INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotion Classification with EEG Responses Evoked by Emotional Prosody of Speech

Zechen Zhang[1], *Xihong Wu*[1], *Jing Chen*[1,2]

[1]School of Intelligence Science and Technology, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China
[2]National Biomedical Imaging Center, College of Future Technology, Peking University

1800013048@pku.edu.cn, wxh@cis.pku.edu.cn, chenj@cis.pku.edu.cn

Abstract

Emotion classification with EEG responses can be used in human-computer interaction, security, medical treatment, etc. Neural responses recorded via EEG can reflect more direct and objective emotional information than other behavioral signals (i.e., facial expression...). In most previous studies, only features of EEG were used as input for machine learning models. In this work, we assumed that the emotional features included in speech stimuli could assist in emotion recognition with EEG when the emotion is evoked by the emotional prosody of speech. An EEG data corpus was collected with specific speech stimuli, in which emotion was represented with only speech prosody and without semantic context. A novel EEG-Prosody CRNN model was proposed to classify four types of typical emotions. The classification accuracy can achieve at 82.85% when the prosody features of speech were integrated as input, which outperformed most audio-evoked EEG-based emotion classification methods.

Index Terms: emotion classification, EEG, emotional prosody, multi-modal learning

1. Introduction

Emotion comes from people's interoception. It is a series of subjective psychological experiences formed from a combination of people's feelings, perceptions, and behaviors. It is an important part of people's cognitive process [1, 2].

Traditional emotion recognition methods mainly use external signals such as facial expressions, body movements, voice and text [3, 4]. Although these signals are relatively easy to collect, they also have some obvious defects. These signals are greatly affected by personal habits and cultural background. They can also be easily disguised. It has been reported that these signals are not reliable because they are not specifically linked with certain emotions. For example, the same face on different bodies can induce different emotions [5]. Besides, such methods are not applicable to specific disabled persons [6].

1.1. EEG Emotion Recognition

To avoid the defects above, physiological signals were used to decode emotions, including respiratory signals, electroencephalogram (EEG), electrocardiogram (ECG), eye movement, etc. These physiological signals can reflect the emotional state of the subjects more objectively [7]. Researchers have reported that EEG-based emotion recognition is more stable, reliable, and accurate than other physiological signals through a large amount of literature research since the EEG signals were collected from the central nervous system of the brain that represents the emotion processing directly and objectively [8].

However, EEG-based emotion recognition also has some setbacks. Firstly, the accuracy of emotion classification still needs to be improved from the current SOTA models on four-category classification [3]. Previous studies on audio-evoked EEG emotion recognition generally used machine learning methods like SVM, kNN, or MLP, and the performance is limited [9, 10]. Secondly, cross-subject emotion recognition becomes a difficult task due to the diversity of people's responses and emotional states under certain conditions [11]. Lastly, rich emotional information in visual and audio stimuli are not fully exploited and used in the EEG-based emotion recognition, while other physiological signals such as ECG or eye movement have been used to form multi-modal emotion recognition model with EEG in previous studies [12].

Speech communication is a common application scenario for emotion recognition, and speech prosody is highly dominated by the speaker's emotions. However, studies or datasets on EEG emotion recognition evoked by emotional prosody of speech are quite rare. Meanwhile, a lot of speech emotion databases are not fully exploited in EEG experiments [13]. In daily conversations, only auditory stimuli are available in many scenarios. For example, when emotion surveillance is applied to drivers, the emotion evoked by passengers' speech could be recognized with the help of speech prosody. Hence, it was worth including speech prosody in EEG studies.

