From Intended to Subjective: A Conditional Tensor Fusion Network for Recognizing Self-Reported Emotion Using Physiology

Hao-Chun Yang and Chi-Chun Lee
* Department of Electrical Engineering, National Tsing Hua University, Taiwan
† MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
E-mail: hgy@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

***Abstract*—Previous studies have shown that an individual's subjective emotional evaluation involves cognitive processing that tends to be different from the directly-measured affective-physio signals, which creates a bias in the emotion labelings. Research has shown that this bodily-physiological signals are shown to be more related to the intended emotion elicitation type (-*Int*), yet correlated less to an individual's subjective emotion feelings (-*Sb*). Hence in this work, we suggest that this intended emotion elicitation status from the original stimuli (-*Int*) should be incorporated as an explicit regularization in achieving a more robust subjective emotion recognition system using physiology. To be more specific, we propose a novel conditional tensor fusion network in which the stimulation's emotion type -*Int* is firstly learned, then this learned intended annotation would then act as an explicit conditional regularization toward the final subjective emotion labeling. Our experiments indicate that this additional regulation helps to improve the overall emotion recognition on self-reported labels using physiology. We achieve an unweighted recall of 69.8% using ECG-EDA multimodal fusion, which is a relative improvement of 6.3% over the vanilla DNN method. Further feature analysis shows that several descriptors from ECG signals are indicative of the differences between these two emotion annotation schemes.**

I. Introduction

Automatic Emotion Recognition (AER) has seen a growing interest from various domains for its wide potential applications. A person's self-assessed emotion state (subjective, denoted as -*Sb*) are often used as the ground truth label and being recognized by computing a variety of expressive signals. For example, pitch contours, energy component, and vocal tract descriptors have all been demonstrated to carry high modeling power in developing speech-based AER [1]. It has been applied in applications of interactive agents [2] and psychological disease detection [3]. On the other hand, the Facial action coding system (FACS), i.e., computed based on facial muscle movement, has long been utilized as inputs to advance emotion recognition from face [4]. Recently, the advancement of miniaturized sensors and ubiquitous computing techniques have enabled the just-in-time bodily signal monitoring. These physiological measures is a more intrinsic bio-indicator, i.e., compared to face or speech modality, revealing activation of the autonomic and somatic nervous system (ANS and SNS), which has been shown to be closely related to emotion fluctuation [5]. This non-invasive and easy-to-use property also draws increasing interest in developing AER systems using physiological signals.

To develop a physiological AER system using machine learning approaches, many datasets have been collected under a similar setup [6]–[9]: during the experiment, a set of pre-selected affect-rich multimedia materials (usually video clips) are expected to elicit intended emotion states (-*Int*) as they being delivered to the receiving participants. Meanwhile, the participant's physiological signals are collected simultaneously as they watch these video stimuli. After each stimuli session, the participants are asked to describe how do they feel about themselves and report their own subjective emotion states (-*Sb*). This subjective self-reported emotion states have been used extensively as labels when learning to recognize emotion from physiology [10], [11].

However, obtaining high accuracy in recognizing one's own subjective emotion states (-*Sb*) remains challenging. It is largely due to the mechanism of self-reporting one's emotional states that can be quite complex as it involves layers of cognitive assessment beyond spontaneous emotional responses. An emotional experience is a psychophysiological process triggered by conscious and/or unconscious stimuli, and the formation of internal emotion is a result of a complex interaction between individual perceptual status and their bodily responses [12], [13]. In other words, one's self-reported emotional feelings could be biased by emotion-irrelevant physiological status and even other confounding factors that may be oriented toward *what should be felt* instead of *what has been felt*. Second, the occurrences of emotion may be completely neglected. In Ivonin's study [14], the idea of an unconscious mental process has been proposed. Specific types of implicit emotional elicitation may be ignored during self-evaluation procedures while showing a significant influence in affecting a subject's physiological responses. Yang's [15] study also suggests that stimulated physiological signals are more correlated with stimuli' affective status than self-disclosed emotion labels. All these researches suggest that subjective emotional state is rather complicated by its nature and it would usually lead to unsatisfying modeling especially using physiology solely.

Several methods have been proposed to enhance physiology AER using self-reported labels. For example, in Li's et al.

