Computers in Biology and Medicine 149 (2022) 105907

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Multi-modal emotion recognition using EEG and speech signals ☆

Qian Wang, Mou Wang, Yan Yang *, Xiaolei Zhang

Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China

ARTICLE INFO

Keywords:
Multi-modal emotion database
EEG emotion recognition
Speech emotion recognition
Physiological signal
Data fusion

ABSTRACT

Automatic Emotion Recognition (AER) is critical for naturalistic Human–Machine Interactions (HMI). Emotions can be detected through both external behaviors, e.g., tone of voice and internal physiological signals, e.g., electroencephalogram (EEG). In this paper, we first constructed a multi-modal emotion database, named Multi-modal Emotion Database with four modalities (MED4). MED4 consists of synchronously recorded signals of participants' EEG, photoplethysmography, speech and facial images when they were influenced by video stimuli designed to induce happy, sad, angry and neutral emotions. The experiment was performed with 32 participants in two environment conditions, a research lab with natural noises and an anechoic chamber. Four baseline algorithms were developed to verify the database and the performances of AER methods, Identification-vector + Probabilistic Linear Discriminant Analysis (I-vector + PLDA), Temporal Convolutional Network (TCN), Extreme Learning Machine (ELM) and Multi-Layer Perception Network (MLP). Furthermore, two fusion strategies on feature-level and decision-level respectively were designed to utilize both external and internal information of human status. The results showed that EEG signals generate higher accuracy in emotion recognition than that of speech signals (achieving 88.92% in anechoic room and 89.70% in natural noisy room vs 64.67% and 58.92% respectively). Fusion strategies that combine speech and EEG signals can improve overall accuracy of emotion recognition by 25.92% when compared to speech and 1.67% when compared to EEG in anechoic room and 31.74% and 0.96% in natural noisy room. Fusion methods also enhance the robustness of AER in the noisy environment. The MED4 database will be made publicly available, in order to encourage researchers all over the world to develop and validate various advanced methods for AER.

1. Introduction

Emotions contain information about people's intentions and reactions and have a significant impact on the perceptions and decisions of the people they communicate with [1,2]. Precise recognition of emotions in interpersonal and human–computer interactions can lead to more harmonious and natural communication [3,4]. Therefore, emotion cognition and recognition play an important role in our everyday social communication.

In neuroscience, emotions are defined as complex psycho-physiological processes that reflect the reactions of our nervous system toward external relations [5]. The reactions involve changes in the external behaviors such as facial expressions, body gestures, frequency and speed of voice, as well as internal physiological responses such as electroencephalography (EEG), electrocardiogram (ECG), respiration and pulse [1]. For example, anger leads to louder voice, staring and frowning, raised blood pressure, body temperature and increased heart rate. In daily life, we understand emotions better when we combine a speaker's facial expressions, body gestures and the tones of voice together, as the expression and perception of emotions are essentially multi-modal [6]. In this case, all the three channels of information contribute to the understanding of emotions. Therefore, multi-modal emotion expressions which integrate the complementary information among different emotion modalities can provide a better interpretation of emotions.

Access to annotated multi-modal database is a prerequisite for developing Automatic Emotion Recognition (AER) algorithms. There are some published databases aiming to providing benchmark methods for AER [7–10]. Since speech and facial expressions are the most intuitive modalities for emotion recognition, the existing multi-modal databases mainly focus on audio and video signals [11,12]. However, this may not be sufficient since humans can involuntarily or intentionally conceal their emotions, known as social masking [13]. For instances, people can perceive and understand emotional information from speech signals, such as special mood words and intonation changes and can also hide

☆ The work of Yan Yang was supported in part by NSFC under Grants 61771403 and N2018KF0157.

* Corresponding author.
E-mail addresses: wangqian203714@mail.nwpu.edu.cn (Q. Wang), wangmou21@mail.nwpu.edu.cn (M. Wang), y.yang@nwpu.edu.cn (Y. Yang), xiaolei.zhang@nwpu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.compbiomed.2022.105907
Received 8 February 2022; Received in revised form 29 June 2022; Accepted 16 July 2022
Available online 22 July 2022

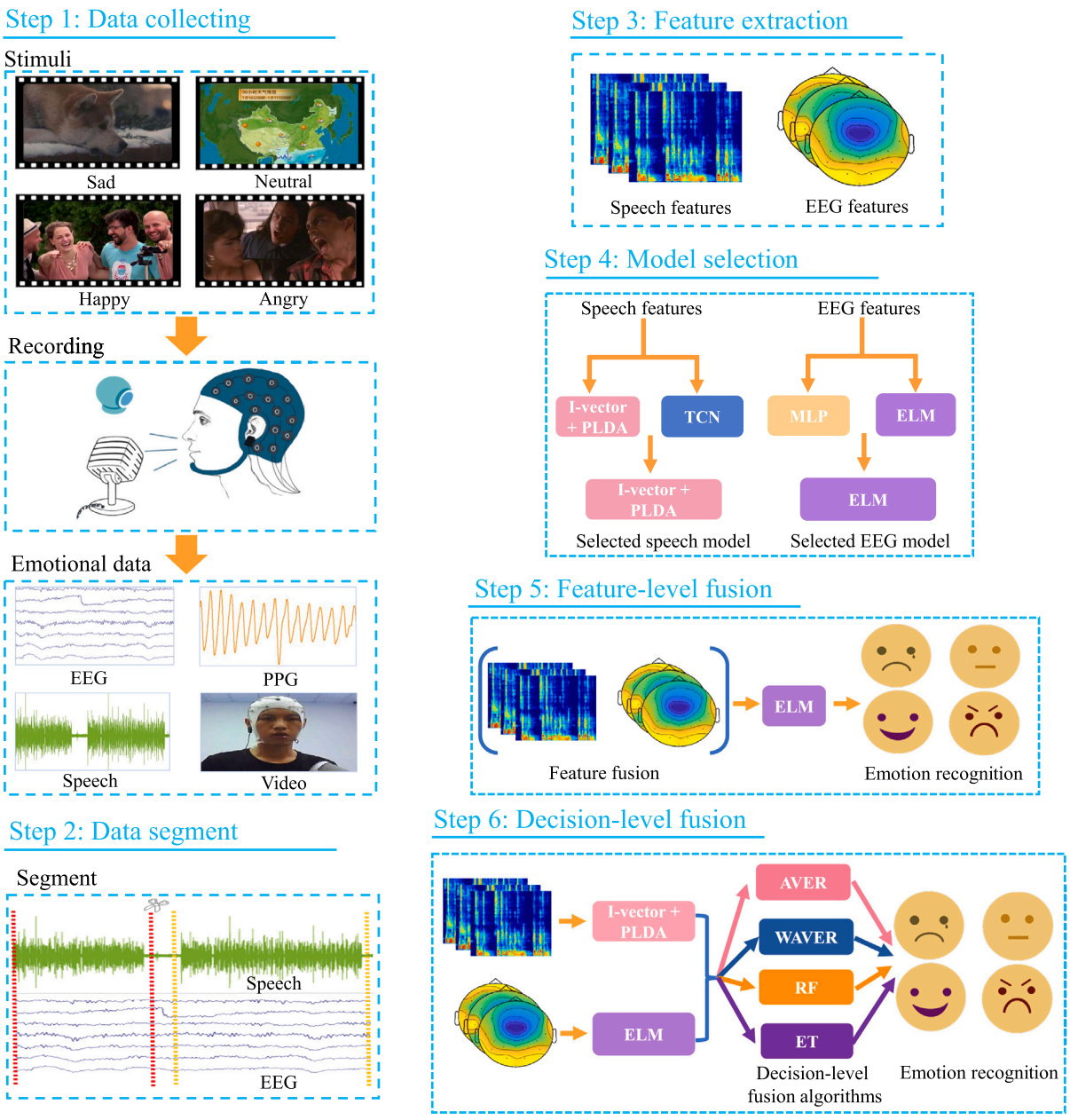


Fig. 1. The framework of emotion recognition. EEG, PPG, Speech and Video signals were simultaneously recorded while subjects watching emotional video clips and read emotional contextual script materials afterwards. EEG and Speech data were segmented and pre-processed. Features were then extracted and fed into baseline algorithms. Finally, the emotion recognition was given through both feature-level and decision-level fusion methods.

true emotions by regulating emotional information in speech [14]. In contrast, physiological information can provide some complementary advantages for AER. Signals from the Autonomic Nervous System (ANS) and Central Nervous System (CNS) signals, such as ECG and EEG, are not easily affected by conscious or intentional control, thus are considered more robust and objective. EEG signals [15] have been widely used in AER since they capture the emotional information from its origin. However, less work is available in combining internal physiological responses into audio/video channels. Therefore, to consider the characteristics of each modality, a comprehensive multi-modal emotion database is needed.

Among different channels of emotional expressions, speech signals are easily corrupted with environmental factors, such as background noise and the reverberation in natural environments, which in turn affect the performance of speech-based AER [16,17]. Schuller et al. suggested that the accuracy degrades could be as high as 74.5% for clean speech to 54.9% at −10 dB signal-to-noise ratio (SNR) on Danish emotional speech corpus (DES) for five emotions (anger, joy, sadness, surprise and neutrality) [18]. Therefore, it is necessary to develop robust and secure AER systems to resist the inevitable noise and adversarial environment (S. Zhao et al.) [13]. Many methods were developed to tackle this issue. However, in most studies, speech signals were collected in one noisy environment [16,17] or simulated by adding different types of noise to clean speech [19]. There is limited access for databases that include both clean and noisy signals in controlled experiment conditions. Therefore, in this work we considered environmental noise as an independent factor in the experimental design.

