conversation scenario, participants deliberately selected scripts that strongly evoked specific emotions, allowing us to infer that our class labels reliably represent their genuine emotional states.

Ethic statement. This study was conducted with approval from the Institutional Research Ethics Committee of Nazarbayev University (NU-IREC 777/09102023) and the Institutional Review Board of Korea University (KUIRB-2021-0248-02). All participants provided written informed consent prior to their involvement and were fully informed about the nature of the research, including the recording of audio, video, and EEG data. They were also made aware of the data distribution, which is restricted to research purposes under a strictly supervised Data Usage Agreement by the IREC. Furthermore, stringent protocols were implemented to ensure the protection and confidentiality of all participant data, in accordance with established ethical guidelines for research involving human subjects.

Participants. The study enlisted a cohort of 42 individuals from Nazarbayev University, including both students and members of the general population. These participants were proficient English speakers who had either lived in or studied in English-speaking countries. The age range of the participants fell between 20 and 30 years.

Database	Primary Modalities	Language	Subjects	Elicitation Method	Types
MAHNOB-HCI[6]	EEG, Face, Audio	—	27 subjects	Videos/Pictures	S, I
SEED-IV[11]	EEG and EM	—	15 subjects	Videos	S, I
DREAMER[10]	EEG & ECG	—	23 subjects	Movies	S, I
MPED[9]	EEG, GSR, RR, ECG	—	23 subjects	Videos	S, I
ASCERTAIN[8]	EEG, ECG and GSR	—	58 subjects	Videos	S, I
AMIGOS[7]	EEG, GSR and ECG	—	40 subjects	Movies	S, I
DEAP[12]	EEG, PS, Face	—	32 subjects	Music videos	S, I
IEMOCAP[15]	Face, Speech, Head	English	10 professional actors	Conversations	S, I
SEMAINE[16]	Face, Speech	English	150 subjects	Conversations	S, I
NNIME[19]	Audio, Video, ECG	Chinese	44 subjects	Conversations	P, N
RAVDESS[20]	Audio, Video	English	24 professional actors	Speech, Song	P, I
BAUM-1[21]	Face, Speech	Turkish	31 subjects	Images/Videos	S, I
SAVEE[26]	Face, Speech	English	4 subjects	Videos/Texts/Pictures	S, I
K-EmoCon[14]	Face, Speech, 1ch EEG	Korean	32 subjects	Conversations (Debate)	S, N
PEGCONV[13]	EEG, GSR, PPG	English	23 subjects	Conversations	S, N
EAV (ours)	EEG (30ch), Audio, Video	English	42 subjects	Conversations	S, I

Table 1. The performance of emotion recognition was evaluated in each modality (EEG, Visual, and Audio data). Specifically, accuracy and AUC scores for 5 balanced classes were calculated across individual subjects.

A high level of language proficiency was a requirement for all participants, given the nature of the experiment, which involved engaging in dialogues and discerning emotional nuances within the English language.

Participants were provided with prior information about the screening and primary experiment dates, allowing them to schedule their involvement accordingly. Before participating, all participants offered informed consent and were given a comprehensive explanation of the study's objectives and procedures. The importance of their right to withdraw from the study at any point was emphasized to ensure ethical compliance throughout the research.

Experiment Protocol. The primary objective of this study is to investigate five distinct emotional classes, which are represented as follows: A for anger, H for happiness, S for sadness, C for calmness, and N for a neutral emotional state. These emotional classes can be further categorized into two overarching dimensions: positive and negative valence, as well as high and low arousal levels.

Our study involved the creation of conversation scenarios in which human participants engaged in emotional dialogues with a conversational system. These dialogues encompassed multiple iterations of listening and speaking interactions. Initially, the dialogues were generated using OpenAI's ChatGPT and were subsequently meticulously reviewed and edited by human researchers to ensure emotional clarity and appropriateness for the study.

It is important to note that the dialogues were intentionally designed to be universally accessible and straightforward, resembling common, everyday conversations. The length of sentences in each interaction was thoughtfully adjusted to ensure they were of an appropriate duration, allowing the speaker to engage in speech for more than 20 seconds.

In our study, the dialogues are structured as paired listen/speak sets. To accomplish this, psychologist reviewed the entire collection of the scripts to ensure that the emotions were correctly induced and to verify that the script did not contain any excessively harsh sentences that could potentially cause adverse effects on the participants.

The experiment was designed to include 100 repeated interactions. In each interaction, participants initially watched a pre-recorded video and then responded to the corresponding dialogue for a duration of 20 seconds. When participants spoke, the dialogue was presented in the central area of the monitor, guiding them to read and deliver the script as instructed (see Fig. 1).

Each iteration commenced with a neutral class dialogue, followed by four interactions corresponding to specific emotional classes, as demonstrated in Table 3. The introduction of a neutral dialogue in the first interaction was intended to establishing a baseline, ensuring that participants initiated the experiment in a neutral emotional state before transitioning to emotionally charged conversations. This approach also contributed to maintaining a balanced distribution of the five emotional class trials, guaranteeing that each class was represented in 20 interactions. In total, our dataset encompasses 20 iterations, comprising a cumulative total of 100 interactions.

Despite providing participants with ample time to rest and return to a neutral emotional state, they reported encountering challenges when transitioning between contrasting emotions, such as moving from anger to happiness. This issue has been documented in previous studies as well. Consequently, we thoughtfully arranged the sequence of emotional classes, strategically placing the 'calm' class between such contrasting emotions to facilitate smoother transitions.

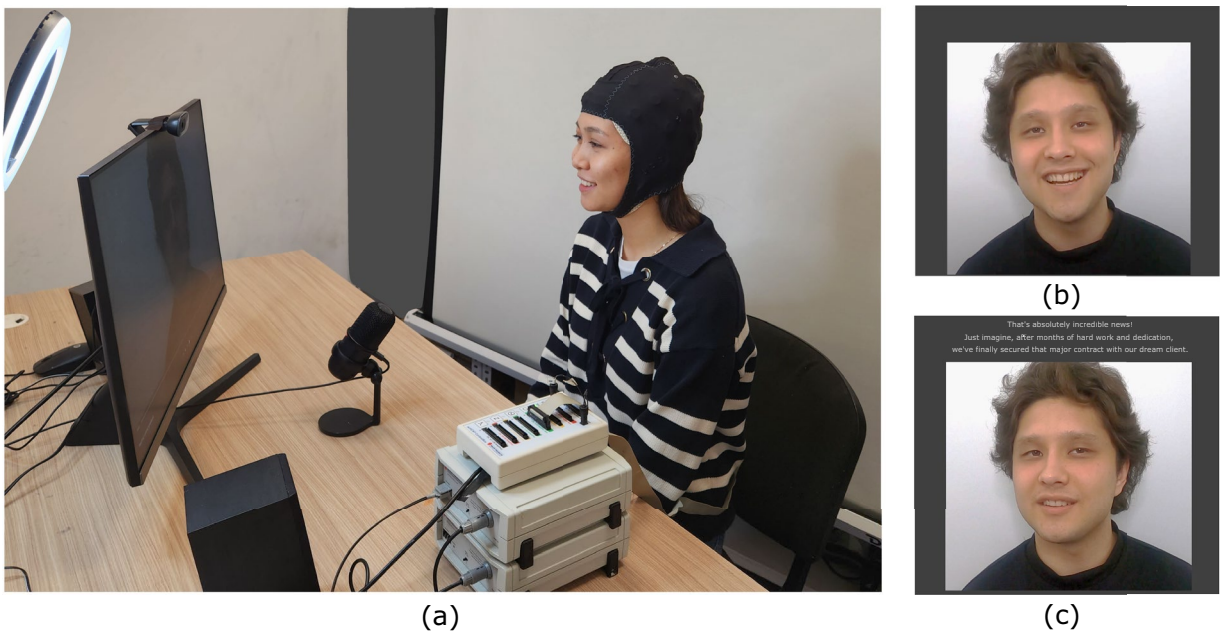


Fig. 1 Illustration of the experimental setup. (**a**) The participants wore a cap for EEG recordings and multimodal data were synchronously recorded, (**b**) Listening condition: a prerecorded video is displayed to the participant, prompting their interaction, (**c**) Speaking Condition: Scripts are provided at the center of the monitor. Participants were encouraged to freely express themselves beyond the confines of the given script. The individual(s) depicted in this figure provided informed consent to have their image openly published.

