Annotation Matters: A Comprehensive Study on Recognizing Intended, Self-reported, and Observed Emotion Labels using Physiology

Hao-Chun Yang
Electrical Engineering
National Tsing Hua University
Taiwan
hgy@gapp.nthu.edu.tw

Chi-Chun Lee
Electrical Engineering
National Tsing Hua University
Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Studies have shown that the formation of emotion as self-awareness and cognitive appraisal process is complicated and can lead to idiosyncratic differences. Subject's self emotion evaluation process could be biased due to factors of environment, personal experience, and one's own cognitive ability, and the true affective state may be neglected (un-noticeable) due to an unconscious mental process. In this work, we present a comprehensive study to investigate the emotion recognition accuracy obtained using physiology with respect to different annotation schemes, i.e., *intended*, *self-reported*, and *observed* emotion labels. We found that when performing recognition across these three different labeling schemes using the same physiological parameters, the accuracy of the self-reported emotion labels results in about 10.3% and 3.1% drop when compared to two other annotation schemes. It indicates that self-assessed emotion labels may be noisier and induces a larger mismatch with respect to the affect-stimulated physiological responses. Further analysis shows that the electrodermal activity signal has the highest recognition rate with respect to the intended emotion of the stimuli. Finally, our error analysis reveals that there may exist a bias in the self-annotated label that is conditioned on the intended stimuli's valence polarity.**

***Index Terms*—emotion recognition, annotation, physiology, affective computing, mental process**

I. Introduction

Human's affective response is a psychophysiological process triggered by conscious and/or unconscious stimuli and is often manifested through observable behavior channels [1]. The field of affective computing has advanced tremendously in developing algorithms for detecting, modeling, and synthesizing emotion mostly through modeling of observable behavior signals collected using audio-video devices, such as facial expressions, speech, and linguistic contents. Recently, the advancement of miniaturized sensors has enabled continuous monitoring of various human internal physiological signals, such as electroencephalography (EEG), electrocardiography (ECG), and electrodermal activity (EDA). This has drawn increasing interest for researchers to investigate the internal affect-related physiological responses and further device computational strategy in automated modeling of emotion using physiology [2].

The most common scenarios in these research works are based on using emotionally-rich audio-visual data as stimuli in order to trigger a subject's internal responses, which is then captured in the physiological measurements from these devices. For example, Koelstra et al. collect a physiological response database by showing highlights of music videos as the emotion-triggering sources [3]. Other related research also demonstrates that short video clips from movies can be used as stimuli to understand the physiological variations of different induced emotion states [4], [5]. While growing research has suggested that the variations of physiology are indicative of a subject's internal emotion states through this particular type of stimuli-based experimental protocol, most if not all of these research studies leverage the subject's *self-assessed* emotion annotations as the *ground-truth* emotion states to perform their corresponding analyses.

Self-assessed emotion annotation comes intuitively as a first priority in order to investigate the relationship between physiology and emotion processes. However, there could be issues in interpreting emotion states only through the viewpoint of self-reported emotion labels. The first issue comes from the fact that previous emotional appraisal theory has demonstrated that the formation of internal emotion is through a series of complicated cognitive appraisal processes, and this particular mechanism is complex and potentially induces individual biases. In Schachter and Singer's experiment [6], the awareness of emotion is shown to be masked by the arousal of one's physiological status, and later on, Mandler extends the theory showing that this arousal results from the occurrences of individual perceptual and/or cognitive discrepancy [7]. Second, the occurrences of emotion could be neglected. In Ivonin et al. study [8], the idea of an unconscious mental process has been proposed. Through their studies, they show that specific types of implicit elicitation (archetypes that represent prototypical experiences associated with objects, people, and situations in the study) may be ignored during self-evaluation procedure, this unconscious mental process, however, shows significant influence in affecting a subject's physiological responses. This study highlights the limitation of our cognitive-based self-