In this paper, we integrated the emotional external behaviors and internal physiological signals to recognize human emotions. The main framework of our proposed AER method is shown in Fig. 1. The main contribution of this work are as follows. We first constructed a multi-modal and multi-environmental emotion database which includes speech, video, EEG and photoplethysmography (PPG) signals

Table 1
Summary of the main Characteristics of Emotional Database Reviewed.

Dataset	Subjects	Modalities	Hours	Labels	Language	Environment
BAUM-1	31	Audio, Visual	–	Happy, confusion, sad, disgust, angry, fear, boredom, interest	Turkish	Natural
SEMAINE	150	Audio, Visual	~80 h	Valence, arousal, and action units	English	Natural
IEMOCAP	10	Audio, Visual	~12 h	Happy, angry, sad, neutral and valence, activation, dominance	English	Natural
SEED-IV	44	EEG, eye movement	~105 h	Happy, sad, fear, neutral	\	Natural
DEAP	32	Visual, EEG, EMG, EOG, RA, BVP, GSR, ST	~21 h	Arousal, valence, liking, dominance, familiarity	\	Natural
NNIME	44	Audio, Visual, ECG	~7h	Happy, sad, angry, neutral, surprise and frustration	Chinese	Natural
RAMAS	10	Audio, Visual, Posture, ECG	~7 h	Happy, sad, angry, fear, disgust, surprise	Russian	Natural
MAHNOB-HCI	27	Audio, Visual, GSR, RA, BVP, ST, Eye Gaze, EEG, ECG	~9 h	Valence, dominance, arousal, predictability, and emotional keywords	English	Natural
MED4	32	Audio, Visual, EEG, PPG	~15 h	Happy, sad, angry, neutral	Chinese	Natural, Anechoic

The last row is our database. The acronyms in the table are: EMG is Electromyogram, EOG is Electrooculogram, GSR is Galvanic Skin Response, ST is Skin Temperature, BVP is Blood Volume Pressure, RA is Respiration Amplitude. The "~" symbol means approximate value. The natural environment is the room that is not equipped with professional sound absorbing materials.

recorded from natural noisy room and anechoic chamber. Based on the MED4 database, we analyzed the effect of the window size for EEG sampling on AER. We then performed single modality AER based on EEG and speech signals separately from different environment through four baseline algorithms. Finally, to identify whether there is complementary information between speech and EEG signals, we performed feature- and decision-level fusion strategies. We concatenated features from EEG and speech signals as feature-level fusion method and applied average sum (AVER), weighted average sum (WAVER), random forest (RF) [20] and extremely randomized trees (ET) [21] as decision-level fusion methods to evaluate the strength and weakness of each approach.

The rest of this paper is organized as follows. Section 2 reviews the related multi-modal emotional databases. Section 3 introduces MED4 database in detail. Section 4 explains the pre-processing and features extraction of speech and EEG signals. Section 5 gives four baseline algorithms. The experimental results are demonstrated in Section 6. Finally, the paper is concluded in Section 7.

2. Multi-modal emotional databases

Creating a high quality multi-modal emotional database that prompts an advance in AER is an important step and requires a deep understanding of existing databases. This section focuses on the publicly available multi-modal emotional databases related to audio and physiological signals modalities. In Table 1, we summarize the characteristics of the reviewed databases.

Existing multi-modal emotional databases including audio signals mainly use three ways to stimulate target emotions: acting, spontaneous responding to designed scenarios, and eliciting emotions via stimuli with emotional contents. BAUM-1 [11] is an acted audio–visual database involving 8 emotions and mental states, where 31 subjects were asked to read aloud scripts in Turkish while imagining specific scenarios. The SEMAINE [7] database contains naturalistic audio–visual expressions from 150 subjects, who engaged in an emotional conversation with a sensitive artificial listener. A total of 959 conversations were recorded, each lasting approximately five minutes. Elicitation techniques were selected in IEMOCAP [22] database : the use of scripts. Ten actors were divided into five dyadic pairs and were asked to perform scripts with clear emotional contents under the supervision of an experienced professional. Approximately twelve hours of audio–visual data were included.

Using emotional film clips, pictures or music videos to elicit subjects' target emotions are widely used in collection of spontaneous physiological responses. SEED-IV [23] and DEAP [8] are two commonly used emotional benchmark datasets related to physiological signals. SEED-IV synchronously recorded 44 subjects' EEG data and Eye movements during they watched video clips which induced happy, sad, fear and neutral emotions. Totally 72 film clips were shown to each subject and the duration of each film clip was approximately two minutes. The DEAP database comprises EEG and peripheral physiological recordings of 32 subjects while watching 40 different music videos of one minute duration.

Multi-modal databases that combine external emotional behaviors and internal physiological signals can provide a more comprehensive data source for AER. In NNIME [24], 44 subjects were paired into dyadic groups to spontaneously perform a short hypothesized scene of about three minutes. They were instructed to interact freely with each other in order to collect audio–visual and ECG responses of the six pre-specified emotions (Happy, sad, angry, neutral, surprise and frustration). A similar database that uses improvised dyadic interactions is the RAMAS [25] database, which includes approximately seven hours of recordings of audio–visual, posture and ECG signals from ten semi-professional actors. However, the above two multi-modal emotional databases did not record EEG responses for emotion expression. MAHNOB-HCI [26] is one of the few publicly available databases which contains both audio–visual channels, CNS and ANS signals. It synchronously recorded face video, audio, eye gaze data and physiological signals of 27 subjects while watching videos and images. However, the audio signals were some natural utterances and laughter that expected occur during experiments and just used for video tagging. Therefore, in this paper we adopted the similar elicited technique that use film clips to evoke subjects' emotions, then collected the emotional responses that includes signals from audio–visual, CNS and ANS channels while subjects read scripts.

From Table 1, it is observed that MED4 is the only one multi-environmental and multi-modal emotional database that contains

Table 2
The video clips and their sources.

Code	Emotion labels	Video clips sources
1	Neutral	Weather forecast
2	Neutral	Weather forecast
3	Happy	Just For Laughs Gags
4	Happy	Just For Laughs Gags
5	Sad	Wedding Dress
6	Sad	Hachi-A Dog's Tale
7	Angry	Falling Down
8	Angry	Life as a House

audio–visual channels, EEG and peripheral physiological signals recorded in natural noisy room and anechoic chamber.

3. MED4 database establishment

The MED4 database has been recorded separately in two different environments: a lab with natural noises and an anechoic room. 32 subjects first watched film clips contain target emotions (happy, sad, angry and neutral) and then read aloud pre-scripted text materials as the expression of such emotions. Speech, video, EEG and PPG signals were recorded during the experiment, but only the signals when subjects expressing emotions were used to construct the MED4 database.

3.1. Stimuli video clips

Emotional video clips are effective and widely used to induce emotions. To elicit the target discrete emotions (happy, sad, angry, neutral), sixteen films were selected based on the public views on the types of emotions related (e.g., tragedy, comedy). We also referred to the films used in other relevant work such as [27]. The film clips were then edited based on the following criteria: (a) the duration of each film clip should be between one and seven minutes which is suitable for emotion evoked and avoids fatigue caused by excessive duration [28], (b) the emotion and content of the video clips should be easily understood, (c) the clips should show the most emotional arousing part of the film, (d) the video clips should stimulate one single and consistent target emotion. Based on these criteria, sixteen video clips were preliminary collected (four clips for each type of emotion), with average duration of four minutes. Then, five evaluators were invited to rate the emotional content of these sixteen video clips. They were asked to assess the emotions they perceived from the sixteen movie clips presented at randomized order, using Post-Film Questionnaire which consists five items (four related to the target emotions and one question of "What other emotion did you feel except the four emotions listed above") on a five-point scale (one = "not at all", five = "extremely") to indicate the emotion intensity. Two film clips received the highest scores based on the Questionnaire from each emotion were chosen as emotional stimuli, ultimately eight film clips were selected to be shown which are listed in the Table 2.

To study the effect of environmental noise for different modalities of emotional expression, we conducted the experiment separately in natural noisy room and anechoic chamber. The eight film clips were divided into two sessions with no repetition. Half of the subjects were shown session one stimuli in natural noisy room and session two in anechoic chamber, the other half subjects were shown the opposite session.

3.2. Environmental setting

The natural noisy environment condition was conducted in a human factor laboratory. The speech signals collected in this environment contain various types of environmental noises and reverberation. Anechoic chamber is 11.8 *m* long *4.2 *m* wide * 3.8 *m* high. It is equipped with sound absorbing material inside the wall that greatly reduces noise, reverberation and electromagnetic interference. The background noise in the anechoic chamber is less than 17 dB. The speech signals collected from the anechoic chamber are high quality and regarded as more suitable for studying the emotional clues of speech [29].

Previous studies have reported that indoor lighting and temperature can influence affective and cognitive processes [30]. Increasing in illumination levels can increase the level of arousal, and mild heat (up to ~27 ℃) can reduce it. To make sure that environmental noise is the single variable, we set the illumination and temperature of the two labs to be the same and comfortable for the subjects to avoid external intervention.