We employed a pseudo-random class-iteration sequence: [A, A, C, C, S, S, H, A, C, H, H, S, S, A, A, C, C, H, H, S] to optimize the emotional transitions experienced by the participants. In this sequence, each emotional class corresponds to a specific iteration, and the neutral class consistently begins each iteration. This approach serves two purposes: it maintains the participant's current emotional state and facilitates a gradual transition to contrasting emotional classes. To prevent any potential bias resulting from rapid emotional shifts in participants and to ensure data consistency, we applied this pseudo-randomized sequence uniformly to all participants.

We must note that, while the sequence set was the same for all participants, the dialogues within each iteration varied as participants selected different sets of dialogues for each emotional class.

The experiment was carried out in two separate sessions, separated by a one-day interval between them. This approach was adopted to ensure optimal participant engagement and to minimize the effects of fatigue. For a comprehensive overview of the procedures implemented in each session, please refer to Table 2.

Session I: Screening. The primary purpose of the screening session was to familiarize participants with the experiment's overview and the research format. During the first session, participants were presented with a selection of dialogues and tasked with choosing the dialogues that they believed would most effectively elicit and reflect their emotional responses, as detailed in Table 2. A pool of 50 dialogues was offered, with 10 dialogues available per emotional class. Each participant was required to choose five dialogues for each emotional class. Each dialogue consisted of four interactions, encompassing both listening and speaking components.

In previous studies, class labels were assigned after participants had experienced emotions from specific stimuli in a natural setting. This approach presented challenges in achieving consistent and balanced data, as label assignments could vary significantly among individuals. Variability in labels for the same stimulus could introduce conflicts in machine learning models. In contrast, our study provided distinct dialogues for each emotion class. Participants then voluntarily selected the scripts they found most evocative, ensuring that class labels were established prior to the experiment.

Following the selection of dialogues, participants followed an experimental protocol based on their choices. They engaged in a practice session without data recording, allowing them to become familiar with the procedure under the guidance of psychologists. If a chosen dialogue and its associated video failed to elicit the desired emotion, participants were given the option to select a different script.

Although our paradigm did not replicate a naturalistic setting and involved predetermined conversation scripts, there was a potential risk of participants merely reading the provided text without genuine emotional engagement. To mitigate this, participants were explicitly instructed to emotionally align with specific scenarios. Those who did not adequately exhibit this alignment or emotional comprehension were excluded from the study.

Session II: Main Experiment. The second session, which constituted the main experiment, spanned approximately 2 hours. Before commencing the experiment, we assessed each participant's psychological condition. If

Session 1: Screening		
Procedure	**Time (min.)**	**Description**
Introduction to experiment	10	We provided an overview of the experiment, including purpose of the study and experimental procedures.
Informed consent form	10	The participant read the consent form, which included information on the study's aims and procedures
Scripts selection	30	The participant selected the scripts for five emotion classes, the scripts that elicited the strongest emotional response
Preliminary experiment	30	The participant was asked to practice the experiment paradigm with a few scripts to ensure their understanding of the task
Feedback	5	Participants were asked to provide feedback regarding their emotional response, or any uncomportable feelings

Session 2: Recording		
Pre-experiment assessment	5	Prior to the experiment, the participant's condition was evaluated
Calibrating equipment	5	We placed a computer screen, camera, and microphone in optimal proximity to the participant
EEG setting	20	We put conductive gel to electrodes and examined the quality of the brain signal measurement
Baseline recording	10	The participant had a brief practice session to confirm their readiness for the experiment
Data collection	100	During the main experiment, a total of 200 interactions were conducted while recording multimodal EEG+audiovisual data

Table 2. The experimental procedures for data measurement consisted of two sessions: the screening and recording. Each session lasted for approximately 1 and 2 hours respectively.

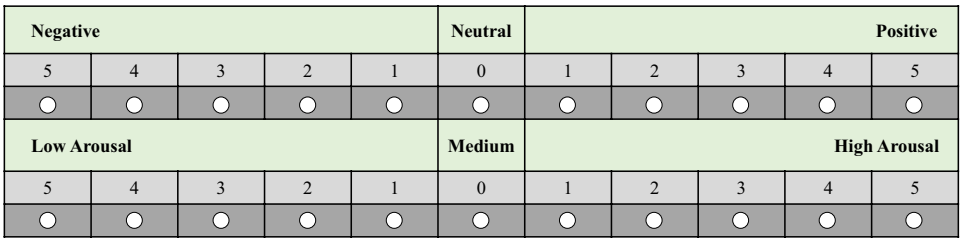


Fig. 2 Subjective self-assessment of emotion on valence and arousal levels.

a participant exhibited signs of biased mental states, such as high levels of fatigue or stress, we either excluded them from the study or rescheduled their session.

During the main experiment, participants sequentially engaged in 20 iterations of conversation, with each iteration lasting around 4 minutes. To allow participants to return to their baseline mental state, we provided ample rest time between iterations. The subsequent conversation only commenced once the participant signaled their readiness for the next trial. At this point, the experimenter alerted the instructor to initiate the next dialogue phase.

After each iteration, participants were asked to provide self-reported ratings for their arousal level and valence score pertaining to the dialogue they had just experienced. The rating scale ranged from -5 to 5, enabling participants to indicate their perceived emotional arousal and the valence (positive or negative) associated with the dialogue (refer to Fig. 2). Additionally, our instructor independently provided a score to assess the participant's emotional state. Upon completing the experiment, we collected scores from both the participant and the instructor for all 20 iterations.

Participants received explicit instructions to actively convey their emotions, both mentally and through their vocal and facial expressions, throughout the conversations. This guidance was intended to ensure that the physiological data (EEG signals), visual data (facial expressions), and audio data (voice modulation) all correspond to the intended emotional class labels. During the experiment, participants were encouraged to make movements and gestures freely to express their emotions.

It is crucial to emphasize that participants had the autonomy to halt the recording at any point during both sessions if they experienced any discomfort. Consequently, eight participants chose to withdraw from the study due to personal reasons.

Experimental Settings. Participants were comfortably seated in front of a computer monitor during the experiment. They interacted with a 27-inch monitor featuring a 60 Hz refresh rate, which displayed the dialogues and other visual stimuli (see Fig. 1a).

The experiment was conducted in a controlled environment to ensure optimal data collection. The controlled settings encompassed the following aspects:

1. **Lighting and Background:** To minimize visual distractions in facial recordings, a white screen was positioned behind the participant. Additionally, a light kit was used to provide uniform illumination of the participant's face.