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

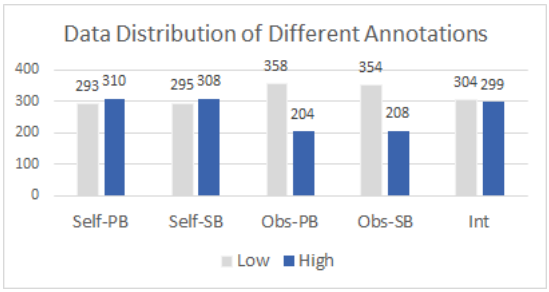


Fig. 1: Data distribution of each labeling method. *-SB* and *-PB* refer to binarized according to subject mean or dataset's mean respectively. Note that due to poor physiological data quality or missing labels (videos of participants P8, P28, and P33 are not annotated in the original dataset), there are totally around 560˜600 data for each labeling scheme.

assessment strategy in performing annotation and points to the fact that by merely exploring the relationship between physiological responses to self-assessed emotion annotation can lead to a restricted interpretation.

In this work, we conduct a comprehensive study to examine the variations of physiological responses as a function of different emotion annotation schemes. More specifically, during an emotion elicitation episode, we can imagine that a subject's true internal affective state is hidden yet could be *labeled* through observation of an subject's explicit behaviors such as facial expression (*observed*), self-assessed evaluation by the subject (*self-reported*), or linking directly to the intended emotion stimuli itself (*intended*). Each of these realizations of the annotation scheme gives us a different view of the true internal states. Hence, with this particular abstraction, we can study the variation in the recognition accuracy using physiological parameters as a function of different annotation labels. Our recognition results show that the differences in the accuracy obtained between these annotations are up to around 10% and 5% for arousal and valence attributes respectively. Besides, the EDA signal consistently and uniquely achieves the best recognition accuracy for the *intended* emotion labels out of the three types of physiology signals (EEG, ECF, EDA). To further understand the potential latent interaction between the physiological modalities and these annotations, we conduct a model-based feature importance analysis. Finally, an error analysis is applied to understand the potential reason behind a lower accuracy obtained when learning to recognize the self-reported labels using physiology measurements.

The rest of the paper is organized as follows: section II illustrates the database and the computational methodology, section III reports the recognition results toward distinct annotations and section IV details the feature and error analysis.

II. RESEARCH METHODOLOGY

A. AMIGOS Dataset

This study uses the AMIGOS Dataset [9]. The dataset is composed of 40 subjects with each watching 16 short emotional videos (duration<250s) and 4 long videos (duration>14min) designed to evoke the participants' affective response. These video stimuli are carefully chosen from two other databases aimed at studying physiological reactions to emotional content, and the intended stimuli type are labeled as (*Int*). EEG, ECG and EDA signals are recorded simultaneously during each stimuli episode. For each trial, the subject's self emotion annotations (*Self-*) are performed right after watching each video stimuli. As for observed external annotations (*Obs-*), the video recordings of the participant's facial reactions are sliced into 20 seconds clips and randomly delivered to three external annotators to indicate the arousal and valence scores offline. Each annotator labels all clips, and the mean of these ratings are regarded as the final score of each stimulus for every subject. Since both self and observed annotations are continuous values, here we perform two mapping strategies to binarize scores in this work. The first one is termed as Self-Based (*-SB*) which refers to binarizing the original scores using a subject-dependent mean value, while the second one would be Public-Based (*-PB*) which simply binarizes scores using the database mean. Hence, under these scenarios, there are a total of five different annotations schemes to a single episode of a subject's physiological measurement. We only use the physiological data from short video stimulation to provide fair comparisons between labels. Fig 1 gives the detailed distribution of each annotation method.

TABLE I: A summary of physiological low-level descriptors. "F*" indicates 15 statistical functions.[1] EEG functions are calculated for each channel then concatenated as a single feature vector. CVSD: The coefficient of variation of successive differences, the RMSSD divided by meanNN. SCR: skin conductance response.