3.3. Transcript for emotion recognition

There are two main types of transcripts in the acted and elicited databases in speech emotion research. Some studies use emotionally neutral utterances, where subjects read sentences by applying different emotions. The others use emotionally biased sentences, where subjects express their emotion according to contextual information of the scripts. The purpose of MED4 database is to obtain as natural audio, video and physiological signals as possible in a controlled experimental setting. Contextual information is critical for the naturalness of speech [31]. Meanwhile, genuinely emotional speech is likely to contain emotionally biased content. Hence, utterances corresponding to the video stimuli were scripted as the reading script. Considering the stimuli that evoke happy emotions have no words, we edited two dialogs with contextual information, of which some sentences contained emotional words toward happy state, such as " We are so happy and cheerful".

3.4. Apparatus and synchronization

Speech data were captured by one condenser microphone (Audio Technica ATR2500) with a sampling rate of 44.1 kHz and presented in Adobe Audition software. EEG data were continuously collected using a 32-channel EEG module (NeurOne) with electrodes arranged according to international 10–20 system with a sampling rate of 500 Hz. PPG was collected by an ear clip sensor which attached to the earlobe with a sampling rate of 64 Hz and displayed in ErgoLab software. Speech, EEG and PPG signals were recorded on a dedicated PC. E-prime software was installed in another PC that controls the protocol of the experiment, including presenting the stimuli, managing the procedure and synchronizing the data recording. When subjects were ready to read the script, they were asked to press any key on the keyboard to start the recording of speech. The E-prime software sent synchronization markers directly to the EEG, PPG collection software to align the EEG, PPG and speech signals. Facial videos of subjects were collected via webcam of the stimuli PC with a sampling rate of 60 Hz. Fig. 2 illustrates the placement of equipment for data collection.

3.5. Subjects

The subjects who engaged in emotional data collection experiment should have the ability to accurately perceive and express emotions. The ability to feel and share the emotional experiences to another is known as empathy [32], while the Chinese version of Interpersonal Reactivity Index (IRI-C) questionnaire [33] which consists of 22 questions measures the empathic ability of participants. The higher score of IRI-C questionnaire means the better ability of empathy. In addition, Toronto Alexithymia Scale (TAS-20) [34] is the most widely used 20-item self-report questionnaire for measuring the difficulty in distinguishing emotions, describing emotions to others and the absence of externally oriented thinking. The score of TAS-20 less than 51 is considered as free of alexithymia. Thus, we chose the subjects who had IRI-C > = 55 and TAS-20 < = 51 point to participant the data collection experiment. 32 healthy students (sixteen males, sixteen females) from Northwestern Polytechnical University participated in this experiment. Average age of subjects is 24.28 years old (sd = 2.69 years).

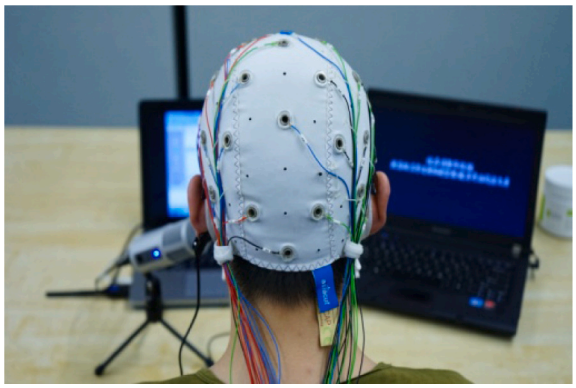


Fig. 2. The placement of equipment.

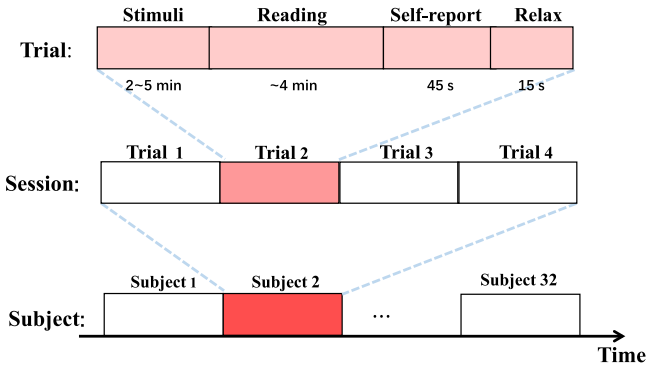


Fig. 3. Protocol of our designed emotion experiments.

3.6. Experimental protocol

From the night before, subjects were informed to ensure adequate sleep and avoid excitable substances, e.g., coffee and alcohol. Upon arrival, prior to each experiment, each subject signed a consent form, which includes the basic demographic information and consents the usage of their personal/experimental data. Then the subject was explained the purpose, the procedure and the self-assessment form in detail, followed by an introduction of the equipment used in the experiment. They were also informed that they can pause the experiment if emotions disappear due to sleepy during reading scripts. By recalling an emotional event [35] or the previous film clip to arouse the target emotion, they can then continue the experiment. In addition, they were asked to watch the film clips attentively and minimize head movements to avoid artifacts in EEG recording. After the sensors were properly placed, their signals were checked before data collection. Then the experimenter existed the lab and the subject started experiment with pressing any key on the keyboard. Each experiment lasted about 60 min.

There was a baseline recording of two minutes, during which EEG and PPG signals were recorded with subject's eyes closed in first one minute and eyes open in the second. Then the four film clips (one session) were presented in four trials, each trial consisting of the following steps:

- Presentation of two-five minutes film clips to induce the target emotion.
- Speech recordings of about four minutes according to the script with the emotion felt from previous video clip.
- 45 s self-assessment for emotions during watching the stimuli and reading.
- 15 s relaxation for the following trial.

EEG, PPG and video signals were recorded throughout each experiment, while speech signals were recorded only during the subject uttering the script. Self-assessment was conducted immediately after each trial. In the self-assessment, participants were asked to assess emotional state that they experienced rather than emotions that were designed to induce. After four trials, the experiment ended, and the subject left the room without talking about the content of experiments to others. The detailed protocol is shown in Fig. 3. We first conducted the experiments in natural noisy room, subsequently experiments in anechoic chamber using the same participants and protocol.

3.7. Database organization

The database contains the EEG, PPG, video and speech emotional signals collected in a lab with natural noises as well as in an anechoic chamber. The signals were annotated with four target discrete emotions based on subjects' self-assessment report after each trial. A brief summary of the experiment content is shown in Table 3.

Table 3
Experiment Content Summary.

Subjects and modalities	
No. of subjects	32, sixteen male and sixteen female
Recorded modalities	32-channel EEG (500 Hz), PPG signals (64 Hz), Speech (44.1 kHz), Video (60 Hz)
Evoke target emotions using video clips	
No. of film clips	eight, four for each environment
Expressing emotions based on scripts	
No. of scripts	eight, two for each emotion
Self-report	Happy, Sad, Angry, Neutral or other
Recorded modalities	Speech, EEG, PPG, Video

Within 64 experiments, data from nine experiments were eliminated, due to poor EEG signals (less than 60% data available). In addition, only the data with consistent target emotion and self-assessed emotion were considered valid. In addition to the self-assessment, we also conducted emotion perceptual evaluation for speech utterances. Two native Chinese annotators who did not participate in the data collection carried out the perceptual evaluation. Since contextual information influences the judgment of evaluators [36], our aim is to understand acoustic cues to decode emotional behaviors, thus each utterance was evaluated by the two annotators in random order. Each utterance was presented to the annotators, and rated on the type of emotions accordingly. Each utterance was given a label only when the two annotators had the same rating and the rating was also consistent with the self-assessment from the subject. Therefore, if the subject's self-assessed emotion was not the same as the target emotion, or if it did not agree with the two annotators' ratings, the data were also discarded. Finally, data from 26 participants in natural noisy room and 29 participants in anechoic chamber were considered as high quality and included in MED4 database, of which 23 participants were recorded in both environments.

Each utterance was considered as one data segment. Synchronizing markers were inserted into EEG and PPG recordings at the beginning and ending of each utterance. The detailed information of the MED4 database is demonstrated in Table 4. In total, the database comprises of 9,504 utterances with an average duration of 2.35 s (sd = 1.09) for natural noisy room and 10,692 utterances in anechoic chamber with an average duration of 2.28 s (sd = 0.99). The length distribution of each utterance is shown in Fig. 4. 90.9% (noisy) and 92.7% (anechoic) of the data duration range from one to four seconds.

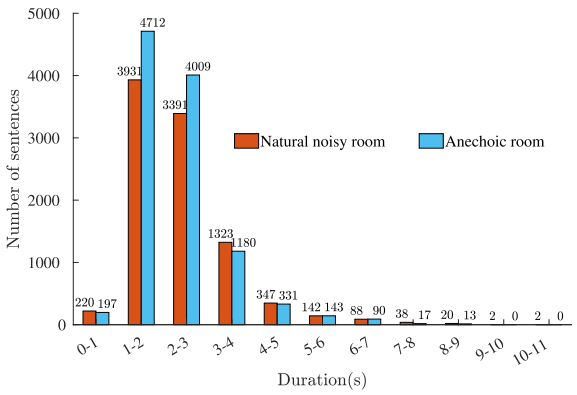


Fig. 4. The duration distribution of utterances in MED4.

Table 4
The detailed information of the database.

	Environments	
	Natural noisy room	Anechoic chamber
Subjects	26	29
EEG (fragments)	4 emotions*26 subjects	4 emotions*29 subjects
PPG (fragments)	4 emotions*26 subjects	4 emotions*29 subjects
Speech (utterances)	Happy:2392, Sad:2575, Angry:2636, Neutral:1901	Happy:2671, Sad:2887, Angry:2942, Neutral:2192

4. Data processing

In this paper, the video and PPG signals were not included in the following AER performance validation algorithms, but will be published along with the rest of the dataset.