Ite.	Class	Task	Time (s)	Script
1	N	L	20	Hey, did you know the sun is the largest star in our solar system? It's much bigger than any other star around. It's like the king of our cosmic neighborhood, providing light to all...
	N	S	20	Yes, the sun is indeed the largest star in our solar system. Its immense size is a result of the vast amount of gas and matter it contains. This size allows it to maintain a stable...
2	H	L	20	Hey, How are you doing? I have an exciting plan for this weekend! How about we bring our friends and have a board game night? We can play all our favorite games. It's been ...
	H	S	20	That sounds like an amazing plan! Count me in! I love board games and spending time with our friends. I have a monopoly, I think I can bring it with me. Also, I guess we need ...
3	H	L/S	20/20	Omitted (Listening+Speaking)
4	H	L/S	20/20	Omitted (Listening+Speaking)
5	H	L/S	20/20	Omitted (Listening+Speaking)
Self-report			~15	
Resting			~25	

Table 3. An example of a single conversation consisting of 5 iterations, including 1 neutral (N) and 4 interactions (Listening (L) + Speaking (S)) on one of four emotion classes (e.g., H: Happiness). During self-reporting a participant rated the strength of the emotional response the conversation evoked.

2. **EEG Equipment:** We utilized a BrainAmp system (Brain Products; Munich, Germany) to record brain activities. EEG data were collected through 30 Ag/AgCl electrodes placed at specific locations on the scalp: Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10. The data were sampled at a rate of 500 Hz. The electrodes were referenced to the mastoids and grounded through the AFz electrode. Throughout the experiment, we ensured that the impedance of the EEG electrodes remained consistently below 10 $k\Omega$. The EEG dataset was initially recorded in BrainVision Core Data Format and was subsequently converted into Matlab (.mat) format for ease of use and analysis.
3. **Audio Equipment:** To accurately capture the participant's verbal responses with minimal distortion or interference, we utilized a high-quality microphone. This choice of equipment facilitated a precise analysis of voice modulations corresponding to various emotional states. The audio data was recorded and stored in the WAV (Waveform Audio File Format) for further analysis.
4. **Video Equipment:** A web camera was securely attached to the monitor and positioned to focus directly on the participant's face. This setup was designed to consistently capture facial expressions and emotions throughout the course of the experiment. The data was initially recorded and stored in the AVI (Audio Video Interleave) format, and subsequently, it was converted into the MPEG-4 (.MP4) format to reduce file size.

To ensure that the EEG electrodes were not visible and to eliminate any unnecessary components from the video recording, the electrodes were concealed beneath a black-colored cap.

The entire experiment was conducted using the PsychoPy software, which is based on the Python programming language. This software played a crucial role in various aspects of the experiment, including the presentation of videos and scripts, the management of recordings, ensuring precise timing throughout the study, and facilitating communication with each modality.

Data Records

All data files are accessible via the Zenodo general-purpose open repository[37]. The data is available under the terms and conditions of the data usage agreement (DUA).

The repository is structured to offer comprehensive insights into the data while maintaining a standardized format. It's worth noting that we have segmented the continuous recordings into 20-second data streams.

The root folder, EVA, contains participant folders labeled as `subject{SUBNUM}`, where `{SUBNUM}` ranges from 1 to 42. Each participant folder includes three unimodal data folders: Video, Audio, and EEG (see Fig. 6).

Video Data: The 'Video' subfolder houses segmented video clips, each 20 seconds in length. The naming convention for these clips is:

$$\{NUM\_INTS\}\ \_\text{Trial}\_\ \{NUM\_TRIAL\}\ \_\text{cond}\_\ \{TASK\}\_\{CLASSNAME\}.\text{MP4}$$

- `NUM_INTS`: Represents the instance index number, ranging 200.
- `NUM_TRIAL`: Denotes the trial index within each conversation, spanning from 1 to 10, corresponding to 5 interactions.
- `Task`: Specifies the activity being performed, either 'listening' or 'speaking'.
- `CLASSNAME`: Indicates the associated emotion class.

Collectively, a single subject folder possesses 200 video clips, derived from [5 emotion classes $\times$ 2 tasks $\times$ 20 iterations].

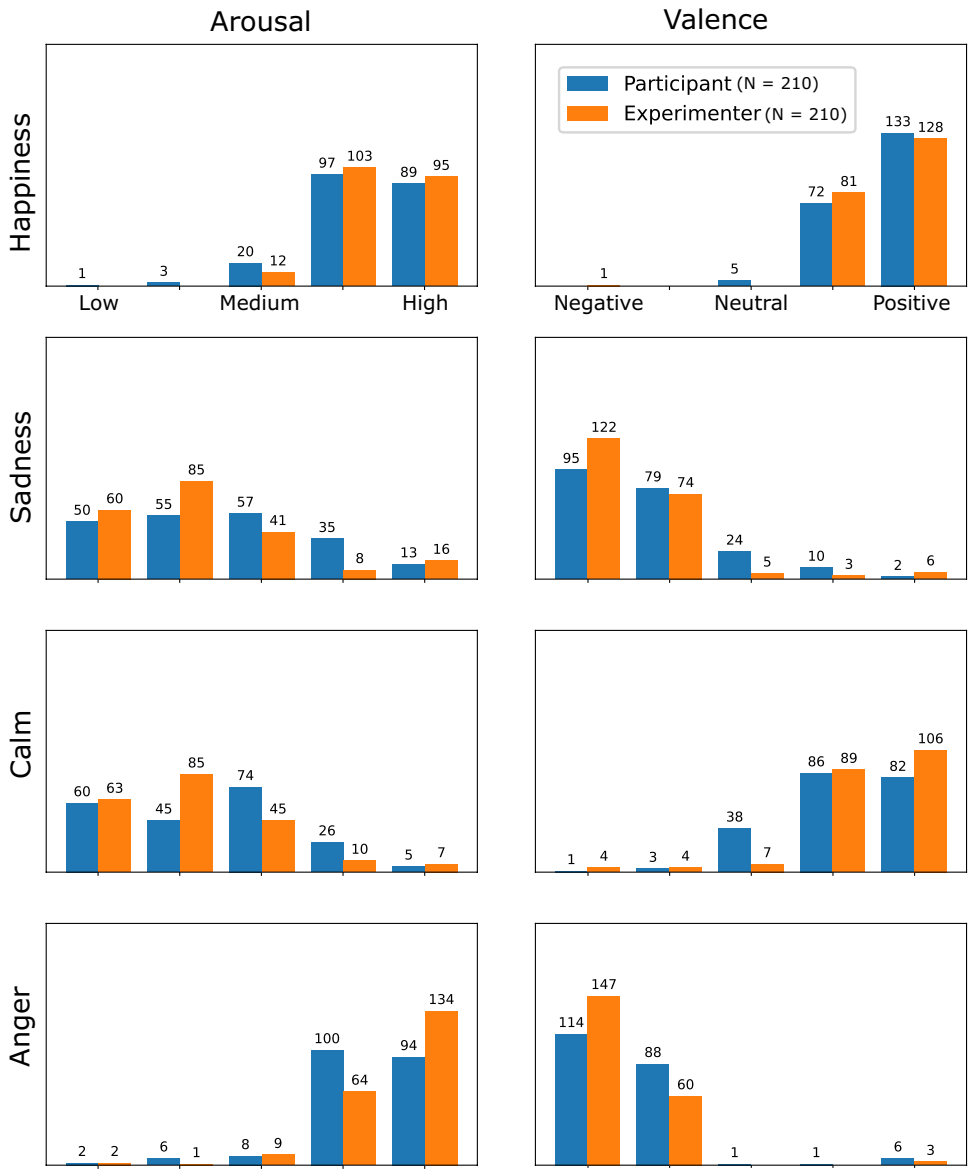


Fig. 3 Comparative Analysis of Emotion Ratings between Participants and Experimenters: A study on Arousal and Valence levels across four emotional states - Happy, Sad, Calm, and Angry. N - total number of data points for rating across all participants.