978-988-14768-8-3/20/$31.00 ©2020 APSIPA

Fig. 1. Data distribution of each labeling method. Due to poor data quality or missing labels, there are 629 samples left for each labeling scheme.

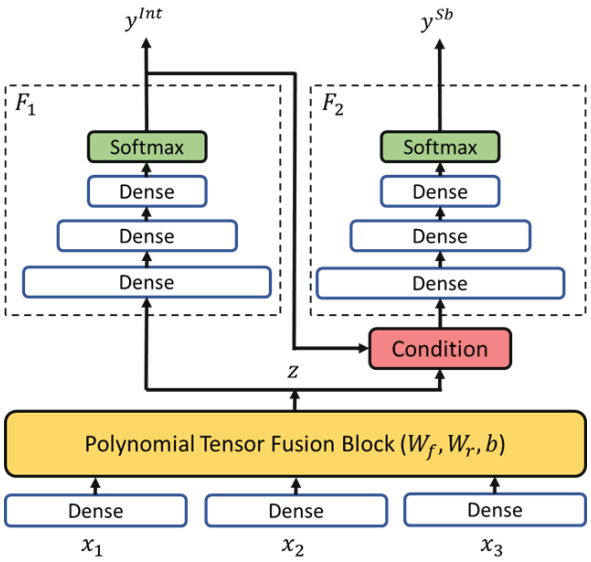

Fig. 2. Our proposed CPFN architecture. Note that each dense block is concatenated by a linear model, a Leaky-Relu as activation function, and a dropout layer.

study [16], the Hjorth parameter of mobility from EEG signals has been extracted as a key indicator describing emotion status. Shukla et al. also present a comprehensive study on calculating EDA features for AER [17]. Besides, personality attributes have started to be taken into consideration toward a more robust self-reported AER using physiology [18], [19]. However, while these methods improve recognition by taking advantage of fusing external information or sophisticated feature engineering, they usually neglect the fact that these measured bodily signals are in fact originally triggered by the affective stimuli (*-Int*). In other words, these signals would contain emotional responses from both emotion appraisal (*-Sb*) and the direct and intended emotion stimuli (*-Int*). Hence in this research, we propose that by explicitly constraint the *-Int* into the network learning, it would mimic the process of self emotion appraisal and improve the challenging subjective emotion recognition by learning a better representation. To be more specific, we propose a Conditional Polynomial Fusion Network (CPFN) which the automatically predicted stimuli' emotion label (*-Int*) is applied as explicitly conditional regulation for predicting the subjective emotion feelings (*-Sb*). We validate our proposed model on a large physiological emotion recognition dataset [9]. Our experiments show that this latent control prevents the potential self emotion cognitive bias and result in a more robust subjective AER system. We achieve an unweighted recall of 69.8% using ECG-EDA multimodal fusion, which is a relative improvement of 6.3% over the vanilla DNN method.

The rest of the paper is organized as follows: section II details the database and the computational methodology, section III reports the recognition results and illustrates potential feature discrepancy under different emotion annotation scheme. Finally, section IV concludes our findings and points out potential future directions.

II. Research Methodology

A. AMIGOS Dataset

In this study, we conduct our study on a large public physiological dataset AMIGOS [9]. In this dataset, 16 emotional videos with intended emotional stimuli (annotated with high/low arousal or valence, *-Int*) were delivered as multimedia elicitation to arouse the participants' affective responses. A total of 40 participants aged between 21 and 40 (mean age 28.3) were recruited to self-disclose their subjective feelings (*-Sb*) at the end of each video, while their physiological responses (ECG, EDA, and EEG) were recorded with biosensors throughout the time. We binarize this subjective emotion rating using the mean of each participant's rating as previously done in [15]. Fig.1 demonstrates the detailed label distributions.

B. Computational Framework

To evaluate the subjective emotion recognition conditional on the intended stimuli, we perform a binary emotion classification task using physiological features as our experimental setting. The detailed processes are described below.