4.1. EEG data pre-processing

Raw EEG signals are prone to noise. The signals are pre-processed using MATLAB EEGLAB toolbox to remove artifacts [37]. In this work, a band-pass filter with a bandwidth range from 1 to 50 Hz was first applied, and then a notch filter at 50 Hz was applied to remove linear trends and minimize artifacts. Head moving and scalp sweat cause EEG signals to drift. In an off-line analysis of event-related potential, lack of partial data does not have a significant impact on the results, therefore the segments with EEG data drift can be discarded. However, in some special applications, e.g., in real-time emotion recognition, discarding the data affects AER performance. Thus, an additional median filter was applied to carry out a baseline correction for drifted EEG data. A spherical interpolation method was then used to interpolate the bad channels. Finally, independent component analysis algorithm was applied to remove artifacts related to eye blinks and muscle movements.

For feature extraction of EEG signals, most researches use sliding window to divide the continuous EEG data into segments with equal length [9,38]. Features are then extracted from each segment. In this work, considering the practical application of EEG and speech signal fusion in AER, we cut the pre-processed EEG data into segments with varying length using the Hamming window, where each segment corresponding to the duration of each utterance. The distribution of the duration of each segment would then be the same as speech utterances, showing in Fig. 4.

In previous work [23,39], differential entropy (DE) feature has shown superiority on EEG-based AER. It is the entropy of continuous random variable and measures the complexity of EEG signal. Definition as follows:

$$h(\mathbf{x}) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx = \frac{1}{2}\log 2\pi e\sigma^2 \quad (1)$$

where $\mathbf{x}$ is a random variable with the Gaussian distribution $N(\mu, \sigma^2)$. DE features were then calculated into five frequency bands for each channel: delta rhythm (1–3 Hz), theta rhythm (4–7 Hz), alpha rhythm (8–13 Hz), beta rhythm (14–30 Hz) and gamma rhythm (31–50 Hz). Finally, we extracted 160-dimensional DE features for each sample.

4.2. Speech data processing

Pre-process for speech signals was conducted before feature extraction.

- *Pre-processing*: Pre-emphasis was first conducted to remove the influence of lip radiation with a finite impulse response (FIR) high-pass filter. The transfer function of the pre-emphasis filter is usually given by

  $$H(z) = 1 - az^{-1}, a \in [0, 1]$$

  Here $a = 0.97$, the pre-emphasis can increase the high frequency resolution of the speech signal. Next, voice activity detection (VAD) was applied to remove the silence segment to improve the global accuracy of AER [40]. Since the mel-frequency cepstral coefficient (MFCC) of silence is close to 0, which has no contribution to AER. Due to the short-term stationarity of speech signal, we also need to divide the speech into frames through a sliding window. In addition, the method of overlapping between frames was adopted to make smooth transition. In this work, each utterance was divided into frames of 256 points using the Hamming window with an overlap of 128 points. After pre-emphasis, VAD and framing, each utterance was segmented into several frames, and acoustic features were extracted for each frame.
- *Acoustic feature extraction*: MFCC was first introduced in [41] which is consistent with the human ear perception of sound frequency characteristics. It has been widely and successfully applied to automatic speech recognition and speech-based AER [42,43]. For each frame, the first 12 MFCC parameters and the associated delta- and double-delta MFCCs describing local dynamics were extracted to form a 36-dimensional frame-level feature vector.

Because of the duration of utterances vary, shown in Fig. 4, each utterance has a variant number of frames that results in unfixed dimensionality of feature vector. Generally, by calculating the statistical values like maximum, minimum, mean and variance of all frame-level features, we can get the fixed dimensional representation of utterance-level. However, this process will discard the temporal clues of speech signals, which might be important to recognize emotions [44]. In this paper, I-vector with PLDA algorithm can capture the temporal clues of continuous frames. The final feature vector for one utterance was hence a matrix of $36 \times nframes$ (number of frames in one utterance).

5. Emotion recognition methods

5.1. Identification-vector + probabilistic linear discriminant analysis

I-vector with PLDA algorithm is widely used in speaker verification, and in this study, we used it for speech-based AER [45,46].

- Standard I-vector system use a Universal Background Model (UBM) in conjunction with acoustic features to collect sufficient statistics for I-vector extraction. Firstly, we use all training data to produce an emotion-independent model which named UBM that is a weighted sum of several Gaussian components. Then, the specific emotion-dependent Gaussian Mixture Models (GMMs) were created by adapting the UBM to each emotion using Maximum A Posteriori (MAP) adaptation with the corresponding emotion training data. In this process, we only needed to adapt the mean vectors of UBM (rather than weights or covariances) by

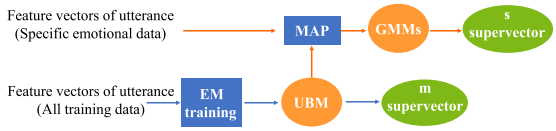


Fig. 5. The model structure of GMM-UBM.

a shift to get the target emotion characteristics. The target GMM supervector was generated from concatenating the mean values of all components GMM. The structure of GMM-UBM model is shown in Fig. 5.

- Since the dimension of the emotion-dependent GMM-UBM supervectors is very high, redundant information may exist. Therefore, dimension was reduced to extract more compact features. In the I-vector paradigm, the mean supervectors of GMM-UBM can be modeled as the sum of the emotion-independent mean supervector and the total variability vector. Formally, the supervector of a given utterance j can be modeled as:

$$\boldsymbol{s}_j = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{\omega}_j \tag{2}$$

where $\boldsymbol{m}$ represents the UBM mean supervector, $\boldsymbol{T}$ is a low dimensional total variability matrix and $\boldsymbol{\omega}_j$ is the I-vector of utterance j and contains specific emotional and channel information.

- Since the existence of channel information during speech transmission in I-vector interferes with AER, channel compensation method should be applied to minimize the influence. PLDA is a popular method to remove the channel attribute from I-vector. PLDA assumes that, for the ith emotion, the I-vector $\boldsymbol{\omega}_{i,j}$ extracted from the jth utterance can be formulated as:

$$\boldsymbol{\omega}_{i,j} = \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{x}_i + \boldsymbol{e}_{i,j} \tag{3}$$

where $\boldsymbol{\mu}$ is the mean vector of all training data, $\boldsymbol{\Phi}\boldsymbol{x}_i$ represents the inter-emotion variability and the residual term $\boldsymbol{e}_{i,j} \sim N(0, \Sigma)$ represents the intra-emotion variability which need to be minimized.

- Given two I-vectors $\boldsymbol{\omega}(j_1)$ and $\boldsymbol{\omega}(j_2)$ extracted from utterances j_1 and j_2, and a trained PLDA model, we can compute the likelihood ratio between the hypothesis H_{same} that the I-vectors belong to the same emotion, versus the hypothesis H_{diff} that the I-vectors were produced by different emotions, with factors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The PLDA score is the log of this quantity, namely

$$score = \ln \frac{p((\boldsymbol{\omega}(j_1), \boldsymbol{\omega}(j_2))|H_{\text{same}})}{p(\boldsymbol{\omega}(j_1)|H_{diff})p(\boldsymbol{\omega}(j_2)|H_{\text{diff}})'} \tag{4}$$

If the score is greater than a given decision threshold, the two utterances belong to the same emotion.

5.2. Extreme learning machine

ELM [47] is known for its fast training and good generalization performance. It has shown competitive accuracy in many pattern recognition applications such as disease diagnosis [48,49], motor imagery [50], face recognition [51] and distraction detection [52]. The standard ELM model is shown in Fig. 6.

Considering a set of N training samples $(\boldsymbol{x}_i, \boldsymbol{t}_i), i = 1, \ldots, N$ for m emotion classes, where each sample and its corresponding label vector are respectively as $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^{\mathrm{T}} \in R^d$ and $\boldsymbol{t}_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^{\mathrm{T}} \in R^m$. For ELM with multi-output nodes, if $\boldsymbol{x}_i$ belongs to the class p, the label vector is denoted as $\boldsymbol{t}_i = [0, \ldots, \underset{p}{1}, \ldots, 0]^{\mathrm{T}}$. In ELM, the input weights $\boldsymbol{\omega}$ and the biases $\boldsymbol{b}$, are randomly generated and fixed, which leads to the analytical calculation of the network outputs weights $\boldsymbol{\beta}$. It can be obtained by solving the following objective function which

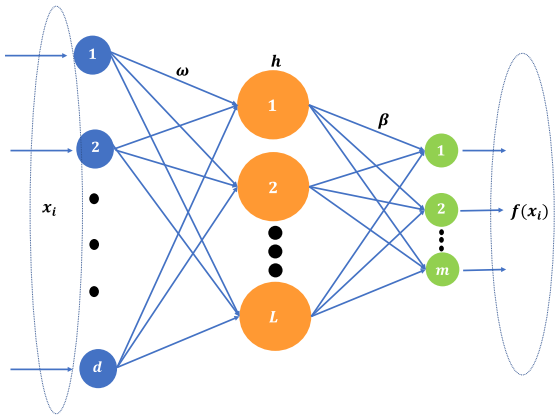


Fig. 6. The model structure of standard ELM.