Audio Data: In the 'Audio' folder, the audio files follow the same naming format as the video files but we appended '_aud' at the end of the file name. It's essential to note that the audio files focus exclusively on the 'speaking' task and omit the 'listening' task. This design choice is due to the audio recordings capturing only the actress's voice. Summarily, each subject in this category boasts 100 audio files, deduced from [5 classes $\times$ 1 task $\times$ 20 conversations].

To offer an intuitive understanding, a single iteration typically comprises the subsequent three video/audio files:

- '001_Trial_01_Listening_Neutral.MP4'
- '002_Trial_02_Speaking_Neutral.MP4'
- '002_Trial_02_Speaking_Neutral_Aud.WAV'

EEG Data: The EEG data were initially recorded continuously with dimensions [Time $\times$ Channels]. For preprocessing, we applied a high-pass filter set above 0.5Hz using a fifth-order Butter-worth filter and band-pass filtered at 50 Hz to mitigate facial noise and electrical line noise. Subsequently, the data were segmented using time markers for each event (trigger), resulting in a structure of [Instances $\times$ Time $\times$ Channels].

Given an initial sampling rate of 500 Hz, our processed EEG data adopts the structure: [200 instances $\times$ 10,000 time points (20s $\times$ 500 Hz) $\times$ 30 channels].

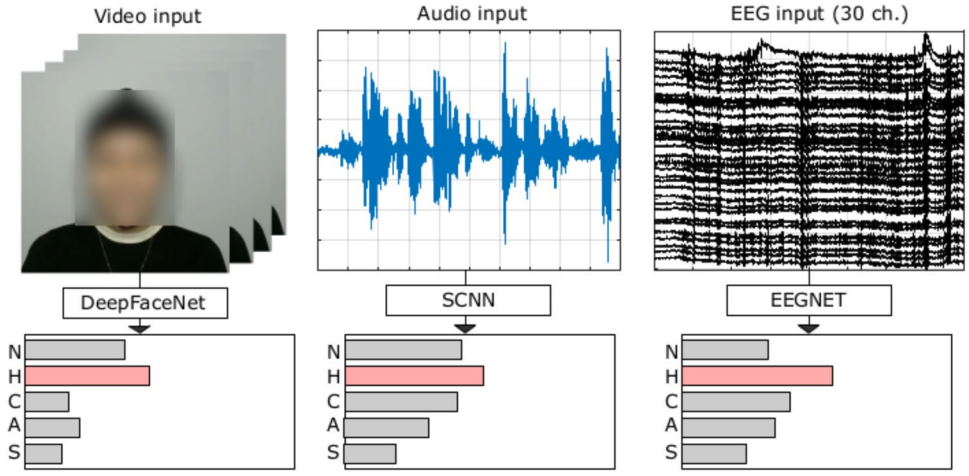


Fig. 4 Multimodal input data and their corresponding processing pipelines for emotion classification. The single trial has 5 seconds of duration. The audio data is preprocessed to create input images while raw video/EEG data were fed to each CNN model. The resulting outputs from these CNN models provide softmax predictions for five emotional states: Neutral (N), Happiness (H), Calmness (C), Anger (A), and Sadness (S). The individual(s) depicted in this figure provided informed consent to have their image openly published.

The labels for this data use a one-hot encoding format, structured as 200 trials by 10 classes (5 emotions multiplied by 2 tasks). The files adhere to the following naming conventions:

subject {*NUM_SUB*} _eeg.mat
subject {*NUM_SUB*} _eeg_label.mat

Note that the label information can be applied across all modalities since all recordings, regardless of the modality, were conducted synchronously. This ensures uniform annotations throughout the dataset.

Missing Data. During the course of the experiment, we encountered several challenges, including technical malfunctions with the equipment, unexpected interruptions, and instances where participants missed the task, among other issues. These missing segments, particularly the iterations in which they occurred, were diligently documented. Subsequently, after the initial experiment was completed, these specific iterations were revisited and conducted again.

This approach was undertaken to maintain data integrity, and the re-positioned iterations were placed in their correct order. Consequently, the entire dataset maintains consistent class labels and formats. It's worth emphasizing that all participants had a balanced dataset, ensuring that there are no discrepancies in the data format across any of the participants.

Technical Validation

The principal objective of this study is to explore open topics in emotion recognition within a single or multimodal setting, rather than to conduct a performance comparison across each modality. It is crucial to recognize the distinct characteristics of EEG, Audio, and Visual signals, as each exhibits unique signal patterns that necessitate specialized analysis pipelines. Performance in these modalities can vary significantly and can be sensitive to the specific techniques employed in each domain. For instance, while certain augmentation techniques in computer vision can substantially enhance performance, their application to audio or EEG data may not consistently yield similar results.

Given the complexity of the task and the diverse nature of the modalities involved, we adopted a widely-used, straightforward end-to-end CNN model and, additionally, more recent transformer architectures[38,39]. We used a simple data split methodology to ensure robust and consistent results.

The dataset was initially annotated with ten evenly distributed class labels: happiness, sadness, anger, relaxation, and neutrality, which were available in both listening and speaking conditions. However, for consistent analysis, only the speaking trials were considered, as the audio data exclusively captures these trials.

For each modality, a single participant was represented by 100 trials, each lasting 20 seconds. However, a 20-second duration is relatively lengthy and not ideal for emotion recognition classification. To address this, we divided the 20-second data streams into 5-second intervals. This division effectively expanded the initial 100 trials and their labels into 400 trials. The dataset was then split into 70% training and 30% testing subsets, with the first 280 trials designated for training and the remaining 120 for testing. Consequently, the performance metrics reported in this study were derived from this consistent set of training and testing data across all modalities. We designed a separate CNN model for each modality, and to establish the baseline performance of these CNN models on the original dataset, we refrained from employing advanced techniques such as knowledge transfer[40], augmentation[41], or signal optimization[42,43].