1) Physiological Low-Level Descriptors (LLDs): We first apply a low-pass filter cut-off at 60Hz on ECG and EDA signals for noise reduction. Then, several standard Heart Rate Variabilities (HRVs) in time and frequency domain are calculated which has been studied as an important marker of autonomic nervous system (ANS) modulation [20]. As for EDA data, we fetch the tonic and phasic component which has previously shown as an important measure linking the physiological status toward affective responses [21]. Finally, EEG features like "Hjorth" and "ARMPB" parameters which have been studied and known as key emotion indicators [22] were also extracted to represent emotional brain activities. The exact features and dimensions are listed in Table I, and we use the open-source toolkit [23] for feature extraction. A subject-wise z-normalization is then applied to each feature dimension to mitigate the issue of individual differences.

TABLE I
AN OVERVIEW OF PHYSIOLOGICAL LOW-LEVEL DESCRIPTORS EXTRACTED FROM [23]. "F*" INDICATES 15 STATISTICAL FUNCTIONS[1].

Modality	Low-Level Descriptors
ECG(51)	RMSSD, meanNN, sdNN, cvNN, CVSD, medianNN, madNN, mcvNN, pNN50, pNN20, Triang, Shannon_h, ULF, VLF, LF, HF, VHF, Total_Power, LFn,HFn, LF/HF, LF/P, HF/P, DFA_1, DFA_2, Shannon, FD_Higushi
EDA(68)	F*SCR_Onsets, F*SCR_Peaks_Amplitudes, F*EDA_Phasic, F*EDA_Tonic
EEG(378)	Hjorth, Kurtosis, Skewness, 1DiffMean, 1diffMax, 2DiffMean, 2DiffMax, SlopeMean, SlopeVar, Wavelets, MaxPwelch, Entropy, AutoRegressiveParameters

2) Polynomial Tensor Pooling (PTP): Previous studies have shown that tensor-based multimodal fusion could improve the multimodal emotion recognition [24], [25]. Hence in this study, we utilize the novel polynomial tensor pooling block [26] for multimodal physiological signal integration. We first concatenate M modality features into a d^{in}-dimensional long vector $\boldsymbol{x}$:

$$\boldsymbol{x}^T = [1, x_1^T, x_2^T, ...x_M^T] \tag{1}$$

Here, a low-rank tensor network factor $\mathcal{W}_f$ with dimension $[d^{in}, d^{rank}, d^{out}]$ is multiplied to approximate the original polynomial multiplication with tensor factorization trick [27], preventing the dimension explosion due to the outer product operation:

$$\widetilde{\boldsymbol{x}} = \boldsymbol{x}^T \times \mathcal{W}_f \tag{2}$$

where the dimension of $\widetilde{\boldsymbol{x}}$ is $[d^{rank}, d^{out}]$. And then the element-wise multiplication within $\widetilde{\boldsymbol{x}}$ would equivalent toward the p-th order polynomial outer products:

$$\widetilde{\boldsymbol{x}}^{\boldsymbol{p}} = \underbrace{\widetilde{\boldsymbol{x}} * \widetilde{\boldsymbol{x}} * ...\widetilde{\boldsymbol{x}}}_{p} \tag{3}$$

Finally, the higher-order multimodal fusion embedding $\boldsymbol{z}$ is calculated by weight matrix $\mathcal{W}_r$ with shape $[1, d^{rank}]$ and bias b_r $[1, d^{out}]$:

$$\boldsymbol{z}^T = \mathcal{W}_r \times \widetilde{\boldsymbol{x}}^{\boldsymbol{p}} + b_r \tag{4}$$

This embedding would be regarded as the super encoding vector from multi-view physiological signals and would be further forward toward two separated dense layers $\mathcal{F}_1, \mathcal{F}_2$.

3) Multilinear Conditioning: To properly model subjective emotion states while preventing latent cognitive bias from bodily signals, we integrate the intended emotion label *-Int* as explicit control when predicting subjective labels *-Sb*. Inspired from the idea of conditional modelings [28], [29], the subjective label *-Sb* could be reparameterized conditioned on the prediction of *-Int*:

$$\begin{aligned} \widetilde{y}^{Int} &= \mathcal{F}_1(z) \\ \widetilde{y}^{Sb} &= \mathcal{F}_2(z \times y^{int}) \end{aligned} \tag{5}$$

[1]max, min, mean, median, std, skewness, kurtosis, min position, max position, 25_percentile, 75_percentile, 75_percentile-25_percentile, 1_percentile, 99_percentile, 99_percentile-1_percentile