1.2. Emotional Prosody

Emotional prosody is a non-verbal expression of emotion, which has important voice clues and unique acoustic features, including fluctuations in pitch, loudness, speech rate, etc. [14]. These features can reflect certain emotion encoded in speech. Studies have found significant differences in the statistical magnitudes of pitch (i.e., mean value, standard deviation, etc.), loudness, and other features across different emotional speeches in English [15]. In 1996, Banse and Scherer proposed that the speech of each emotion category has its unique "acoustic profile" [16]. Since emotion evoked by speech could result from both the speech semantic and the emotional prosody, researchers have created meaningless sentences to enable closer studies of single emotional prosody. The sentences produced conform to the general rules of natural languages in syntactic structure and pronunciation, but they don't have any semantic content. Scherer et al. used phonemes in different languages to form non-existent words, mixed them with ordinary sentences, and recorded speech spoken by actors with emotional speech prosody. Subjects' emotion was successfully induced without the interference of semantic context [17].

Recently, a similar emotional prosody speech corpus based on Mandarin Chinese was released [18]. The method of con-

10.21437/Interspeech.2023-412

structing meaningless sentences in Chinese is slightly different from that in English [19]. Researchers replace the content words in sentences with words that have no meaning or are irrelevant to the content of sentences while retaining function words to ensure that meaningless sentences still conform to Chinese grammar [20]. The recording of the emotional prosody stimuli using meaningless sentences was produced by native speakers (emotion encoders) who have experience in broadcasting. Then the recorded speeches were verified by a group of listeners who don't know the emotion labels of these speeches. According to the listeners' feedback, speech stimuli with high emotional intensity are selected for a data set [7, 21].

1.3. Multi-modal Emotion Recognition

Researchers have been trying to use emotion cues from different modalities to achieve higher performances on emotion recognition tasks. Visual and audio multi-modal emotion recognition is mostly adopted, due to the plenty of databases and well-performed models. There are many open audio-visual emotion databases available, such as IEMOCAP [22]. Audio and text are also used together with visual information to achieve better results. As for physiological signals, the DEAP dataset (A Database for Emotion Analysis using Physiological Signals) recorded EEG, electrooculogram (EOG), and many other signals collected from the human body with subjects' annotation scores of valance and arousal. Multi-modal emotion recognition using physiological signals can reach an accuracy higher than 90% on binary classification related to valance and arousal [23]. Based on the good performances of existing multi-modal emotion recognition models, it is reasonable to assume that the speech prosody included can enhance emotion classification performance in EEG experiments.

2. Material and Methods

2.1. Data set

A new EEG emotion data set was built by using speech selected from the Mandarin Chinese Speech Emotional Stimulation Database (MCAESD) [18]. This speech corpus consists of different emotion speech of Chinese meaningless sentences. Six students (3 female) with broadcasting experience encoded emotions and recorded audio stimuli according to sentence patterns (declarative sentences or interrogative sentences) and emotional intensity (strong or normal) for each of the sentences. The data set includes the annotation results of 40 subjects (decoders) of each speech stimulus, and statistical results of different emotional speech stimuli are provided.

The speech stimuli contain rich material of four commonly used emotion types in classification [18], neutral, sad, fear, and happy. Its validation accuracy by listeners also meets the requirements for EEG emotional stimuli. Thus, stimuli of these emotion types are selected for the emotion prosody dataset.

Speech stimuli evoking EEG responses were selected from MCAESD according to the criteria as follows: First, declarative sentences are preferred instead of interrogative sentences to avoid prosodic differences of the same emotion speech. Second, the speech stimuli with higher accuracy in the listeners' annotation are preferred. As the duration of each emotional speech stimulus is too short to induce stable emotional EEG responses, every piece of speech stimuli was concatenated with several selected speech sentences for each of the four selected emotion types. Each piece lasts about 1 minute, and each emotion consists of 5 pieces. The speech stimuli are all with 16-bit quantization and a 44.1 kHz sampling rate.

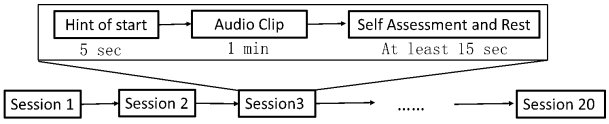

Figure 1: *Protocol of EEG data recording experiment.*

The EEG data recording experiment included a total of 12 healthy participants (7 female) aged from 19 to 27 years. All subjects are native speakers of Mandarin Chinese, right-handed with normal audiometric hearing, normal or corrected vision, and have no experience of brain injury or cognitive impairment. During the EEG experiment, the picture of a black background with a calibration cross in the centre is presented as a controlled visual stimulus. 12 subjects responded well to the stimuli materials after completing the experiment. Their emotion annotation results show that the stimulus materials can induce expected emotional states.