Modality	Low-Level Descriptors
EEG(378)	Hjorth, Kurtosis, Skewness, First_diff_mean, First_diff_max, Sec_diff_mean, Sec_diff_max, Slope_mean, Slope_var, Wavelets, MaxPwelch, Entropy, ARMPB
ECG(51)	number_of_artifacts, 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, Sample_Entropy, Correlation_Dimension, Entropy_Multiscale_AUC, Entropy_SVD, Entropy_Spectral_VLF, Entropy_Spectral_LF, Entropy_Spectral_HF, Fisher_Info, FD_Petrosian, FD_Higushi, Average_Signal_Quality, F* Cardiac_Cycles_Signal_Quality
EDA(68)	F*SCR_Onsets, F*SCR_Peaks_Amplitudes, F*EDA_Phasic, F*EDA_Tonic_Component

B. Computational Framework

To evaluate the discriminability of physiological data to the three labeling methods, we perform a binary emotion classification task using physiological features as our experimental setting. The detailed processes are described below.

1) Low-Level Physiological Descriptors (LLDs): First, we apply filters to the physiological data for noise suppression. For EEG, a bandpass filter from 4-45Hz is applied, while for

[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

TABLE II: A summary of prediction UARs. We also report the multi-modal fusion results based on feature concatenation method. Note that the bold numbers are the highest among different labeling methods within the same modality, and number with * are marked as the global maximum for each emotion label. The baseline (random guess) UAR would be 0.5.

	Arousal					**Valence**				
	Self-PB	Self-SB	Obs-PB	Obs-SB	Int	Self-PB	Self-SB	Obs-PB	Obs-SB	Int
ECG	0.514	0.533	0.568	**0.589**	0.552	0.542	**0.557**	0.532	0.547	0.517
EDA	0.566	0.555	0.567	0.560	**0.628**	0.544	0.535	0.550	0.540	**0.567**
EEG	0.577	0.578	0.636	**0.681**	0.618	0.586	0.565	**0.620**	**0.620**	0.574
ECG_EDA	0.562	0.585	0.556	0.598	**0.659**	0.566	**0.602**	0.548	0.557	0.600
ECG_EEG	0.576	0.573	0.662	**0.685**	0.632	0.588	0.600	0.639	**0.650**	0.584
EDA_EEG	0.568	0.569	0.664	**0.685**	0.652	0.623	**0.632**	0.615	0.623	0.613
ECG_EDA_EEG	0.568	0.587	0.650	**0.690***	0.650	0.637	0.613	0.624	**0.644***	0.633

ECG and EDA are both filtered by a low-pass filter with 60Hz cut-off frequency. Then, several standard LLDs are extracted using NeuroKit [10], i.e., an open-source feature extractor for neurophysiological signals. Features like heart rate variability (HRVs) from ECG and peak related features from EDA are calculated using this tool. As for EEG, we extract features of Hjorth parameters, wavelet analysis, and entropy related features. The detailed list of features is shown in Table I. Finally, we apply a z-normalization on each feature dimension for each subject to mitigate the issue of individual difference.

2) Recognition: We perform a binary classification task using features extracted in the previous section as our main experimental setting. The classifier we use is a linear support vector machine (SVM) [11], with penalty parameter c is set to 1 and other parameters are left as default. Our experiments are based on strict leave one person out (LOO) cross-validation scheme. The evaluation metric reported is the unweighted average recall (UAR).