Sub.	Gender	EEG		Audio		Visual	
		ACC [%]	F-score	ACC [%]	F-score	ACC [%]	F-score
1	M	67.5	0.66	49.2	0.48	37.5	0.38
2	F	80.0	0.78	73.3	0.73	55.8	0.57
3	F	52.5	0.52	67.5	0.67	76.7	0.76
4	M	77.5	0.77	71.7	0.71	79.2	0.79
5	M	52.5	0.52	65.8	0.66	74.2	0.73
6	F	47.5	0.47	77.5	0.78	74.2	0.73
7	F	59.2	0.59	67.5	0.67	77.5	0.76
8	F	56.7	0.55	55.8	0.56	75.8	0.75
9	M	44.2	0.44	70.0	0.69	74.2	0.74
10	M	66.7	0.66	55.0	0.53	64.2	0.63
11	M	46.7	0.46	66.7	0.66	63.3	0.60
12	M	65.8	0.67	52.5	0.53	52.5	0.53
13	F	64.2	0.63	57.5	0.58	80.0	0.80
14	M	47.5	0.48	66.7	0.67	55.0	0.54
15	F	59.2	0.58	66.7	0.65	83.3	0.82
16	F	55.0	0.50	60.8	0.61	61.7	0.55
17	M	78.3	0.78	80.0	0.80	96.7	0.97
18	M	69.2	0.68	66.7	0.64	75.0	0.74
19	F	55.0	0.54	63.3	0.63	77.5	0.77
20	F	77.5	0.77	66.7	0.67	79.2	0.79
21	F	68.3	0.66	59.2	0.57	75.0	0.75
22	F	76.7	0.77	68.3	0.66	76.7	0.76
23	M	63.3	0.62	68.3	0.69	61.7	0.59
24	F	75.0	0.70	50.8	0.51	73.3	0.72
25	F	49.2	0.50	66.7	0.66	70.0	0.69
26	F	58.3	0.56	59.2	0.59	81.7	0.80
27	F	70.0	0.70	64.2	0.63	88.3	0.88
28	F	63.3	0.63	59.2	0.59	83.3	0.83
29	M	51.7	0.49	40.8	0.39	78.3	0.77
30	F	59.2	0.58	61.7	0.63	68.3	0.67
31	M	48.3	0.48	58.3	0.58	67.5	0.62
32	F	60.0	0.60	51.7	0.52	62.5	0.59
33	F	73.3	0.72	84.2	0.84	87.5	0.88
34	F	54.2	0.48	44.2	0.42	77.5	0.76
35	F	30.8	0.29	48.3	0.46	63.3	0.63
36	M	58.3	0.54	65.0	0.62	65.0	0.63
37	F	54.2	0.50	45.0	0.45	60.0	0.60
38	F	64.2	0.62	52.5	0.51	82.5	0.82
39	F	59.2	0.58	68.3	0.69	64.2	0.64
40	M	40.0	0.39	61.7	0.60	58.3	0.52
41	F	26.7	0.23	65.0	0.64	85.8	0.86
42	F	72.5	0.73	50.8	0.54	79.2	0.80
Mean		60.0	0.58	61.9	0.61	71.4	0.70

Table 4. The performance of emotion recognition was evaluated independently in each modality. The accuracy and f-scores for 5 balanced classes were calculated across individual subjects. The represented CNN methods were EEGNet, DeepFaceNet, and SCNN for EEG, Visual, and Audio data, respectively.

To evaluate model performances in multi-class classification, we calculated metrics such as mean accuracy and the F1 Harmonic Mean Score across all participants.

EEG Data Analysis. The segmented EEG data was downsampled to 100 Hz and then band-pass filtered within a frequency range of 0.5-50 Hz. Then the EEG data can be formatted as (400, 30, 500, 1), which represents the total number of trials (training+test), channels, data points (5 seconds), and depth.

DeepConvNet[44], and ShallowConvNet[44] models are straightforward CNN networks that directly take the input EEG data without considering the spatial structure. A recent study applied preprocessing steps to find user-specific frequency ranges[43] then the out images were fed to the CNN model. Lawhern *et al.*[45] demonstrated

Modality (method)	Mean Accuracy [%]	Mean F1-score
EEG (EEGformer)[51]	53.5	0.52
Transformer (AST)[38]	62.7	0.62
Transformer (Vivit)[39]	74.5	0.72

Table 5. Performance comparison of Transformer models with EEG, audio, and visual modalities.

the successful implementation of spectral-spatial optimization into an end-to-end CNN model. In this study, the baseline validation for EEG data was calculated based on the EEGnet[45].

The EEGNet architecture is configured with standard parameters: it includes a 2D-convolutional layer (F1 = 8, kernel-Length = 300), a depthwise convolutional layer (D = 8), and a separable convolution layer (F2 = 16). After these layers, there is a dense layer with a dropout rate of 0.5, followed by a softmax layer. The model utilized categorical cross-entropy as the loss function and employed the Adam optimizer. The training was performed with a batch size of 32 across 100 epochs.

Video Data Analysis. The 5-second video clip contains 150 frames, recorded at 30 frames per second. We extracted 25 frames by selecting every 6th frame, resulting in a total of 10,000 frames (combining training and testing) across all emotion categories. For training, we used the DeepFace model[46], a deep neural network comprising a series of convolutional, pooling, and fully connected layers, totaling nine layers. We adjusted the final softmax layer to accommodate our class number. A dropout layer with a rate of 0.5 was incorporated into the first fully connected (FC) layer. The model was trained using the Adam optimizer with a learning rate of 0.001, and the loss function was cross-entropy. The training was conducted with a batch size of 32 over 100 epochs.

For performance evaluation on each test trial, we predicted the outputs for all frames and then assigned the most frequently occurring class as the prediction.

Audio Data Analysis. To validate the audio data, we employed the Librosa library[47] to preprocess the audio data into a usable format. Subsequently, we used a standard feature extraction method known as Mel-Frequency Cepstral Coefficients (MFCCs)[48]. MFCCs have been widely validated on extensive audio datasets and are a reliable choice for audio feature extraction.

During the pre-processing phase, we extract essential audio features, including MFCCs, Chroma features, and Mel Spectrograms, which are then concatenated for further processing. We utilized a Sequential Convolutional Neural Network (SCNN) architecture[49,50] with one-dimensional data processing capabilities. The model consists of four 1D convolutional layers with Rectified Linear Unit (ReLU) activation, interspersed with dropout layers to enhance generalization. Additionally, L1 and L2 regularization techniques are applied to reduce overfitting. The network starts with 256 filters and progressively decreases to 128 filters, followed by Max pooling for down-sampling. The architecture concludes with a densely connected layer employing softmax activation. Model optimization is performed using the Adam optimizer, with a specific weight decay of $1e-6$ and a learning rate parameter of $1e-3$ to fine-tune the training process.

Figure 4 provides a concise overview of the classification pipelines for each data type. The multimodal data samples depicted in this figure were extracted from one of the participants (*participant30*), representing the happiness class.

Transformer architecture. The Transformer was initially developed for language modeling and has since replaced CNNs due to its superior ability to handle dependencies. In our study, the performance of EEG, audio, and visual data was validated using specific Transformer architectures, namely EEGformer[51], AST[38], and ViViT[39], respectively. The model architectures and parameters were used as outlined in their original publications.

Performance evaluation. The dataset was partitioned into two subsets: a training set and a test set. This division was executed in a manner where the initial 70% of the video/audio/EEG data was designated for training, while the latter 30% was reserved for testing. Specifically, for audio data, each participant contributed 400 audio samples, each lasting 5 seconds. Of these, 280 audio samples were drawn from the first 70% of conversations, with the remaining 120 selected for testing purposes. This ratio can be flexibly adjusted to align with the specific objectives of the study. Commonly, we train the model for a fixed number of epochs without using a validation set.

Table 4 presents the emotion recognition performance in the uni-modal condition. For individual CNN models, the EEG data yielded a mean accuracy of 60.0% $\pm$ 0.09 and an average F1-score of 0.58 $\pm$ 0.09. The audio data yielded a mean accuracy of 61.9% $\pm$ 0.08 and an average F1-score of 0.61 $\pm$ 0.09. The video data model reported a mean accuracy of 71.4% $\pm$ 0.08 and a mean F1-score of 0.70 $\pm$ 0.08. It's important to note that the reported scores in Table 4 can be easily varied according to the diverse machine learning techniques and data split configurations.