Finally, the entire Conditional Polynomial Fusion Net-work (CPFN) would be optimized in an end-to-end manner with additional entropy $H(\widetilde{y}^{Int})$ as confidence level of $\widetilde{y}^{Int}$:

$$\begin{aligned} H(\widetilde{y}^{Int}) &= \sum_{c=1}^{C} \widetilde{y}^{Int}\ log\ \widetilde{y}^{Int} \\ \lambda(\widetilde{y}^{Int}) &= 1 + e^{-H(\widetilde{y}^{Int})} \\ \min_{\mathcal{W}_f, \mathcal{W}_r, b_r, \mathcal{F}_1, \mathcal{F}_2} & \mathcal{L}(\mathcal{F}_1(z), y^{Int}) + \lambda(\widetilde{y}^{Int}) \mathcal{L}(\mathcal{F}_2(z, \widetilde{y}^{Int}), y^{Sb}) \end{aligned} \tag{6}$$

where $\mathcal{L}$ states for the standard cross-entropy loss.

III. EXPERIMENT SETUP AND RESULTS

A. Experiment Setup

The exact architecture of our Conditional Polynomial Fusion Network includes several blocks of networks. Several hyperparameters were grid-searched: learning rate among $[0.005, 0.001]$, polynomial order between $[1, 2]$, and tensor rank dimension d^{rank} among $[1, 4, 8]$. Batch size is fixed as 16 and dropout rate at 0.2, the max epoch is 150 with early-stopping, and the optimizer is Adam. To prevent overfitting, we carry out all experiments under a subject independent 10-fold cross-validation setup. The final evaluation metric used is the unweighted average recall (UAR).

1) Comparison Models: We first conduct our experiments utilizing linear SVM and vanilla DNN without consideration of *-Int*. Then we compare results among the following models to validate our idea of multi-annotation conditional control:

- MTL-DNN: Multitask Learning Dnn. Multitask learning has been studied as learning multiple tasks simultaneously expecting knowledge transfer among a variety of tasks [30]. Here we applied the simple two-stream architecture in predicting both *-Int* and *-Sb*. We consider it as a naive method to investigate the potential effect of multi-annotation joint learning. Since we specifically focus on the latent conditional property from *-Int* toward *-Sb*, other MTL architectures would not be discussed in this work.
- PFN: Polynomial Fusion Network. Modalities are fused using the polynomial tensor factorization technique depicted in II-B2. Noted that we apply the same technique under a single modality scenario which we regard it as intra-modality fusion among LLD dimensions.
- CPFN: Our proposed Conditional Polynomial Fusion Network. An enhanced subjective emotion recognition system integrating conditional control of learned intended stimuli emotion level *-Int*.

B. Subjective Emotion Recognition Results

Table II summarizes our emotion recognition results. Our proposed CPFN reaches the highest subjective emotion recognition across almost all ECG and EDA related modalities. Several notable observations can be summarized. Firstly, through the SVM and DNN experiments, we could see that Valence is more better modeled than Arousal in either ECG, EDA, or EEG modalities. Comparing to Valence, there is around 5%

TABLE II
A SUMMARY OF SUBJECTIVE EMOTION RECOGNITION (-SB) RESULTS. THE BOLD ONE WOULD BE A SINGLE MODALITY'S MAXIMUM WHILE * IS THE GLOBAL MAXIMUM. THE CHANCE UAR IS 0.5.

	Arousal						Valence					
	ECG	EDA	EEG	ECG-EDA	ECG-EEG	EDA-EEG	ECG	EDA	EEG	ECG-EDA	ECG-EEG	EDA-EEG
SVM	0.542	0.559	0.543	0.476	0.478	0.495	0.558	0.597	**0.639**	0.537	0.586	0.597
DNN	0.53	0.576	0.562	0.565	0.562	0.578	0.588	0.629	0.597	0.635	0.608	0.629
MTL-DNN	0.556	0.592	0.584	0.593	0.579	0.585	0.607	0.643	0.61	0.66	0.621	0.654
PFN	0.548	0.587	0.582	**0.603***	**0.585**	**0.596**	0.622	0.648	0.624	0.658	0.63	0.653
MTL-PFN	0.562	0.586	0.584	0.592	0.58	0.588	0.624	0.64	0.623	0.676	**0.625**	**0.658**
CPFN	**0.565**	**0.6**	**0.591**	0.601	0.572	**0.596**	**0.631**	**0.655**	0.62	**0.698***	0.623	0.645