aims to not only reach the minimum training error but also minimize the output weights

$$\begin{aligned} &\text{Minimize}: L_{\mathrm{p}} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \mathrm{C}\frac{1}{2}\sum_{n=1}^{N}\xi_i^2 \\ &\text{Subject to}: \boldsymbol{h}(\boldsymbol{x}_i)\boldsymbol{\beta} = \boldsymbol{t}_i^{\mathrm{T}} - \boldsymbol{\xi}_i^{\mathrm{T}}, i = 1, \ldots, N \end{aligned} \tag{5}$$

where $\boldsymbol{h}$ is any nonlinear activation function actually mapping the data from d-dimensional input space to the L-dimensional hidden layer feature space. Based on the Karush–Kuhn–Tucker theorem, the output function of ELM classifier is

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}_i) = \boldsymbol{h}(\boldsymbol{x}_i)\boldsymbol{\beta} &= \boldsymbol{h}(\boldsymbol{x}_i)\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\mathrm{C}} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T} \quad \text{or} \\ &= \boldsymbol{h}(\boldsymbol{x}_i)\left(\frac{\boldsymbol{I}}{\mathrm{C}} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T} \end{aligned} \tag{6}$$

For any testing sample $\boldsymbol{y}$, let $\boldsymbol{f}_j(\boldsymbol{y})$ denote the result of the jth output node, i.e. $\boldsymbol{f}(\boldsymbol{y}) = [\boldsymbol{f}_1(\boldsymbol{y}), \ldots, \boldsymbol{f}_m(\boldsymbol{y})]^{\mathrm{T}}$, then the predicted class of sample $\boldsymbol{y}$ is

$$class(\boldsymbol{y}) = \arg \max_{i \in \{1,\ldots,m\}} \boldsymbol{f}_i(\boldsymbol{y}) \tag{7}$$

In ELM, if the feature mapping $\boldsymbol{h}$ is known, almost all nonlinear piecewise continuous functions can be used as the hidden-layer activation functions. Sigmoid and Gaussian functions are two of the major hidden-layer output functions. If $\boldsymbol{h}$ is unknown, we can apply Mercer's conditions on ELM. A kernel matrix can be defined as

$$\begin{aligned} \boldsymbol{\Omega}_{\mathbf{ELM}} = \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} &: \boldsymbol{\Omega}_{\mathbf{ELM}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{h}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{x}_j)^{\mathrm{T}} \\ &= K(\boldsymbol{x}_i, \boldsymbol{x}_j) \end{aligned} \tag{8}$$

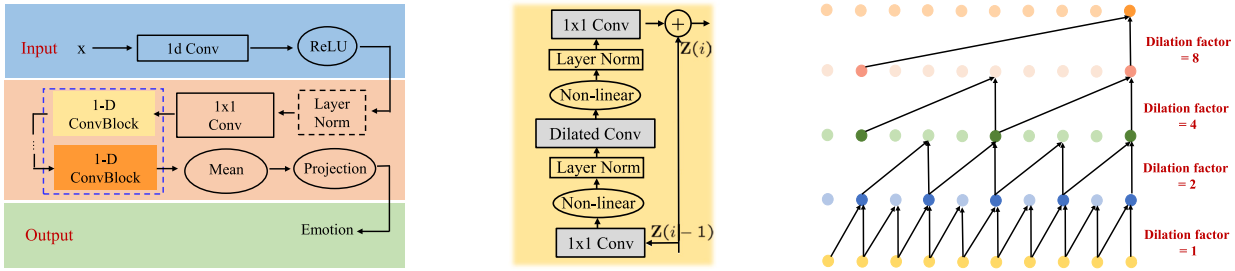
Then, Eq. (6) can be written compactly as

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}_i) &= \boldsymbol{h}(\boldsymbol{x}_i)\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\mathrm{C}} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T} \\ &= \begin{bmatrix} K(\boldsymbol{x}_i, \boldsymbol{x}_1) \\ \vdots \\ K(\boldsymbol{x}_i, \boldsymbol{x}_N) \end{bmatrix}\left(\frac{\boldsymbol{I}}{\mathrm{C}} + \boldsymbol{\Omega}_{\mathbf{ELM}}\right)^{-1}\boldsymbol{T} \end{aligned} \tag{9}$$

From Eq. (9) we can see that the kernel form of ELM classifier is only related to the input data and the number of training samples. The dimensionality L of the feature space (number of hidden nodes) need not be given either.

5.3. Multi-layer perception network

MLP is a widely used artificial neural network and commonly applied in classification task [53,54]. We used MLP as a deep learning method to compare with ELM for EEG-based AER. MLP consists of input layer, hidden layers and output layer. The DE feature extracted from pre-processed EEG signals was fed into MLP as input vector. To



(a) (b) (c)

Fig. 7. The structure of TCN. (a) The flowchart of TCN. The elements in dotted line are repeated 1-D convolution blocks. Different colors in the 1-D convolution blocks denote different dilation factors. (b) The design of 1-D convolution block. It is a residual connection between the input and output. Each 1-D convolution block has a dilated convolution layer. (c) An example of a dilated convolution with dilation factors = 1, 2, 4, 8 and kernel size = 2.

verify the recognition accuracy, several models were tested. For MLP, the number of input nodes was 160 which is the dimension of DE feature, and the number of hidden nodes was selected in {128, 64, 32, 16}, hidden layer from 1 to 2. The parameters which provided best classification performance were set as 2 hidden layers with 128 and 16 hidden nodes separately.

5.4. Temporal convolutional network

Convolutional Neural Network (CNN) and its variation, such as Temporal Convolutional Network (TCN) are commonly used in modeling sequential data [55]. TCN adopts dilated 1-D convolutions to create a large temporal receptive field with fewer parameters. By stacking 1-D convolution blocks with different dilation factors to capture the temporal dependence of various resolutions from the sequential data. Each block has a residual connection between input and output to avoid losing low-level details and to provide hooks for optimization. In addition, the computations in TCN can be performed in parallel to greatly speed up the training process and also significantly reduces the model size.

Due to the excellent performance of TCN-based models in speech enhancement [56] and speech separation [57], we used it for end-to-end speech AER. Referring to the research in [57] which studied the effect of different configurations in TCN for speech separation, we determined the model parameters of TCN with 4 Convolution Blocks and dilation factors are {1, 2, 4, 8} respectively. The structure of TCN that used in this paper is shown in Fig. 7.

5.5. Fusion methods

Multi-modal fusion can improve the performance of AER. To combine the EEG and speech modalities, two fusion strategies were developed in this work, i.e., feature-level and decision-level fusion. Feature-level fusion is achieved by concatenating the features from each modality to form a new feature. Fusion at decision-level is obtained by fusing the output from the single model classifier.

For feature-level fusion, I-vector was extracted as the feature of speech signal and concatenated with the DE vector of EEG, and then fed into the ELM classifier. In the decision-level fusion, the outputs generated by two classifiers of the speech and EEG recognition were combined. To explore an effective method for decision-level fusion, we investigated the following four approaches [58]:

1. AVER: Suppose the outputs of AER classifiers based on EEG and speech signals were normalized to [0,1] and denoted separately as $\boldsymbol{e} = [e_1, e_2, e_3, e_4]$ and $\boldsymbol{s} = [s_1, s_2, s_3, s_4]$, where e_i, s_i can be considered as the probabilities that EEG and speech segments labeled as the ith emotional state. AVER method calculated the average sum of the probabilities which is denoted as p_i, and predicted the final label c as follows:

$$p_i = \frac{1}{2}(e_i + s_i), i = 1, \ldots, 4$$

$$c = \arg\max_i(p_i)$$

2. WAVER: The validation accuracies of models that based on EEG and speech signals on a validation set were denoted as $\boldsymbol{a} = [a_1, a_2]$. Then, WAVER method calculated the weighted sum of the probabilities which is denoted as p_i, and predicted the final label c as follows:

$$p_i = a_1 * e_i + a_2 * s_i, i = 1, \ldots, 4$$

$$c = \arg\max_i(p_i)$$

3. RF: The outputs of the two classifiers in training set were concatenated into a new vector which is denoted as $\boldsymbol{p}_{\text{train}} = [e_1, e_2, e_3, e_4, s_1, s_2, s_3, s_4]$ to training a random forest [20].
4. ET: Similarly, the new score vector $\boldsymbol{p}_{\text{train}}$ was taken as input to training an extremely randomized trees [21].

For the four decision-level fusion approaches, AVER and WAVER are linear methods while RF and ET are nonlinear fusion methods.

6. Results

In this section, we present the performances of single modality and multi-modal AER separately on the MED4 database. In the experiments given below, we employed five-fold subject independent cross-validation in which one-fold data for testing and the remaining data were used for training the classifiers.