Table 5 provides a comparative analysis of the mean accuracy and mean F1-score for Transformer models. The results demonstrate that the pretrained Transformer model enhances classification performance in the audio and video modalities, achieving mean accuracies of 62.7% and 74.5%, respectively. Conversely, the EEGTransformer[51] did not yield comparable improvements, achieving 53.5% accuracy and an F1 score of 0.52. This performance discrepancy may be attributed to the skewness in the availability of pre-trained models.

Figure 5 indicates the confusion matrices for EEG, Audio, and Video modalities, each assessed over 1008 trials across distinct emotion categories for all participants. The EEG modality showcases a marked accuracy in

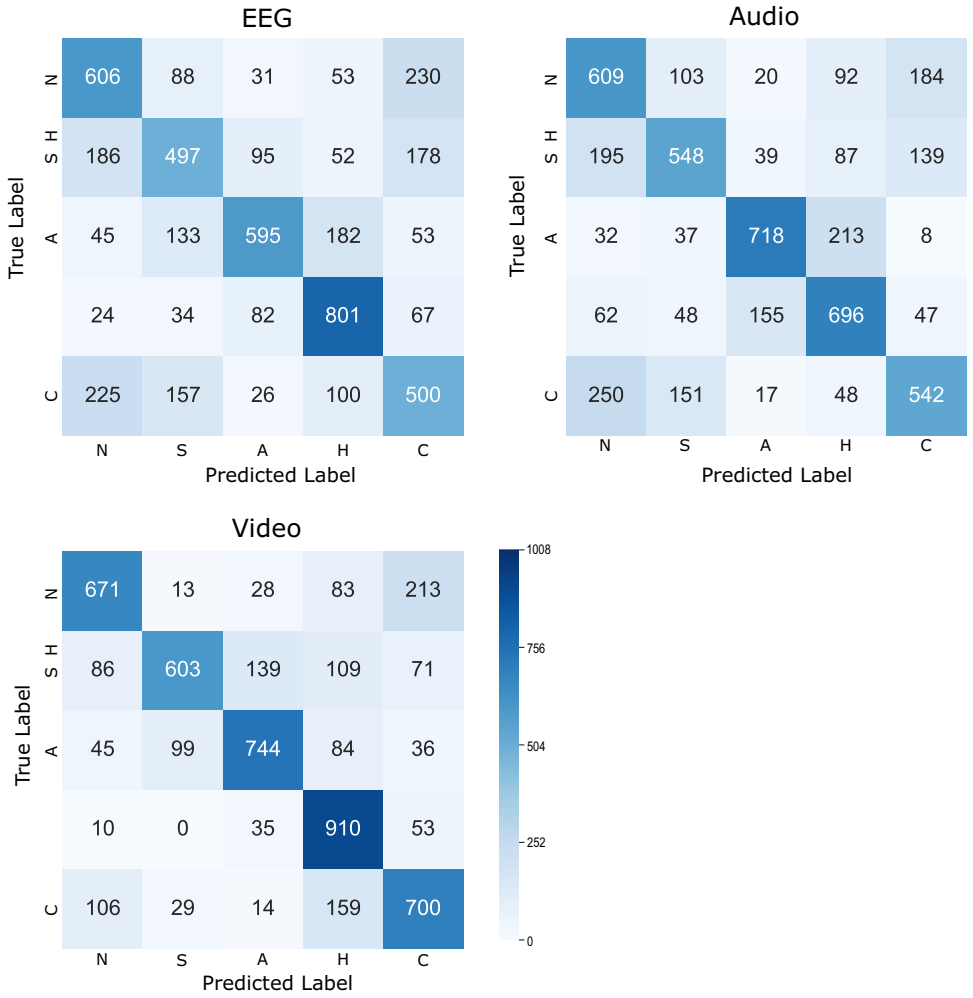


Fig. 5 Accumulative confusion matrix for classifying five emotion classes across all participants using EEG, Audio, and Video data. The matrices represent for each emotion: Neutral (N), Sadness (S), Anger (A), Happiness (H), and Calm (C). The total test trials per class across all participants are 1008 (24 speaking tasks × 42 participants).

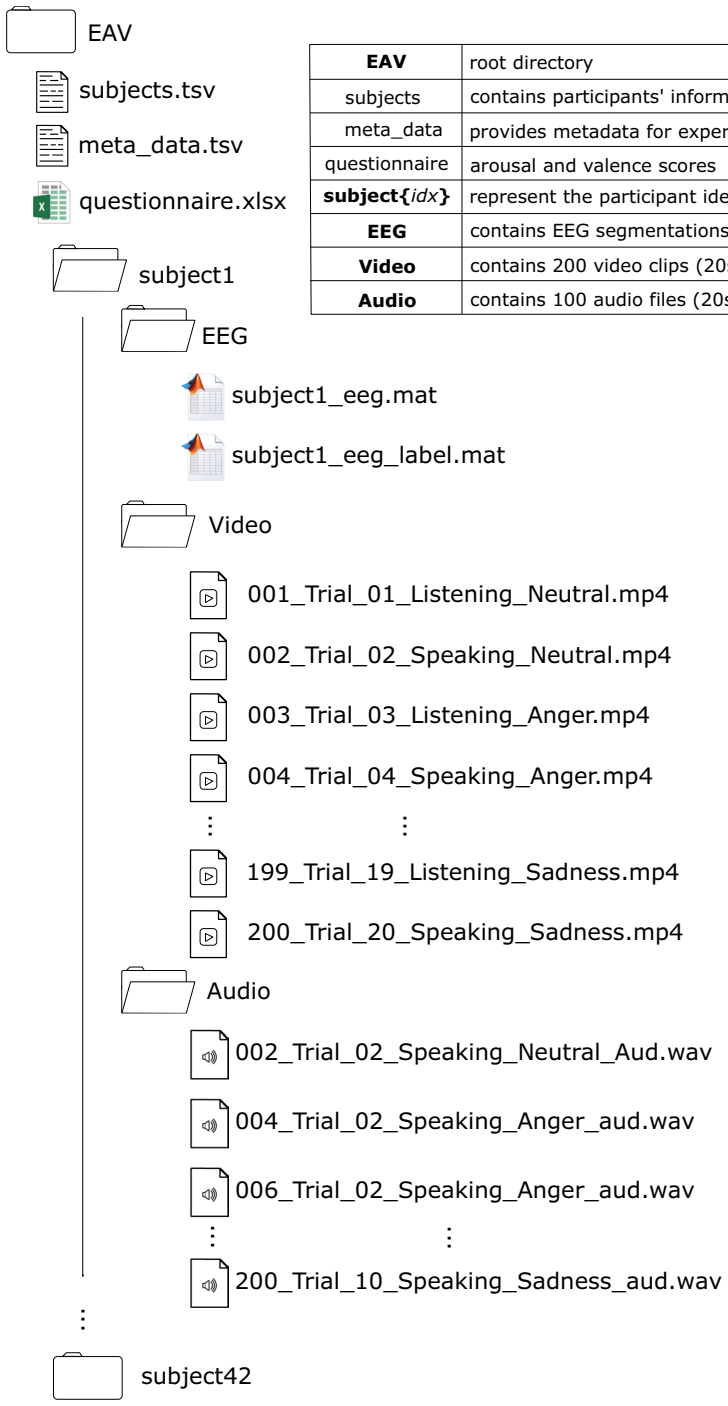
classifying the Happiness and Neutral emotions. The Audio modality distinctly excels in discerning high-arousal states such as Angry and Happiness. Both EEG and Audio modalities manifest a congruent trend, notably mispredicting low-arousal classes like Neutral, Sadness, and Calm – an observation that aligns well with foundational knowledge in the field. In stark contrast, the Video modality reveals significant mispredictions, frequently misclassifying low-arousal classes as high-arousal emotions, specifically Angry and Happiness.