III. RESULTS

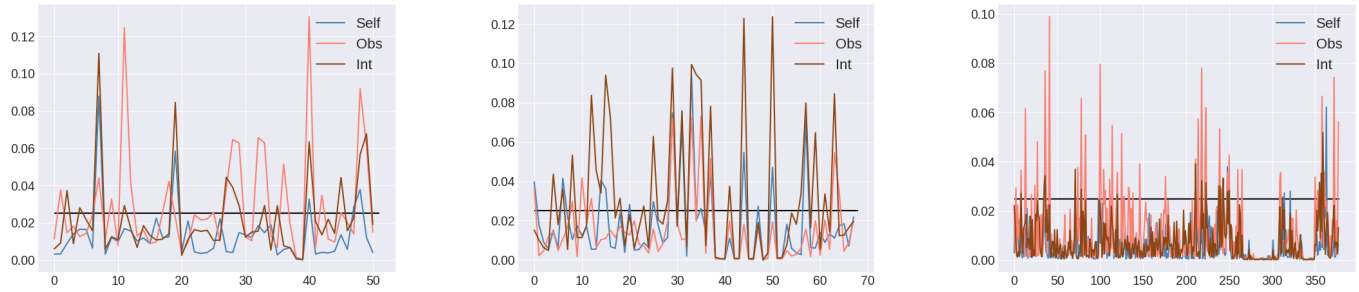
Table II summarizes our emotion recognition results. Several findings could be observed. First, we find that subject-dependent label binarization technique (*-SB*) results in generally a better recognition accuracy. This suggests that a global normalization may neglect the individual baseline creating unwanted bias when learning to recognize these emotion labels using physiological signals. Second, we notice that different labeling methods have a profound impact on recognition accuracy. Specifically, after concatenating all three physiology modalities as our fusion method, the recognition result of arousal ranges from 0.568 to 0.69 and from 0.613 to 0.644 for valence across different labeling strategies. This indicates that although we have identical physiological parameters, merely different annotation methods could lead to inconsistent results. More precisely, from the original experiment settings, we know that the physiological data are recorded from the participants themselves, so intuitively it should be mostly related to their self-annotations. However, from our results, we notice that the predictability of self-labels (*Self-*) using physiology is the worst. In contrast, these physiological data is more discriminative when learning to recognize the external observer's ratings (*Obs-*). This indicates that the observer's assessment of a participant's emotional state can be more consistently modeled by using the participant's physiological data rather than their own self emotion assessment. In the next section, we will conduct further analysis to understand the potential relationship between these internal signals with the observed, intended and self-reported emotion labels.

IV. DISCUSSION

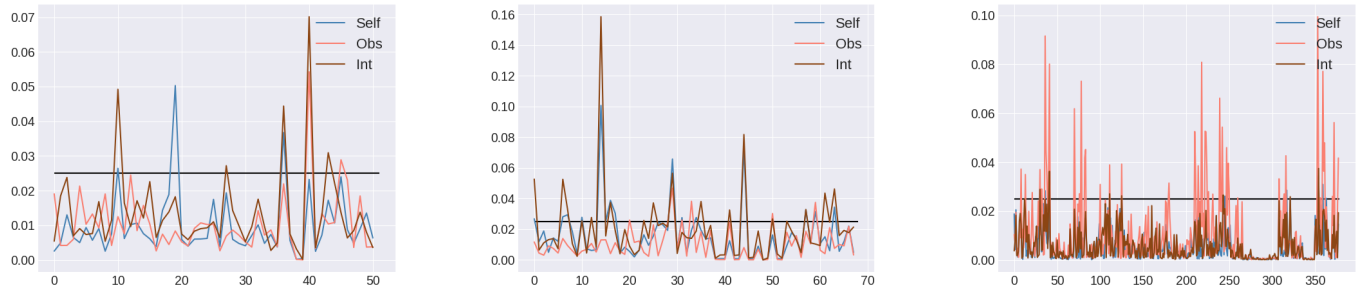
In this section, we present further analyses of the recognition results. First, we perform model-based feature importance analysis to identify key physiological parameters for each type of labels. Then, an error analysis is conducted to examine the relationship between the *intended* versus the *self-reported* label's recognition results. For simplicity, we denote the self-annotations as *Self* and external observations as *Obs*, which refers to the *Self-SB* and *Obs-SB* mentioned in the previous section.

1) Model-based Feature Importance Analysis: We take the approach of interpreting feature importance by directly evaluating the classification model itself. To identify the most discriminative dimensions of physiological parameters for each emotion recognition task, we compute the SHAP (SHapley Additive exPlanations) score [12] for feature analysis in this section. SHAP assigns a feature importance value of a test sample for a given prediction model. SHAP has become the most advanced explainable artificial intelligence approach in understanding how a complicated machine learning model makes a prediction, and under certain criteria, SHAP is proven to identify and rank the feature attributes for a given model to a unique solution. This method has been shown to be more effective and precise compared to previous methods of explaining machine learning prediction. The original method is developed for the decision tree, and we modify it to be used for support vector machine in this work.