6.1. Effect of window size with equal- and variable-length for EEG

Since EEG signals are generally non-stationary and some analysis techniques such as spectral analysis can be performed only for stationary data, signal segmentation by window function was adopted [59]. Aya et al. [60] used KPSS test to determine the best size of window function to segment EEG signals for Seizure detection. Different window sizes were tested on the desired signal, the one with the minimum number of non-stationary segments was chosen as best window size. To analysis the effect of different window size on AER, we compared the classification performance using DE feature across all EEG frequency bands with equal- and variable-length time windows. In the case of equal length segments, the window size ranged from one second to fourteen seconds with 0.5s interval, without overlapping. For the variable length segments, as mentioned in Section 3.7, the window size equals to the duration of corresponding utterance. The ELM classifier

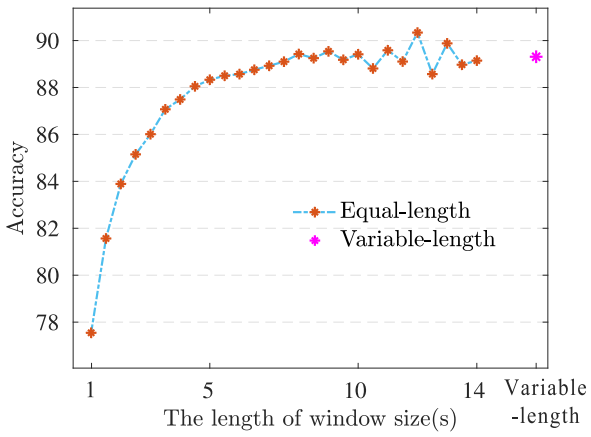


Fig. 8. The effect of window size for EEG signal.

with radial basis function kernel was applied to identify emotions in each dataset. The average accuracies are presented in Fig. 8, the results can be summarized as:

- 1s–5s: When the window size increased from 1s to 5s, and the accuracies rapidly increased from 77.55% to 88.33%.
- 5s–14s: The accuracies stayed stable, between 88.33% (at 5s) and 90.34% (at 12s).
- Variable-length: The average accuracy of the EEG segments intercepted by variable-length window function is 89.31%. The recognition accuracy was higher than that with equal-length window size.

The result in case of equal-length segments is consistent with the previous study of Henry et al. [61], where wavelet features were extracted from EEG at different time segments and fed into SVM algorithm for classifying between high/low arousal and high/low valence. The result demonstrated that the segments with 3–12 s contained the most emotion information. The accuracy using data with 1–3 s duration has a slight reduction. The results from this work showed that EEG signals with variable length corresponding to speech utterances are more appropriate for emotion recognition.

6.2. Classification on single modality

From the above results, variable lengths of EEG segments contain effective information for AER. Therefore, in the following experiment, EEG samples are of variable lengths. The results of single modality are shown in Table 5. Compared to MLP and TCN, ELM and I-vector+PLDA achieve 6.63% and 7.06% accuracies improvements in anechoic chamber, and 8.99% and 7.99% in natural noisy room for EEG- and speech-based AER separately. From the results of ELM and I-vector+PLDA, both speech and EEG signals provided useful information for AER. Average recognition accuracies of 64.67% and 58.92% across four target emotions are achieved in different scenes respectively using speech data. EEG signals achieve relatively high accuracies of 88.92% and 89.70% in anechoic chamber and noisy lab.

We also analyzed the confusion matrices and graphs from the results of ELM and I-vector+PLDA that using EEG and speech modality separately, shown in Fig. 9 and Fig. 10, to further explore the complementary information between EEG and speech signals. In general, EEG achieves higher recognition performances for all four emotions, especially for happy emotion with 91.76% and 91.32% accuracies, compared to speech signals (51.52% and 35.97%). The research in [13] also summarized that valence shows the strongest correlation with EEG signals. Thus, EEG data have better performance in classifying high valence emotion (happy) and low valence emotions (sad and angry). Fig. 9(a) and (b) present the confusion matrices of the speech-based AER. It can be observed that speech signals have advantage in recognizing neutral and angry emotions. In addition, speech are more representative for low valence emotions like angry and sad, but less effective for emotions with high valence levels like happy. For example, the accuracies of angry are much higher than happy, 70.17% vs. 51.52% in anechoic chamber and 64.12% vs. 35.97% in natural noisy room, respectively. The research [62] also verified the same conclusion. This is probably due to the difference of the culture, environment and education of speaker leads happy emotion is easily misclassified as other emotion without the help of linguistic information [63]. This finding has useful implications on developing AER applications.

Table 5
Performance of each signal modality.

Modality	Method	Anechoic	Natural
EEG	ELM	**88.92**	**89.70**
	MLP	82.29	80.71
Speech	I-vector + PLDA	**64.67**	**58.92**
	TCN	57.61	50.93

Table 6
Performances of feature- and the decision-level fusion methods.

Environment	Fusion Method				
	Feature-level	AVER	WAVER	RF	ET
Anechoic chamber	89.04	87.45	**90.59**	82.91	87.22
Natural noisy room	89.65	85.86	**90.66**	82.56	87.47

Meanwhile, Fig. 10 also shows that the misclassification character of these two modalities is similar. Both speech and EEG easily misclassifies sad and neutral emotions, sad and angry emotions. As indicated by these results, speech and EEG signals have some complementary and similar characteristics in emotion representation.

6.3. Effect of environmental noise

Table 5 and Fig. 9 also show the impact of environmental noise on AER. There is a benefit of acquiring clean signals for speech emotion recognition, as indicated by the improved accuracy from 58.92% to 64.67% in anechoic chamber data. From Fig. 9 (a) and (b), it can be seen that environmental noise has a great influence on the happy emotion recognition and there is an increased performance with 15.55% from natural noisy room to anechoic chamber. In natural noise lab, more happy emotion is misclassified into neutral emotion which shown in Fig. 10. This result is similar to that in simulated noisy speech data. Huang et al. [64] studied the influence of white noise on AER by adding the noise to clean speech with different SNRs. They observed that with the reduction of SNRs, the voice quality features may be distorted by white noise which caused the lowest detection accuracy of the positive valence.

However, EEG-based algorithms showed robustness across types of emotions and environments. Compared with other signal channels, such as images, speeches, and other types of biometric, EEG-based systems are more efficient, robust and give higher performance, as its sophisticated and directly reflect brains' inherent status [65–67]. Especially, in Pieper's study [68], it was found that there was no significant difference in the power spectral density of EEG signals in quiet and artificial noise environments. Therefore, it will be beneficial for AER to reduce the noise for speech signals but natural environmental noises are negligible for AER using EEG signals.

6.4. Classification on multi-modality

The average accuracies of feature- and decision-level fusion strategies are shown in Table 6. The decision-level fusion is combining the

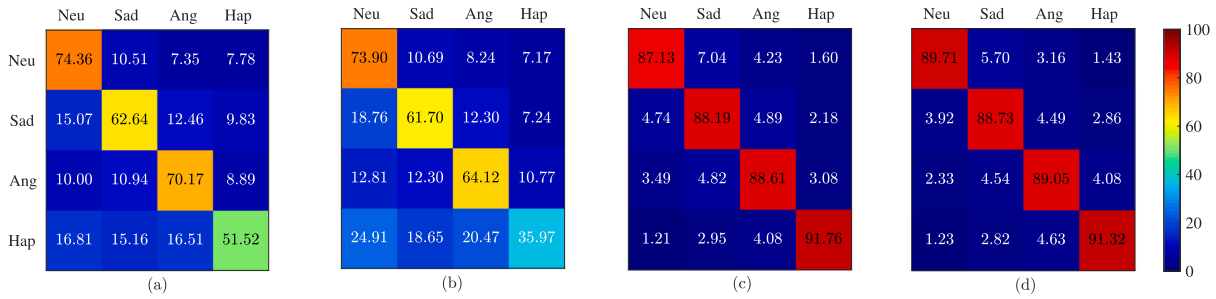
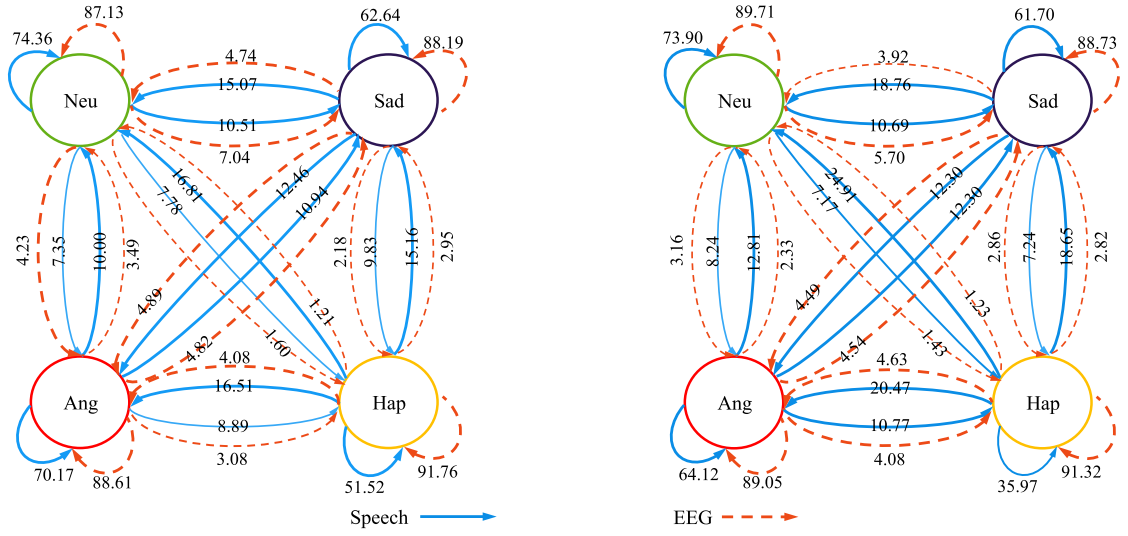


Fig. 9. Confusion matrices of single modality. Each row of the confusion matrices is the true emotion label and each column is the predicted label. Where Neu is Neutral, Ang is Angry and Hap is Happy. (a) Anechoic speech. (b) Natural noisy speech. (c) Anechoic EEG. (d) EEG from natural noisy room.



(a) signals from anechoic chamber (b) signals from natural noisy room

Fig. 10. Confusion graphs of Speech and EEG signals from different environments. The numbers denote the classification accuracy that classify the samples from arrow tail emotion to arrow head emotion. Bolder lines indicate higher values.