TABLE III
FEATURES INDICATIVE TO DIFFERENCE BETWEEN TWO EMOTION ANNOTATION SCHEMES *-Int* AND *-Sb*. CCSQ: CARDIAC CYCLES SIGNAL QUALITY, LF/P: POWER RATIO BETWEEN LOW FREQUENCY AND TOTAL POWER, VLF: POWER OF VERY LOW FREQUENCY, NN_MEAN: MEAN OF RR-INTERVALS

Modality	Arousal	Valence
ECG	CCSQ_low_quar (1.972, p=0.049) CCSQ_std (-2.045, p=0.041) LF/P (-2.393, p=0.017) VLF (2.124, p=0.0343)	CCSQ_min (-2.092, p=0.04) NN_mean (-2.082, p=0.04)
EDA	-	-

drop while using physiology for modeling Arousal. Further experiments on MTL-DNN method confirms our hypothesis that the joint modeling using both *-Sb* and *-Int* emotion annotation scheme could improve the recognition of subjective emotion labeling. We would put more emphasis on Valence recognition in the following section.

We observe that the integration of the polynomial fusion block (PFN) into the model could boost the recognition of ECG and EDA, which is an improvement of around 3% in both single and dual modalities scenario. On the other hand, there are no observable changes in EEG signals. Then similar to the MTL-DNN method, the MTL-PFN model could also help the prediction of -Sb, especially in the ECG-EDA fusion scenario. Lastly, multilinear conditional control enhances the model capability. During the conditional modeling from pseudo--*Int* toward *-Sb*, the outer product operation act as a reparameterization trick toward the original learned physiological hidden features. We believe this step would mimic the process of self emotion appraisal by introducing additional prior knowledge that *"WHICH"* intended emotion elicitation is being delivered, and force the *-Sb* branch model to learn a constrained (better targeted) representation under this controlling mechanism. Comparing to the naive DNN method, our proposed architecture reaches a relative improvement of 4.3%, 2.6%, and 6.3% respectively on ECG, EDA, and ECG-EDA modalities respectively.

C. Feature Discrepancy Analysis

In this section, we conduct statistical tests to examine the discrepancy of emotionally-informative physiological descriptors under two different annotation *-Int* and *-Sb*. We specifically run the tests on ECG-EDA modality due to its modeling power in achieving a higher emotion recognition rates. We first split the entire dataset into two groups according to whether the sample's *-Int* and *-Sb* are the same or different. By applying the two-tailed Student's t-test, we summarize statistically sensitive features ($p - value < 0.05$) toward two different annotation schemes in table III. From this table, first, we could quickly notice that there are no EDA features selected. This implies that comparing with EDA data, people may somehow more likely to reveal through ECG variations when they are experiencing a certain level of emotion appraisal bias (i.e., *-Int* not equals to *-Sb*). This may also explain the reason that when comparing MTL models to our proposed CPFN, the additional information from *-Int* usually helps improve the recognition more on ECG related modalities. Finally, it is surprising to see that some of the ECG quality features are quite sensitive toward the differences between the emotion labels. We hypothesize that when people are under a certain level of emotional cognitive difficulties, this "confusion" status could potentially be reflected in this special ECG feature. However, it would require further specifically designed experiments to investigate the potential underlying mechanism.