First, we perform LOO recognition and utilize the trained model to obtain SHAP scores of each feature dimension for every testing samples. Since the principle of the SHAP analysis is built on the trained models, whether the model is well-trained would affect the validity of our SHAP analysis. We only aggregate SHAP score of the samples that are correctly predicted by our model. And then, in order to estimate the importance of each feature dimension, we average the absolute



(a) Arousal-ECG (b) Arousal-EDA (c) Arousal-EEG



(d) Valence-ECG (e) Valence-EDA (f) Valence-EEG

Fig. 2: The distribution of SHAP scores obtained for each physiological modalities. The vertical axis refers to the absolute SHAP score while the horizontal axis maps to the feature dimensions to each physiology. The horizontal black line indicates the cut-off threshold at 0.025.

SHAP value of each feature dimension. Finally, in order to compare across different prediction models, we set a cut-off threshold (0.025 in our analysis, see fig2) to identify those most discriminative features with respect to each classification task. This feature set is listed in table III.

According to the table, we can see that in ECG modality, there is hardly any features that are consistently important between different labeling schemes and mostly due to the fact that the predictability of *Self* label is low. In predicting arousal, we notice that many of the Heart Rate Variability (HRVs) features are jointly selected in the cases for the intended original video stimuli *Int* and the observers' judgment *Obs*. By referencing the recognition results in table II, we hypothesize that this may indicate that the participants indeed are successfully stimulated by the intended source video, then react internally as evident in the HRV measures, and further manifest in their observable behavior manifestation (for example a smile on their face) as external annotation demonstrates. This joint identification of physiological parameters across labeling schemes is, however, not evident for self-reported emotion, which points to a potential discrepancy between one's self-aware emotion labels and the subject's physiological responses. This effect has also recently been studied in ECG as traces of unconscious mental process [8]. This phenomenon is defined by Bargh [13] as "in terms of a lack of awareness of the influences or effects of a triggering stimulus and not of the triggering stimulus itself". Similarly, they applied an emotion recognition task to provide evidence in their hypothesis and found that the types of stimuli would affect the effect of the physiological variation that goes beyond self-assessment. In this work, we observe that even when using the same set of physiological response, it would still have a varying degree of predictability to different annotations; this may result from some degree of emotional unconscious reaction.

Secondly, we focus on the EDA's modality. We find that for both arousal or valence attribute, "Peaks_Amp" feature computed from skin conductance response (SCR) is consistently selected as a key factor that are correlated to the intended *Int* labels. SCR is defined by the rapid fluctuations in eccrine sweat gland activity, which results from the liberation of acetylcholine by the sympathetic nervous system [14]. Hence, this measure has distinct phenotype from other automatic nervous system signs such as heart rate, since SCR is under the strict control of the sympathetic branch of the nervous system. This discrepancy in the production mechanism may help explain why EDA related features achieve the best predictive power on *Int* label in contrast to other physiological modalities.

Lastly, we examine the EEG modality. EEG reaches the highest single modality UAR for both arousal and valence, and from the fusion outcomes, we could also observe that this modality dominates the emotion recognition task in our experiment. Specifically, EEG achieves high predictability on *Obs*, however, decreases an almost 10% and 5% on *Self* labels of arousal and valence. We notice that many "Hjorth" related dimensions, which are commonly calculated in emo-

TABLE III: It shows feature dimensions that show significantly statistical difference across participants for different affective labels. Columns names with multiple labels show the list of feature dimensions that are selected as the intersection between the annotation labels. CCSQ: Cardiac cycles signal quality. For EDA, "Peaks", "Onsets" indicates the skin conductance response property and "Tonic", "Phasic" are the frequency component of the signal. ARMPB: Autoregressive model parameters using Burg method. Hjorth: Hjorth Parameters.