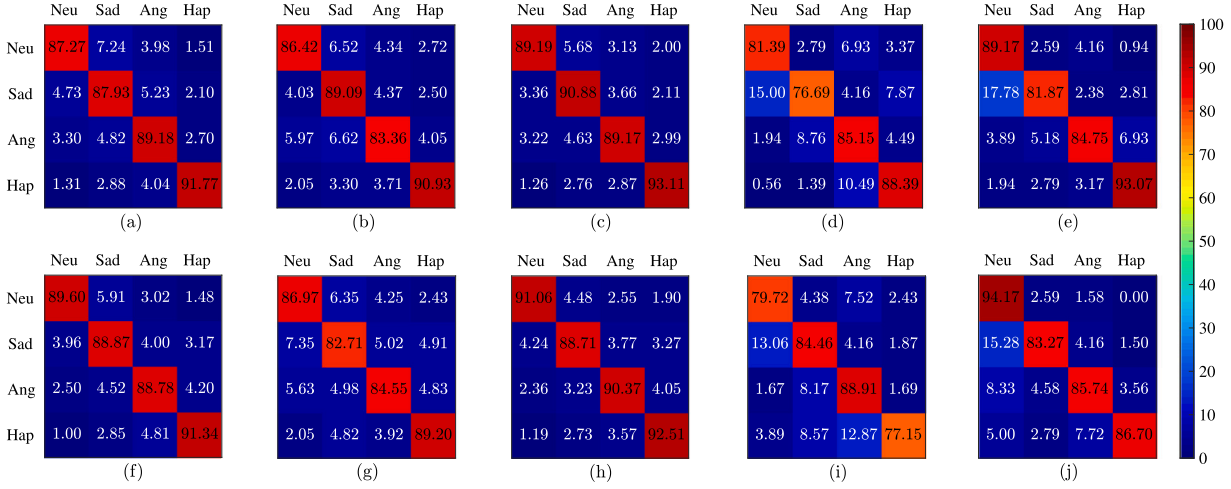


Fig. 11. Confusion matrices of feature- and decision-level fusion methods. The first row represents the fusion results of signals from anechoic chamber, the second row represents the results that signals from natural noisy environment. From left to right, the fusion strategies are feature-level, AVER, WAVER, RF, ET.

outputs from ELM and I-vector+PLDA algorithms. It shows that the WAVER method performs the best in both environments and achieves the average accuracy of 90.59% and 90.66% separately, followed by feature-level fusion, ET, AVER and RF methods.

To further understand the performance of each fusion method, we show the confusion matrices in Fig. 11. From Fig. 11(a) and (f), we can see the performance of feature-level fusion approach is similar to that of single EEG modality. Concatenating MFCCs and DE features do not improve the recognition accuracies. The results of nonlinear decision-level fusion approaches shown in Fig. 11(d), (i), (e) and (j) reveal that RF and ET methods, especially RF, have lower AER accuracies and are easily affected by environmental noise. RF performs best

in recognizing happy and angry emotions using data from anechoic chamber. In addition, the accuracy has a significant decrease from 88.39% (anechoic chamber) to 77.15% (natural noisy room) for happy emotion recognition, which is similar to the performance of speech modality. In the case of linear decision-level approaches, the weights in WAVER method are the proportions of recognition accuracies of EEG and speech single modality, which emphasizes the importance of EEG signals, therefore WAVER outperformed AVER approach.

Compared to single modality, WAVER fusion across EEG and speech signals increase the accuracy further. In anechoic chamber, 25.92% and 1.67% improvements achieved for decision-level fusion method compared to speech and EEG modality, respectively. The improvements for AER are 31.74% and 0.96% in natural noisy room. Furthermore, Fig. 11(c) and (h) shows that WAVER method enhances the classifying accuracy for each emotion, especially has the best performance in classifying happy emotion. Meanwhile, the classification accuracies of each emotion are not affected by environmental noise. The results demonstrate the effectiveness of combining speech and EEG signals for AER in practical application. The WAVER fusion method not only improves the classification accuracy of each emotion, but also integrates the advantages of EEG for recognizing happy emotion while simultaneously overcoming the influence of environmental noise in speech-based AER. The similar result was achieved from the work in [69], they added white Gaussian noise to corrupt the visual data and audio data were left uncorrupted. The decision-level fusion recognition rate of the combined audio–visual modalities is higher than that of single modality and demonstrates the robustness toward noisy visual data.

7. Conclusion

In this paper, we constructed an emotional database MED4. To our knowledge, MED4 is the first multi-modal and multi-environmental emotion database which has four modalities of synchronized speech, video, PPG and EEG signals, recorded in both a natural noisy room and an anechoic chamber.

Based on the MED4 database, we firstly tested the effect of variable-length EEG on AER performance, and then we performed single modality emotion recognition based on speech and EEG signals.

The results showed that EEG signals with variable length corresponding to speech utterances outperforms other methods in emotion recognition. The confusion matrices and graphs for single modality showed that EEG signals achieved high accuracy in AER, especially in recognizing happy emotion, and speech signals were more effective in recognizing neutral and angry emotions. The average classification accuracies based on EEG and speech signals separately were 89.70% vs 58.92% (in natural noisy room) and 88.92% vs 64.67% (in anechoic chamber). Even with a lower recognition accuracy, speech signals can still useful in applications because of its easy accessibility.

Furthermore, the effect of environmental noise for AER using single modality was analyzed. There was a 5.75% average accuracy decrease from clean speech to noisy speech. Especially for the recognition of happy state from speech signals, environmental noise had a great influence with a 15.55% reduction in recognition accuracy. However, the performances were stable using EEG signals across different environments. Thus, we can conclude that EEG modality is a reliable source for AER in suboptimal acoustic environment. These results provide insights to future applications when considering different channels, environments and for different application scenarios. For example, sad and neutral emotions are easily confused through speech AER, and such signals are sensitive to noise. Therefore, such limitations should be bear in mind and optimized for system design.

Finally, we studied the complementary information contained in speech and EEG signals through feature- and decision-level fusion strategies. The experimental results revealed fusion at decision-level with linear models, i.e., WAVER method enhance the emotion recognition accuracies and eliminate the effects of environmental noise. In addition, the WAVER method can ensure the effectiveness of the emotion recognition system even when one modality is missing. Feature-level fusion method did not improve the recognition accuracies over single EEG modality. Nonlinear decision-level methods, RF and ET strategies, did not improve the accuracy of AER and were unstable to environmental noise that is similar to the performance of speech modality.

This work contributes a useful database resource to various communities for AER from single modality to multiple modalities. The results in this paper also provide a benchmark to evaluate the effects of speech noise reduction techniques in various environments. The findings on the interaction among different signal modalities, target emotions and environments reveal useful experimental support for the design of future AER system.

Our future work will include the following directions. Firstly, in this paper, we note that combining EEG and Speech signals by equal weights does not guarantee improved performance. Thus, the information complementarity across all signal modalities including external behaviors, will be thoroughly investigated. Secondly, channel selection and online classification algorithms should be explored with the consideration of performance, portability and system robustness. Thirdly, the influence of environmental noise on AER based on different channels will be conducted to develop systems that are robust to environmental noise. Furthermore, cross-domain research can be conducted by combining MED4 database with other databases through adaptive methods to develop algorithms that are robust to cross-environment conditions.

Given the potential of this database in the field of AER, MED4 will be publicly available for researchers globally.

CRediT authorship contribution statement

Qian Wang: Conceptualization of this study, Methodology, Software. **Mou Wang:** Conceptualization of this study, Methodology, Software. **Yan Yang:** Conceptualization of this study, Supervision, Funding acquisition. **Xiaolei Zhang:** Conceptualization of this study, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Keltner, A.M. Kring, Emotion, social function and psychopathology, Rev. Gen. Psychol. 2 (3) (1998) 320–342.
[2] S. Kaplan, J. Cortina, G. Ruark, K. Laport, V. Nicolaides, The role of organizational leaders in employee emotion management: A theoretical model, Leadership Quart. 25 (3) (2014) 563–580.
[3] Z. Wang, J.B. Walther, J.T. Hancock, Social identification and interpersonal communication in computer-mediated communication: What you do versus who you are in virtual groups, Hum. Commun. Res. 35 (1) (2010) 59–85.
[4] I.B. Mauss, M.D. Robinson, Measures of emotion: A review, Cogn. Emot. 23 (2) (2009) 209–237.
[5] Y. Jiang, W. Li, M.S. Hossain, M. Chen, M. Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, Inf. Fusion 53 (2020) 209–221.
[6] S. Shimojo, L. Shams, Sensory modalities are not separate modalities: plasticity and interactions, Curr. Opin. Neurobiol. 11 (4) (2001) 505–509.
[7] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE Trans. Affect. Comput. 3 (1) (2012) 5–17.
[8] S. Koelstra, DEAP: A database for emotion analysis; Using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2012) 18–31.
[9] M. Soleymani, S. Asghari-Esfeden, Y. Fu, M. Pantic, Analysis of EEG signals and facial expressions for continuous emotion detection, IEEE Trans. Affect. Comput. 7 (1) (2016) 17–28.