IV. CONCLUSION

Previous studies have shown that the bias between physiological features and the self-assessed emotion annotations could degrade the automatic emotion recognition system of self-reported emotion feeling. This bias could result from either intended emotion elicitation or unconscious emotion process, which results in a mismatch between the recorded physiological responses and subjective emotional feelings. Hence in this work, we present a novel idea that the stimuli' emotion level (*-Int*) should also be considered to mitigate this mismatch. We propose a Conditional Polynomial Fusion Network which incorporates additional emotion annotation into joint modeling. In this model, the original multimedia stimuli's intended emotion label *-Int* was firstly learned from the recorded physiology. Then we explicitly apply this automatically-learned stimulation's emotion as a conditional control when predicting the real subjective emotional feelings *-Sb*. We consider it as a mechanism that we automatically learn a prior knowledge describing the context when the emotion

was triggered, then this prior information would be applied to guide the prediction of the subjective emotion labels. Our experiments show that by applying this conditional control, it improves subjective emotion recognition results. To our best knowledge, this is one of the first works that incorporate multi-perspective emotion annotation schemes into physiological emotion modeling. We can foresee several future directions. Immediate work would be to design a new experimental protocol to investigate the underlying mechanism of emotion-appraisal bias. Besides physiological signals, other modalities that may be used as an emotion indicator should re-investigate the potential annotation discrepancy as well. Additionally, other emotion annotation schemes such as *Perceived-Emotion* should also be studied. By better understand the formation of emotion from multiple points of view would help in advancing a variety of human-centered multimedia applications [31], [32].

REFERENCES

[1] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[2] Madiha Anjum, "Emotion recognition from speech for an interactive robot agent," in *2019 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2019, pp. 363–368.

[3] Sahar Harati, Andrea Crowell, Helen Mayberg, and Shamim Nemati, "Depression severity classification from speech emotion," in *EMBC*. IEEE, 2018, pp. 5763–5766.

[4] Byoung Chul Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, pp. 401, 2018.

[5] Walter B Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[6] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014, pp. 647–655.

[8] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

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

[10] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, pp. 2074, 2018.

[11] Patricia J Bota, Chen Wang, Ana LN Fred, and Hugo Plácido Da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019.

[12] Stanley Schachter and Jerome Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, no. 5, pp. 379, 1962.

[13] George Mandler, "Emotion," *Handbook of psychology*, pp. 157–175, 2003.

[14] Leonid Ivonin, Huang-Ming Chang, Marta Diaz, Andreu Catala, Wei Chen, and Matthias Rauterberg, "Traces of unconscious mental processes in introspective reports and physiological responses," *PloS one*, vol. 10, no. 4, pp. e0124519, 2015.

[15] Hao-Chun Yang and Chi-Chun Lee, "Annotation matters: A comprehensive study on recognizing intended, self-reported, and observed emotion labels using physiology," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[16] Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, and Bin Hu, "Exploring eeg features in cross-subject emotion recognition," *Frontiers in neuroscience*, vol. 12, pp. 162, 2018.

[17] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, GC Nandi, and Domenec Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, 2019.

[18] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–18, 2019.

[19] Hao-Chun Yang and Chi-Chun Lee, "An attribute-invariant variational learning for emotion recognition using physiology," in *ICASSP*. IEEE, 2019, pp. 1184–1188.

[20] Richard D Lane, Kateri McRae, Eric M Reiman, Kewei Chen, Geoffrey L Ahern, and Julian F Thayer, "Neural correlates of heart rate variability during emotion," *Neuroimage*, vol. 44, no. 1, pp. 213–222, 2009.

[21] Guillaume Berna, Laurent Ott, and Jean-Louis Nandrino, "Effects of emotion regulation difficulties on the tonic and phasic cardiac autonomic response," *PloS one*, vol. 9, no. 7, pp. e102971, 2014.

[22] Raja Majid Mehmood and Hyo Jong Lee, "Eeg based emotion recognition from human brain using hjorth parameters and svm," *International Journal of Bio-Science and Bio-Technology*, vol. 7, no. 3, pp. 23–32, 2015.

[23] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and Annabel S H Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," 2020.

[24] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel, Eds. 2017, pp. 1103–1114, Association for Computational Linguistics.

[25] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 2018, pp. 2247–2256, Association for Computational Linguistics.

[26] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Advances in Neural Information Processing Systems*, 2019, pp. 12113–12122.

[27] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, and Danilo P Mandic, "Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions," *Foundations and Trends® in Machine Learning*, vol. 9, no. 4-5, pp. 249–429, 2016.

[28] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *ICML*, 2009, pp. 961–968.

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, 2018, pp. 1640–1650.

[30] Yu Zhang and Qiang Yang, "A survey on multi-task learning," *CoRR*, vol. abs/1707.08114, 2017.

[31] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[32] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.