	Self / Obs / Int	Self / Obs	Self / Int	Obs / Int	Self Only	Obs Only	Int Only
Arousal							
ECG	CCSQ_max_pos, VHF, pNN20		DFA_2	CCSQ_min_pos, HF, HF/P, HFn, LF, nmeanNN, pNN50	n_Artifacts	CCSQ_1_per, CCSQ_median, CCSQ_quartile_range, Correlation_Dimension, DFA_1, FD_Petrosian, LF/P, Shannon_h, cvNN	CCSQ_99_per, CCSQ_kurtosis, Shannon
EDA	Tonic_max_pos, Tonic_min_pos, Tonic_skew, Tonic_up_quar	Phasic_1_per	Phasic_VLF, Phasic_min_pos, Phasic_quartile_range, Tonic_VLF, Tonic_median, Onsets_max_pos, Onsets_skew, Peaks_Amp_low_quar	Phasic_low_quar, Phasic_median, Peaks_Amp_min_pos	Tonic_99_per, Onsets_min	Phasic_max_pos, Peaks_Amp_quartile_range	Phasic_LF, Phasic_min, Phasic_skew, Phasic_up_quar, Tonic_LF, Tonic_max, Tonic_quartile_range, Onsets_kurtosis, Peaks_Amp_kurtosis, Peaks_Amp_max_pos, Peaks_Amp_median
EEG	slope_mean*1, wavelet_cD_mean*1	wavelet_cD_mean*1		ARMPB0*1, ARMPB2*1, first_diff_mean*2, Hjorth_mobility*1, sec_diff_mean*3, slope_mean*2	wavelet_cA_mean*2, wavelet_cD_mean*1	ARMPB0*2, ARMPB1*3, ARMPB2*2, f,irst_diff_mean*3, sec_diff_mean*2, Hjorth_complexity*5, Hjorth_mobility*2, maxPwelch0*4, Pwelch1*1, Pwelch3*1, slope_mean*2, slope_var*2, wavelet_cA_mean*2, wavelet_cD_mean*3, wavelet_cD_std*2	ARMPB2*1, slope_mean*2, wavelet_cA_mean*1, wavelet_cD_mean*2
Valence							
ECG			CCSQ_min, Shannon_h	VHF	DFA_2	meanNN	HF, madNN
EDA	Tonic_max_pos	Tonic_median, Peaks_Amp_max_pos	Phasic_1_per, ets_max_pos Peaks_Amp_low_quar, Peaks_Amp_min_pos	Onsets_skew	Phasic_std, Tonic_quartile_range	Tonic_99_per Tonic_min_pos	Phasic_median, Tonic_LF, Tonic_VLF, Tonic_skew, Onsets_kurtosis, Peaks_Amp_1_per, Peaks_Amp_median
EEG	ARMPB2*1, wavelet_cD_mean*1	ARMPB2*1, wavelet_cD_mean*1	wavelet_cA_mean*1	ARMPB2*1, Hjorth_complexity*1, Hjorth_mobility*1	slope_mean*1, wavelet_cA_mean*1, wavelet_cD_mean*1	ARMPB0*2, ARMPB1*1, ARMPB2*4, first_diff_mean*5, sec_diff_mean*6, Hjorth_complexity*1, Pwelch3*1, skew*3, slope_mean*5, slope_var*1, wavelet_cA_mean*3, wavelet_cD_mean*4, wavelet_cD_std*3	ARMPB0*1, slope_mean*1

tion recognition tasks using EEG [15], [16], are exclusively selected for *Obs* and *Int* labels only (not for *Self*). The Hjorth parameters measure the complexity of the signal and are usually considered as a good measure quantifying event related properties of EEG signals. Again similar to ECG, we could infer that there seems to be an unconscious emotion reaction reflected in this modality that is not properly captured in the self annotation. We observe that the external annotator could identify an individual's affective status the most, i.e, the participants are indeed being stimulated (evident in the internal measures) and expressed explicitly (hence observed by annotators), but the participants themselves are not able to report these states via self-appraisal. Besides, we also report the selected subject independent brain regions in table IV. We can see that T7 and T8, which are from temporal lobes of the left and right hemisphere respectively, are repeatedly chosen for both arousal and valence recognition. These temporal regions selected are consistent with past researches in happiness detection [17] and emotional movie response [18]. Most of the insights demonstrated in this section show that participants in this dataset seem to be successfully stimulated and reflected their emotional status externally (i.e., both *Int* and *Obs* achieve

TABLE IV: The number of filtered EEG features from each channel. The bold one refers to features selected over 5 times and '*' indicates the maximum of times selected for the given label.