Our EEG experiment protocol is shown in Figure 1. After being informed of the purpose, protocol, and precautions of the experiment, subjects were asked to respond to audio stimuli in 20 subsequent sessions. In each session, there is a 5 s hint before each clip, then a 1-minute emotional prosody clip would be played, after the clip subjects were given at least 15 s for self-assessment by choosing among four emotion types.

The newly built EEG dataset was validated by high average accuracies of subjects' emotional annotations in the EEG experiment. The EEG emotional dataset and emotional prosody stimuli that induced certain emotion states in experiment contain rich emotional information that can be decoded through computational method.

EEG responses are recoded using 64-channel electrode cap with 10/20 layout and NeuroScan Acquire 4.3 software. In order to eliminate the influence of EOG artifacts (such as blinking and scanning) on scalp electrodes, two additional pairs of bipolar electrodes are used to record vertical electrooculogram (VEOG) and horizontal electrooculogram (HEOG).

Recorded EEG data is pre-processed using EEGLAB toolbox. The recorded EEG data file is cut into segments at same time length and synchronized with audio signals. The sample rate of EEG data was kept at 500Hz in raw and processed signals. Previous work shows that emotional information in EEG signals is mainly distributed in a frequency band from 1 to 50Hz, thus EEG signal passes through a 1-75Hz band-pass filter to remove artifacts. For EEG data of each trial, independent component analysis (ICA) is performed to remove the artifact components and eventually acquire pre-processed signal. Totally, there are about 240 minutes of EEG data, with the recording of four emotions evenly divided (60-min each).

2.2. Feature Extraction

2.2.1. EEG feature extraction

Many EEG emotional features have been developed in previous work concerning EEG emotion recognition, including time domain features, frequency domain features, and time frequency domain features [23]. The spectral power of EEG signal has been shown highly correlated with emotions, in which energy spectrum and differential entropy (DE) are commonly used EEG emotion features, and have been proven effective on SEED-IV and other EEG data sets [24, 25]. The Short Time Fourier Transform (STFT) with a 2-second time window and no overlapping Hanning window was used to extract the DE fea-

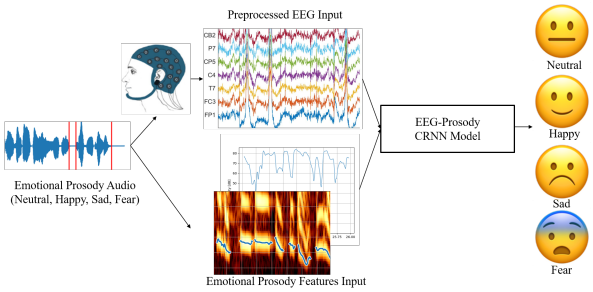

Figure 2: *The system diagram of the proposed method.*

ture in the five frequency bands: Delta (1-4Hz), Theta (4-8Hz), Alpha (8-14Hz), Beta (14-31Hz), Gamma (31-50Hz). Differential entropy features of each frequency band are calculated respectively.

Differential entropy was first applied to EEG fatigue detection, and Duan et al. applied it to EEG emotion recognition, achieving higher accuracy than power spectral density features [25]. It has been validated in recognizing emotions under a similar window length in previous experiments [24, 25, 26]. For a fixed-length EEG signal sequence, its differential entropy in the selected frequency band equals the logarithm of the power spectral density in that frequency band. Extracted EEG features are smoothed by calculating the mean with its nearest features on temporal dimension within a short time window.

Another method of extracting EEG features adopted in the following experiments is preserving time domain information by using EEG time sequences as input of deep learning models. This can be recognized as a simplified spatial-temporal feature of EEG data.