[10] M.Y. Tsalamlal, M. Amorim, J. Martin, M. Ammi, Combining facial expression and touch for perceiving emotional valence, IEEE Trans. Affect. Comput. 9 (4) (2018) 437–449.
[11] S. Zhalehpour, O. Onder, Z. Akhtar, C.E. Erdem, BAUM-1: A spontaneous audio-visual face database of affective and mental states, IEEE Trans. Affect. Comput. 8 (3) (2017) 300–313.
[12] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Topics Signal Process. 11 (8) (2017) 1301–1309.
[13] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Inf. Fusion 59 (2020) 103–126.
[14] S. Zhao, G. Jia, J. Yang, G. Ding, K. Keutzer, Emotion recognition from multiple modalities: Fundamentals and methodologies, IEEE Signal Process. Mag. 38 (6) (2021) 59–73.
[15] M.M. Rahman, A.K. Sarkar, M.A. Hossain, M.S. Hossain, M.R. Islam, M.B. Hossain, J.M. Quinn, M.A. Moni, Recognition of human emotions using EEG signals: A review, Comput. Biol. Med. 136 (2) (2021) 104696.
[16] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: International Conference on Multimodal Interaction, 2015, pp. 423–426.
[17] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, AVEC 2016 : Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International Workshop Audio/Visual Emotion Challenge, 2016, pp. 3–10.
[18] B. Schuller, D. Arsic, F. Wallhoff, G. Rigoll, Emotion recognition in the noise applying large acoustic feature sets, in: International Conference on Speech Prosody, 2006, pp. 276–289.
[19] A.R. Avila, Z. Akhtar, J.a.F. Santos, D. O'Shaughnessy, T.H. Falk, Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild, IEEE Trans. Affect. Comput. 12 (1) (2021) 177–188.
[20] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[21] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.
[22] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.
[23] W.L. Zheng, W. Liu, Y. Lu, B.L. Lu, A. Cichocki, EmotionMeter: A multimodal framework for recognizing human emotions, IEEE Trans. Cybern. 49 (3) (2018) 1110–1122.
[24] H.C. Chou, W.C. Lin, L.C. Chang, C.C. Li, H.P. Ma, C.C. Lee, NNIME: The NTHU-ntua Chinese interactive multimodal emotion corpus, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 2017, pp. 292–298.
[25] O. Perepelkina, E. Kazimirova, M. Konstantinova, RAMAS: Russian multimodal corpus of dyadic interaction for studying emotion recognition, in: International Conference on Speech and Computer, 2018, pp. 501–510.
[26] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2012) 42–55.
[27] W. Zheng, B. Dong, B. Lu, Multimodal emotion recognition using EEG and eye tracking data, in: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 5040–5043.
[28] T. Wang, Y. Zhao, Y. Xu, Z. Zhu, Comparison of response to Chinese and western videos of mental-health-related emotions in a representative Chinese sample, PeerJ 9 (2) (2021) e10440.
[29] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, G. Rigoll, Speaker independent speech emotion recognition by ensemble classification, in: IEEE International Conference on Multimedia and Expo, 2005, pp. 864–867.
[30] X. Dong, Y. Wu, X. Chen, H. Li, B. Cao, X. Zhang, X. Yan, Z. Li, Y. Long, X. Li, Effect of thermal, acoustic, and lighting environment in underground space on human comfort and work efficiency: A review, Sci. Total. Environ. 786 (2021) 147537.
[31] A. Tawari, M. Trivedi, Speech emotion analysis: Exploring the role of context, IEEE Trans. Multimedia 12 (2010) 502–509.
[32] S.D. Preston, F.B.M. De Waal, Empathy: Its ultimate and proximate bases, Behav. Brain Sci. 25 (01) (2002) 1–20.
[33] Z. Fengfeng, D. Yi, W. Kai, Z. Zhiyu, X. Runfang, Study on the reliability and validity of the Chinese version of the interpersonal response indicator scale (IRI-c), Chin. J. Clin. Psychol. 18 (02) (2010) 155–157.
[34] J.Y. Yi, S.Q. Yao, X.Z. Zhu, The Chinese version of the TAS-20: reliability and validity, Chin. Ment. Health (2003) 763–767.
[35] N. Amir, S. Ron, N. Laor, Analysis of an emotional speech corpus in hebrew based on objective criteria, in: Proceedings of ISCA Workshop on Speech and Emotion, 2000, pp. 29–33.
[36] R.T. Cauldwell, Where did the anger go? the role of context in interpreting emotion in speech, in: Proceedings of ISCA Workshop on Speech and Emotion, 2000, pp. 127–131.
[37] A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics, J. Neurosci. Methods 134 (1) (2004) 9–21.
[38] Y. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, Y. Shi, Real-time movie-induced discrete emotion recognition from EEG signals, IEEE Trans. Affect. Comput. 9 (4) (2018) 550–562.
[39] R.N. Duan, J.Y. Zhu, B.L. Lu, Differential entropy feature for EEG-based emotion classification, in: 6th International IEEE/EMBS Conference on Neural Engineering (NER), 2013, pp. 81–84.
[40] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, in: 9th European Conference on Speech Communication and Technology, 2005, pp. 1517–1520.
[41] S.B. Davis, P. Mermelstein, Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences, IEEE Trans. Signal Proces. 28 (4) (1980) 357–366.
[42] S. Wu, T.H. Falk, W.Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Commun. 53 (5) (2011) 768–785.
[43] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Trans. Affect. Comput. 6 (1) (2015) 69–75.
[44] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimedia 20 (6) (2018) 1576–1590.
[45] P. Matejka, O. Glembek, F. Castaldo, M.J. Alam, P. Kenny, L. Burget, J. Cernocky, Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4828–4831.
[46] H.K. Vydana, P.P. Kumar, K.S.R. Krishna, A.K. Vuppala, Improved emotion recognition using GMM-ubms, in: International Conference on Signal Processing and Communication Engineering Systems (SPACES), 2015, pp. 53–57.
[47] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: Theory and applications, Neurocomputing 70 (13) (2006) 489–501.
[48] R. Harikumar, C.G. Babu, M.G. Shankar, Extreme learning machine (ELM) based performance analysis and epilepsy identification from EEG signals, IETE J. Res. (2021) 1–11.
[49] B. Shi, H. Ye, L. Zheng, J. Lyu, C. Chen, A.A. Heidari, Z. Hu, H. Chen, P. Wu, Evolutionary warning system for COVID-19 severity: Colony predation algorithm enhanced extreme learning machine, Comput. Biol. Med. 136 (2021) 104698.
[50] Z. Jin, G. Zhou, D. Gao, Y. Zhang, EEG classification using sparse Bayesian extreme learning machine for brain-computer interface, Neural Comput. Appl. 32 (11) (2020) 6601–6609.
[51] W. Zong, G. Huang, Face recognition based on extreme learning machine, Neurocomputing 74 (16) (2011) 2541–2551.
[52] Q. Wang, Y. Yang, J. Chen, J. He, H. Zuo, W. Zhang, Driver motion detection using online sequential learning, in: 18th COTA International Conference of Transportation, 2018, pp. 315–320.
[53] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, X. Chen, EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network, Knowl. Based Syst. 205 (2020) 106243.
[54] V.M. Joshi, R.B. Ghongade, EEG based emotion detection using fourth order spectral moment and deep learning, Biomed. Signal Process. Control 68 (2) (2021) 102755.
[55] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018, arXiv:1803.01271.
[56] J. Lin, A.J.d.L. van Wijngaarden, K.C. Wang, M.C. Smith, Speech enhancement using multi-stage self-attentive temporal convolutional networks, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 3440–3450.
[57] Y. Luo, N. Mesgarani, Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (8) (2019) 1256–1266.
[58] M. Wang, R. Wang, X. Zhang, S. Rahardja, Hybrid constant-q transform based CNN ensemble for acoustic scene classification, in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 1511–1516.
[59] Y. Shen, Y. Lin, Challenge for affective brain-computer interfaces: Non-stationary spatio-spectral EEG oscillations of emotional responses, Front. Hum. Neurosci. 13 (2019) 366.
[60] A. Alzahy, M. Elgammal, H. Mohammed, H. Mostafa, Optimal EEG window size for neural seizure detection, in: 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), 2019, pp. 1–4.
[61] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, S. Su, Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine, in: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 7250–7253.
[62] A. Amjad, L. Khan, H.T. Chang, Effect on speech emotion classification of a feature selection approach using a convolutional neural network, PeerJ Comput. Sci. 7 (2021) e766.
[63] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D and 2D CNN LSTM networks, Biomed. Signal Process. Control 47 (2019) 312–323.
[64] C. Huang, C. Guoming, Y. Hua, B. Yongqiang, Z. Li, Speech emotion recognition under white noise, Arch. Acoust. 38 (4) (2013) 457–463.

[65] Y. Wang, K.C. Veluvolu, J.-H. Cho, M. Defoort, Adaptive estimation of EEG for subject-specific reactive band identification and improved ERD detection, Neurosci. Lett. 528 (2) (2012) 137–142.
[66] A.A. Alariki, A.W. Ibrahimi, M. Wardak, J. Wall, A review study of brian activity-based biometric authentication, J. Comput. Sci. 14 (2) (2018) 173–181.
[67] F. Llanos, Z. Xie, B. Chandrasekaran, Biometric identification of listener identity from frequency following responses to speech, J. Neural Eng. 16 (5) (2019).
[68] K. Pieper, R.P. Spang, P. Prietz, S. Möller, E. Paajanen, M. Vaalgamaa, J.N. Voigt Antons, Working with environmental noise and noise-cancelation: A workload assessment with EEG and subjective measures, Front. Neurosci. 15 (2021).
[69] K.P. Seng, L.M. Ang, Video analytics for customer emotion and satisfaction at contact centers, IEEE Trans. Hum. Mach. Syst. 48 (3) (2018) 266–278.