In an analysis of valence and arousal levels across 42 participants, the data was analyzed for five distinct emotions. Within each emotion, participants participated in four separate conversations. For each conversation, participants' valence and arousal levels were measured on a scale ranging from -5 (lowest) to 5 (highest) (see Fig. 2). On average, emotions traditionally associated with higher arousal, such as excitement or anger, consistently scored above 3.5, while emotions typically associated with lower arousal, such as sadness or calmness, scored below 2.5 (see Fig. 3). Similarly, positive emotions, like happiness, had a valence level frequently exceeding 4, whereas negative emotions, like sadness, were frequently below 2 on the valence scale.

Usage Notes

This work is under a Data Usage Agreement (DUA) at Zenodo[37]. The approval process and usage restrictions are outlined in the accompanying application form, which can be accessed through the repository. The data access will be granted to applicants who agree to the terms and conditions outlined in the Data Usage Agreement. In order to access and use the EAV dataset, you will be asked to provide your full name, affiliation, position/title, and a brief description of your intended use of the dataset.

For EEG data analysis, the BBCI toolbox (https://github.com/bbci/bbci_public)[52] and OpenBMI[53] offer a wide range of signal processing functions in Matlab, including artifact rejection, spectral/spatial filtering, resampling, and re-referencing. The MNE toolbox (https://min2net.github.io)[45] provides similar functions and visualization methods in Python.



EAV	root directory
subjects	contains participants' information: age, sex
meta_data	provides metadata for experimental setting
questionnaire	arousal and valence scores
subject{*idx*}	represent the participant identifier; {*1 - 42*}
EEG	contains EEG segmentations [30ch x 20s x 200 trials]
Video	contains 200 video clips (20s each)
Audio	contains 100 audio files (20s, speaking only)

Fig. 6 The data repository is structured for intuitive navigation. Under the primary level, folders are named 'subject{*idx*}', where '*idx*' denotes the participant identifier. Each of these participant-specific directories contains three sub-directories: 'Video', 'Audio', and 'EEG'. Correspondingly, data files within these sub-directories adhere to their specific formats: video files are saved as *.MP4, audio recordings are in *.WAV format, and EEG data is stored as *.MAT files.

For audiovisual data analysis, the implemented CNN models in this study can be found in two repositories: Deepface (https://github.com/serengil/deepface)[46] for video analysis and (https://github.com/vandana-rajan/1D-Speech-Emotion-Recognition)[40] for audio analysis.

Limitations. Our conversational scenario is designed using cue-based and posed conditions to standardize the conversation, ensure consistency, and achieve balanced class labels. However, such posed scenarios and our experimental setting come with acknowledged limitations:

- The scripted dialogues may not fully capture the spontaneity and genuine emotional expression that occurs in real-life conversations.
- Human reactions may be exaggerated or downplayed due to cue-based conversation scenarios, which could result in mismatched emotion trials.
- General participants may lack familiarity with expressing emotions facially, our dataset is particularly well-suited for multimodal settings.
- The setup of the EEG cap could constrain the facial expression, and it covered the forehead area.
- EEG cleaning methods, i.e., artifact rejection, spatial filtering, and normalization, were excluded to maintain consistent baseline results.
- Participants' ages were limited to a young age group, and they are not native English speakers. This may create biased results and reduce performance in applying a language model or the use of a pretrained model.
- The predefined emotional classes may not encompass the full spectrum of human emotions, potentially limiting the dataset's representativeness.

Recognizing emotions from videos of conversations presents unique challenges, primarily due to the obstructions introduced by non-verbal cues[54]. In contrast, the audio signal offers a more discernible pathway to emotion identification, driven by nuanced auditory features such as pitch, tone, and voice dynamics. A review of literature on prevalent visual-audio datasets reveals significant disparities in the emotion recognition capabilities between video and audio modalities. For instance, in[21] authors reported accuracies of 71.77% for audio-only and 47.44% for video-only modalities. A similar trend was observed by another study[55], which recorded accuracies of 42.16% for video and 72.95% for audio.

In an EEG-based emotion study, participants provided ratings on scales (e.g., valence) and the ground truth has been determined based on these self-assessed ratings scores[12,18]. On the contrary, in audio/video studies with conversational scenarios, the emotion classes are mostly pre-determined and participants are expected to align their emotions to the given scenario[15,17,20]. Furthermore, the annotation can be created with diverse perspectives[14,19]. Brain-focused studies tend to focus on participants' internal/direct feeling, while vision/audio studies concentrate more on their external/indirect expressions.

In our research, we utilize a multimodal dataset and aim to reduce this gap by synchronizing the internal feelings with their external expressions. Participants selected the scripts that resonated most with them emotionally, and they were then asked to fully express their emotions through voice and facial expressions.

Our dataset included two supplementary class labels. One is based on participant's reported arousal-valence scores, and the other is determined by a supervisor's observation. Feedback from the participants might be more indicative of direct emotions, whereas observations by the supervisor could align more with indirect emotions. Note that our dataset can be annotated differently based on these reports, catering to researchers' interests in specific modalities and emotional foci, whether indirect or direct.

We believe that despite these limitations, our study provides valuable insights into multimodal emotion recognition in controlled conversational settings, providing a strong foundation for further research and the development of emotion recognition models.

In conclusion, this study has introduced a valuable resource in the form of the Emotion in Audio-Visual (EAV) dataset, encompassing EEG, audio, and video recordings collected during cue-based conversations. By addressing the critical need to understand emotional states for human-machine interfaces, we have taken a significant step toward developing AI models with human-like attributes. The EAV dataset is the first of its kind to include these three primary modalities, enabling comprehensive emotion recognition within conversational contexts. We believe that this dataset will play a pivotal role in advancing our understanding of human emotions from both neuroscience and machine learning perspectives, offering a robust foundation for future research and model development.

Future studies can build upon the foundation laid by the EAV dataset and contribute to advancements in emotion recognition, human-computer interaction, and affective computing, ultimately benefiting a wide range of applications and industries. Expanding the dataset to include participants from various cultural backgrounds can lead to insights into how emotions are expressed and recognized differently across cultures, enabling more culturally sensitive emotion recognition models. Moreover, exploring practical applications of emotion recognition, such as in online education, customer service, and mental health support, can be a relevant future research direction. Implementing and evaluating emotion recognition systems in real-world scenarios can provide valuable insights and enhance the practicality of this technology.

Code availability

The GitHub repository, available at https://github.com/nubcico/EAV, includes various methodologies, data split, configuration, preprocessing, and implementations of CNNs and Vision Transformers along with the complete training pipeline. While details of the model parameters and training procedures are omitted in this paper, they can be fully replicated using the resources provided in our repository.