All two types of EEG input were examined in the experiment, in the following paragraphs spatial-temporal input refers to EEG time sequence, while spatial-spectral input refers to handcraft EEG features such as DE.

2.2.2. *Prosodic feature extraction*

In research on emotional prosody, many prosodic features related to emotion have been mentioned and examined, including fundamental frequency, loudness, etc. Also, many sets of features have been developed to solve the problem of speech emotion recognition in ordinary speech with semantic meanings. These feature sets include eGeMAPS (The extended Geneva Minimalistic Acoustic Parameter Set), PyAudio, IS13, and so on [27, 28, 29]. Based on existing feature sets, fundamental frequency, loudness, spectral flatness, spectral centroid, zero crossing rate, and MFCCs are selected as emotional prosodic features. These features also emerge frequently in previous studies on emotional prosody [15, 20].

2.3. Multi-modal Emotion Classification

The proposed multi-modal emotion recognition model is an end-to-end model. Both the EEG spatial-temporal signals directly extracted from pre-processed EEG data of every subject and the 6 down-sampled features of emotional prosody were provided as the input. Features of emotional prosody stimuli are extracted and down-sampled to match the sample rate of EEG features so that feature fusion between two modalities can be performed.

The model consists of a spatial-temporal EEG stream and a spectral features stream of emotional prosody, which are independent yet with similar network structures. As shown in Figure

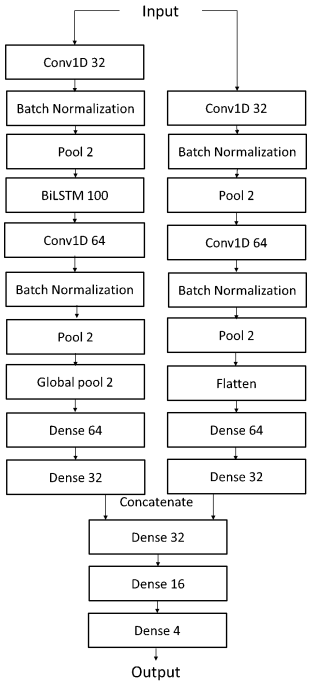

Figure 3: *The architecture of multi-modal EEG-Prosody CRNN model.*

3, the spatial-temporal EEG stream contains two 1-Dimension convolutional layers (kernel size: 3) followed by two dense layers. The spectral features stream contains two 1-D convolutional layers (kernel size: 3), a Bidirectional LSTM layer in the middle, and dense layers same as the EEG stream. In the end, the EEG spatial-temporal stream and emotional prosody features are fused by concatenating together along the temporal axis and classified into four emotion categories through an output layer with a softmax activation function.

3. Experiments and Results

3.1. Experiments

Experiments are based on the data set composed of emotion prosody stimuli and EEG data collected in experiments in Section 2.1.

The EEG-Prosody CRNN model and baseline deep learning models were trained and tested on 2 NVIDIA RTX 3090 GPUs. The Adam optimization is used to minimize the categorical cross entropy loss function and the learning rate is set to 0.0001. We use 5-fold cross validation (train test split ratio is 4:1) for 20 trials of each subject, which means EEG data and corresponding emotion prosody features of 16 trials (4 of each emotion category) are used for training, while the rest 4 trials from all 4 emotional categories formed the test data set.

3.2. Results and Discussions

We compare our model to baseline models, including SVM, CNN, and similar models, using differential entropy features (spatial-spectral feature) of EEG on our data set.

As Table 1, the CRNN model not only outperforms EEG single modality but also outperforms SVM, the most used method in previous literature on audio-evoked EEG emotion recognition. The model proposed in this study also outperforms machine learning methods in previous studies [9, 10, 30].

The CRNN model also reached a smaller cross-subject difference (standard deviation) compared with single modality models. Due to the dimension difference between prosody and EEG spectral features, features can't be concatenated as input of the SVM model.