	AF3	F7	F3	FC5	T7	P7	O1	O2	P8	T8	FC6	F4	F8	AF4
Arousal														
Self / Obs / Int	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Self / Obs	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Self / Int	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Obs / Int	1	2	0	0	**5**	0	0	0	1	1	0	0	0	0
Self Only	0	0	1	0	0	0	0	0	0	2	0	0	0	0
Obs Only	0	3	3	1	**6**	1	1	1	3	**8***	5	2	2	0
Int Only	0	0	0	2	0	0	0	0	0	0	1	0	1	2
Valence														
Self / Obs / Int	0	0	0	0	0	0	0	0	0	0	1	0	1	0
Self / Obs	0	0	0	0	1	1	0	0	0	0	0	0	0	0
Self / Int	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Obs / Int	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Self Only	0	0	0	1	0	1	0	0	0	1	0	0	0	0
Obs Only	4	2	4	2	**6***	2	2	2	**5**	4	2	1	3	0
Int Only	0	0	0	0	0	0	0	0	1	0	0	0	0	1

TABLE V: Error indexes in each video stimuli for self labels. Note that the VideoID refers to the video clips ID described in [9] Int Labels: Original intended stimuli type during the experiment, 0 maps to low while 1 maps to high stimuli level.

VideoID	10	13	138	18	19	20	23	30	31	34	36	4	5	58	80	9
Arousal																
ER	0.33	0.35	0.26	0.48	0.45	0.44	0.41	0.5	0.59	0.46	0.59	0.3	0.22	0.44	0.45	0.33
ER_Diff_OBS	0.14	0.19	0.13	0.33	0.26	0.21	0.11	0.16	0.11	0.28	0.14	0.05	0.08	0.13	0.29	0.19
ER_Ratio_OBS	0.42	0.54	0.5	0.68	0.59	0.47	0.27	0.32	0.18	0.61	0.23	0.18	0.38	0.29	0.65	0.58
Valence																
ER	0.42	0.41	0.26	0.6	0.21	0.41	0.38	0.34	0.19	0.46	0.35	0.38	0.54	0.33	0.47	0.33
ER_Diff_OBS	0.14	0.05	0.18	0.1	0.08	0.33	0.3	0.21	0.14	0.23	0.22	0.27	0.35	0.13	0.34	0.28
ER_Ratio_OBS	0.33	0.13	0.7	0.17	0.38	0.81	0.79	0.62	0.71	0.5	0.62	0.71	0.65	0.38	0.72	0.83
Int Labels																
INT_ARO	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1
INT_VAL	1	1	0	1	0	0	0	0	0	0	0	1	1	1	1	1

higher recognition rates when learning from the subject's physiology), but for some reason, there exists a higher level of mismatch between awareness of their self affective status to their own bodily internal physiological response.

2) Error Analysis: From the previous section, we observe that self-evaluation of mental state has a larger mismatch from the participants' physiological responses. We would like to further investigate a question: in what circumstances would this happen more often? In this section, we examine the incorrectly predicted samples of the *Self*, specifically examining the prediction model learned using a concatenation of three modalities. Given a set of data with N samples, we derive several indices for this error analysis:

- **Error Rate (ER)**: $(fp + fn)/N$, fp: false positive, fn: false negative
- **Error Rate Diff OBS (ER_Diff_OBS)**: $(fp_{obs} + fn_{obs})/N$, fp_{obs}: false positive for samples with different label classes between *Self* and *Obs*
- **Error Rate Diff Ratio OBS (ER_Ratio-OBS)**: ER_Diff_OBS/ER