Received: 27 December 2023; Accepted: 29 August 2024;
Published online: 19 September 2024

References

1. Salovey, P. & Mayer, J. D. Emotional intelligence. *Imagination, cognition and personality* **9**, 185–211 (1990).
2. Lopes, P. N. *et al*. Emotional intelligence and social interaction. *Personality and social psychology bulletin* **30**, 1018–1034 (2004).
3. Etkin, A., Büchel, C. & Gross, J. J. The neural bases of emotion regulation. *Nature reviews neuroscience* **16**, 693–700 (2015).
4. Jazaieri, H., Morrison, A. S., Goldin, P. R. & Gross, J. J. The role of emotion and emotion regulation in social anxiety disorder. *Current psychiatry reports* **17**, 1–9 (2015).

5. Gunes, H. & Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* **31**, 120–136 (2013).
6. Soleymani, M., Lichtenauer, J., Pun, T. & Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* **3**, 42–55 (2011).
7. Miranda-Correa, J. A., Abadi, M. K., Sebe, N. & Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* **12**, 479–493 (2018).
8. Subramanian, R. *et al*. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* **9**, 147–160 (2016).
9. Song, T. *et al*. MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access* **7**, 12177–12191 (2019).
10. Katsigiannis, S. & Ramzan, N. 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**, 98–107 (2017).
11. Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L. & Cichocki, A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics* **49**, 1110–1122 (2018).
12. Koelstra, S. *et al*. DEAP: A database for emotion analysis using physiological signals. *IEEE transactions on affective computing* **3**, 18–31 (2011).
13. Saffaryazdi, N. *et al*. Emotion recognition in conversations using brain and physiological signals. In *27th International Conference on Intelligent User Interfaces*, 229-242 (2022).
14. Park, C. Y. *et al*. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* **7**, 293 (2020).
15. Busso, C. *et al*. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* **42**, 335–359 (2008).
16. McKeown, G., Valstar, M., Cowie, R., Pantic, M. & Schroder, M. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing* **3**, 5–17 (2011).
17. Busso, C. *et al*. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* **8**, 67–80 (2016).
18. O'Reilly, H. *et al*. The EU-emotion stimulus set: A validation study. *Behavior research methods* **48**, 567–576 (2016).
19. Chou, H.-C. *et al*. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 292-298 (2017).
20. Livingstone, S. R. & Russo, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* **13**, e0196391 (2018).
21. Zhalehpour, S., Onder, O., Akhtar, Z. & Erdem, C. E. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing* **8**, 300–313 (2016).
22. Haq, S., Jackson, P. J. & Edge, J. Speaker-dependent audio-visual emotion recognition. In *AVSP*, vol. 2009, 53–58 (2009).
23. Li, Y., Tao, J., Chao, L., Bao, W. & Liu, Y. CHEAVD: A Chinese natural emotional audio-visual database. *Journal of Ambient Intelligence and Humanized Computing* **8**, 913–924 (2017).
24. Poria, S. *et al*. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536 (2019).
25. Chen, J. *et al*. HEU emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications* **33**, 8669–8685 (2021).
26. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W. & Zafeiriou, S. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing* **11**, 1301–1309 (2017).
27. Ji, X. *et al*. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14080–14089 (2021).
28. Lei, Y. & Cao, H. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing* (2023).
29. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. Emotion, attention, and the startle reflex. *Psychological review* **97**, 377 (1990).
30. Eskimez, S. E., Zhang, Y. & Duan, Z. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia* **24**, 3480–3490 (2021).
31. Tu, G., Liang, B. Jiang, D., & Xu, R. Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing* (2022).
32. Wang, Y. *et al*. Tacotron: Towards End-to-End Speech Synthesis. In *Proceedings of the Interspeech 2017*, 4006–4010 (2017).
33. Zhou, K., Sisman, B., Rana, R., Schuller, B. W. & Li, H. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing* **14** (2023).
34. Yuan, X. *et al*. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6995–7004 (2021).
35. Sun, Y. *et al*. Long-form video-language pre-training with multimodal temporal contrastive learning. *Advances in neural information processing systems* **35**, 38032–38045 (2022).
36. Oh, T.-H. *et al*. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7539–7548 (2019).
37. Lee, M.-H. *et al*. EAV: EEG-Audio-Video Dataset for Emotion Recognition in Conversational Contexts. *Zenodo*. https://doi.org/10.5281/zenodo.10205702 (2023).
38. Gong, Y. *et al*. AST: Audio Spectrogram Transformer. In *Proceedings of the Interspeech 2021*, 571–575 (2021).
39. Arnab, A. *et al*. VIVIT: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836-6846 (2021).
40. Zhao, J., Mao, X. & Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control* **47**, 312–323 (2019).
41. Liu, K., Perov, I., Akhtar, Z. & Gao, D. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition* **141** (2016).
42. Bang, J.-S., Lee, M.-H., Fazli, S., Guan, C. & Lee, S.-W. Spatio-spectral feature representation for motor imagery classification using convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems* **33**, 3038–3049 (2021).
43. Kwon, O.-Y., Lee, M.-H., Guan, C. & Lee, S.-W. Subject-independent brain-computer interfaces based on deep convolutional neural networks. *IEEE transactions on neural networks and learning systems* **31**, 3839–3852 (2019).
44. Schirrmeister, R. T. *et al*. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* **38**, 5391–5420 (2017).
45. Lawhern, V. J. *et al*. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of neural engineering* **15**, 056013 (2018).
46. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1708 (2014).
47. McFee, B. *et al*. Librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, 18-25 (2015).

48. Logan, B. *et al*. Mel frequency cepstral coefficients for music modeling. In *Ismir* 270, 11 (2000).
49. Yang, H., Yuan, C., Xing, J. & Hu, W. SCNN: Sequential convolutional neural network for human action recognition in videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, 355-359 (2017).
50. Issa, D., Demirci, M. F. & Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* **59**, 101894 (2020).
51. Wan, Z. *et al*. EEGformer: A transformer-based brain activity classification method using EEG signal. *Frontiers in Neuroscience* **17**, 1148855 (2023).
52. Blankertz, B. *et al*. The Berlin brain-computer interface: Progress beyond communication and control. *Frontiers in neuroscience* **10**, 530 (2016).
53. Lee, M.-H. *et al*. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience* **8**, giz002 (2019).
54. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J. & Gedeon, T. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI)*, 423-426 (2015).
55. Martin, O., Kotsia, I., Macq, B. & Pitas, I. The eNTERFACE' 05 Audio-Visual Emotion Database. In *22nd International Conference on Data Engineering Workshops (ICDEW)*, 8-8 (2006).

Acknowledgements

This work is supported by the Ministry of Science and Higher Education of the Republic of Kazakhstan for Prof. Dr. Adnan Yazıcı under the grant titled "Smart-Care: Innovative Multi-Sensor Technology for Elderly and Disabled Health Management" (AP23487613, duration 2024-2026) and by the Faculty Development Competitive Research Grant Programs of Nazarbayev University with reference number 20122022FD4109: "Intention Estimation from Behavior and Emotional Expression". This work was also partially supported by the National Research Foundation of Korea (NRF) grant funded by the MSIT (No. 2022-2-00975, MetaSkin: Developing Next-generation Neurohaptic Interface Technology that enables Communication and Control in Metaverse by Skin Touch) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University)).

Author contributions

M.-H.L. conceived the experiment, wrote the paper, designed the paradigm, and conducted data analysis. A.S. developed the experimental paradigm. B.B. and A.N. were responsible for data collection. Z.K. was involved in CNN model development and creating the GitHub code. A.Y. contributed to writing the paper and data analysis. S.-W.L. supervised the experiment and handled ethical considerations. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.-W.L.