Table 1: *Accuracy mean% and cross-subject standard error% of baseline methods and EEG-Prosody CRNN for emotion recognition models on the data set*

Model \ Input	Spatial-Temporal	Spatial-Spectral
EEG-SVM	71.89 (12.95)	57.34 (8.09)
EEG-CNN	79.79 (2.84)	64.24 (11.11)
EEG-Prosody SVM	49.51(3.16)	—
EEG-Prosody CRNN	**82.85 (3.34)**	65.16(4.93)

Ablation experiments have been performed to prove the necessity of combining EEG and prosody features. Prosody features achieved low accuracy on the single modality as features extracted were selected according to previous literature on emotional prosody. As Table 2, the combination of EEG and prosody features outperforms the models with EEG or prosody features only.

To identify EEG signals' critical frequency bands, the DE feature of EEG signals on five frequency bands has been examined separately and gamma band achieved the highest accuracy as input of the SVM model, beta band achieved the second best. The result is consistent with previous experiments on other datasets [24]. As the experiments have validated, EEG signals of beta and gamma bands have better emotion classification ability than delta, theta, and alpha bands.

Ablation experiments also have been performed to identify critical features in speech prosody. The results showed that the classification accuracy decreased dramatically once loudness or spectral flatness was removed from the prosodic features. This suggest that loudness and spectral flatness were critical features for the emotional prosody of speech in this experiment.

It is safe to conclude that features of emotional prosody are effective in enhancing the performance of EEG emotion recognition and show a statistically significant increase in accuracy according to our test result. Confusion graph in Figure 4 shows that the classification between sad and fear needs to be improved.

Compared with speech prosody stimuli, audio-visual stimuli like movie clips contain complex emotional information including visual context, text, speech prosody, and so on. These can hardly be quantified or structured, and the corresponding neuro responses are too complicated to decode. With controlled stimuli of emotional prosody in this paper, high-quality prosodic features have been used for emotion recognition with EEG signals and achieved good performance.

It is noteworthy that the classification result with EEG spatial-temporal data as input outperforms EEG spatial-spectral features input of differential entropy under same CNN model

Table 2: *Accuracy mean% and cross-subject standard error% of single modality and two modality fusion for emotion recognition models on the data set*

Model \ Input	Spatial-Temporal
EEG-CNN	79.79 (2.84)
Prosody CRNN	67.94
EEG-Prosody CRNN	**82.85 (3.34)**

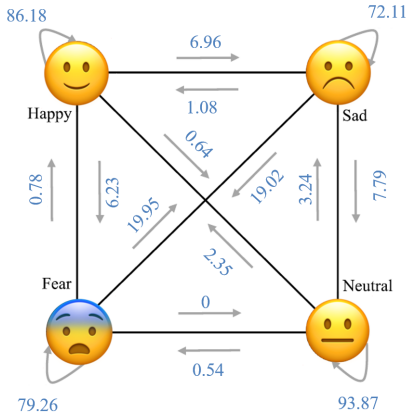

Figure 4: *The confusion graph of emotion classification result based on EEG-Prosody CRNN model.*

structure. Similarly, EEG spatial-temporal data achieves better performance than EEG spatial-spectral features in multi-modal emotion recognition with features of emotional prosody, as it is better to fuse high-level features of EEG and prosody along temporal dimension than concatenate spatial dimension of EEG features and temporal dimension of emotional prosody features.

Commonly used EEG spectral features in emotion recognition like differential entropy, although proved effective in many studies, contain signals from many aspects that are hard to be explained or disentangled [25]. EEG data aligned with prosody on temporal dimension achieved better results on deep learning models in the experiment, which indicate the better potential of EEG spatial-temporal features in EEG emotion recognition tasks, especially when emotional prosody features are included.

4. Conclusion

In this work, the emotional prosody of speech was used to evoke emotions in EEG experiments, and a novel framework based on multi-modal learning was proposed for emotion classification. Under this framework, a multi-modal CRNN network based on EEG spatial-temporal data and emotional prosody features was trained to enhance the performance of EEG emotion classification. The experiment results suggest that our method outperforms audio-evoked EEG emotion recognition models in previous studies and provides additional information to cross-subject emotion recognition tasks.

Theoretically, our framework has the potential to make use of more emotional information from more modalities like text to further enhance the performance of emotion classification, as this information commonly exists in both daily conversations and emotional stimuli databases. It also correlates with speech-evoked neural responses. So far, we only take emotional prosody features into consideration to simplify our experiment. It is also important to re-examine EEG spectral features, especially when the features of other modalities are used. In addition, how to select more effective emotional prosody features to achieve higher emotion classification accuracy in daily conversation scenarios should be furtherly studied.

5. Acknowledgements

This work was supported by the STI 2030-Major Projects (Grant No. 2021ZD0201500), a National Natural Science Foundation of China (Grant No. 12074012), and the High-performance Computing Platform of Peking University.

6. References

[1] L. F. Barrett, *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.

[2] J. William, "What is an emotion?" *Mind*, no. 34, pp. 188–205, 1884.

[3] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, 2022.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[5] H. Aviezer, R. R. Hassin, J. Ryan, C. Grady, J. Susskind, A. Anderson, M. Moscovitch, and S. Bentin, "Angry, disgusted, or afraid? studies on the malleability of emotion perception," *Psychological science*, vol. 19, no. 7, pp. 724–732, 2008.

[6] D. Nie, X. Wang, R. Duan, and B. Lv, "A survey on eeg based emotion recognition," *Chinese Journal of Biomedical Engineering*, vol. 31, no. 4, pp. 595–606, 2012.

[7] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[8] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2017.

[9] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Computers in Human Behavior*, vol. 65, pp. 267–275, 2016.

[10] I. Daly, D. Williams, A. Kirke, J. Weaver, A. Malik, F. Hwang, E. Miranda, and S. J. Nasuto, "Affective brain–computer music interfacing," *Journal of Neural Engineering*, vol. 13, no. 4, p. 046022, 2016.

[11] Z. He, Y. Zhong, and J. Pan, "Joint temporal convolutional networks and adversarial discriminative domain adaptation for eeg-based cross-subject emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3214–3218.

[12] W. Liu, W.-L. Zheng, Z. Li, S.-Y. Wu, L. Gan, and B.-L. Lu, "Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people," *Journal of Neural Engineering*, vol. 19, no. 2, p. 026012, 2022.

[13] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[14] W. F. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Journal of the International Association for Semiotic Studies*, 2006.

[15] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[16] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[17] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and emotion*, vol. 15, pp. 123–148, 1991.

[18] B. Gong, N. Li, Q. Li, X. Yan, J. Chen, L. Li, X. Wu, and C. Wu, "The mandarin chinese auditory emotions stimulus database: A validated set of chinese pseudo-sentences," *Behavior Research Methods*, pp. 1–19, 2022.

[19] P. Liu and M. D. Pell, "Recognizing vocal emotions in mandarin chinese: A validated database of chinese vocal emotional stimuli," *Behavior research methods*, vol. 44, pp. 1042–1051, 2012.

[20] S. Paulmann and A. K. Uskul, "Cross-cultural emotional prosody recognition: Evidence from chinese and british listeners," *Cognition & emotion*, vol. 28, no. 2, pp. 230–244, 2014.

[21] I. Darcy and N. M. Fontaine, "The hoosier vocal emotions corpus: A validated set of north american english pseudo-words for evaluating emotion processing," *Behavior Research Methods*, vol. 52, pp. 901–917, 2020.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[23] M. R. Islam, M. A. Moni, M. M. Islam, M. Rashed-Al-Mahfuz, M. S. Islam, M. K. Hasan, M. S. Hossain, M. Ahmad, S. Uddin, A. Azad *et al.*, "Emotion recognition from eeg signal focusing on deep learning and shallow learning techniques," *IEEE Access*, vol. 9, pp. 94 601–94 624, 2021.

[24] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[25] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[26] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for eeg-based vigilance estimation," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 6627–6630.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[28] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[30] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "Eeg based emotion recognition: A tutorial and review," *ACM Comput. Surv.*, vol. 55, no. 4, 2022.