Here we calculate the above index for every video stimuli (a stimulus is used for multiple subjects), and the results are presented in table V. By conducting a two-tailed Student's t-test, we observe that Arousal's ER and Valence's ER are slightly ($t = 2.09, p = 0.05585$) and significantly ($t = -2.27, p = 0.039601$) correlated to the original stimuli's valence label (INT_VAL) respectively. In other words, for arousal stimuli containing a lower degree of pleasant content, it raises the probability of mismatch between subjects physiological responses and their self-assessments. On the other hand, when the video stimuli include contents of higher valence level, it corresponds more to the errors in recognizing a participant's self-evaluation using their physiology.

In addition, we also see that valence's ER_Ratio_OBS is related to stimuli's arousal label ($t : -2.06, p : 0.05861$). This indicates that while there exists a mismatch between physiological response and the self-assessment, the external observer's evaluation tends to be more alike to the participant's physiological reaction, only if the source stimuli encourage to stimulate higher arousal of the participant. In conclusion, these analyses show that the polarity of the emotion-triggering source would lead to a potential bias in the self emotion assessment process, which may also underscore the reason that we observe a significant discrepancy (degradation) of recognition results obtained between self-assessment and external observations in this work.

V. CONCLUSION

Previous research works have shown that there could be either bias or neglection of self-assessment on emotion. Hence, it is important to investigate emotion responses from distinct labeling viewpoint in order to better understand the relationship between physiology and the emotional reaction when stimulated. In this work, we comprehensively inspect the predictability of emotion annotations from three different perspectives: self-assessment, external observation, and intended stimuli. Our experiments show that there exist several interesting patterns on the recognition results. ECG and EEG data consistently obtain better discriminative power for the external observations while the EDA signal tends to work better on original stimuli's label. Furthermore, feature importance analysis shows that several well-known emotion-related key physiological parameters are selected on observed and intended emotion labels; however, the same effect is not evident in the prediction model for self emotion annotation. Finally, from our error analysis, we show that the mismatch between self-assessment and the physiology may be conditioned on the type of original emotional stimuli's valence polarity.

To our knowledge, this is one of the first works in providing comprehensive recognition and analyses on multi-perspective of emotion annotations using physiology. We can foresee several future directions. An immediate work would be to take the personal attributes of the participant into consideration. These attributes like age, gender or implicit measurements like personality and moods, possibly are additional hidden modulating factors during the cognitive-appraisal process of one's self emotion assessment. Second, our error analysis suggests that the types of stimuli could also be a key component in affecting the physiological responses and potentially inducing the bias in the self emotion assessment. Through better understanding the relationship of these multiple perspectives of emotion annotations and the measured physiological responses could help enhance the robustness of affective recognition module that can be integrated for many human behavior modeling applications [19], [20].

REFERENCES

[1] J. T. Cacioppo, L. G. Tassinary, and G. Berntson, *Handbook of psychophysiology*. Cambridge University Press, 2007.

[2] S. Wioleta, "Using physiological signals for emotion recognition," in *2013 6th International Conference on Human System Interactions (HSI)*. IEEE, 2013, pp. 556–561.

[3] 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, 2012.

[4] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[5] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2018.

[6] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state." *Psychological review*, vol. 69, no. 5, p. 379, 1962.

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

[8] L. Ivonin, H.-M. Chang, M. Diaz, A. Catala, W. Chen, and M. Rauterberg, "Traces of unconscious mental processes in introspective reports and physiological responses," *PloS one*, vol. 10, no. 4, p. e0124519, 2015.

[9] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *arXiv preprint arXiv:1702.02510*, 2017.

[10] M. D, "Neurokit: A python toolbox for statistics and neurophysiological signal processing (eeg, eda, ecg, emg...)," Day, 01 November 2016, paris, France.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[13] J. A. Bargh and E. Morsella, "The unconscious mind," *Perspectives on psychological science*, vol. 3, no. 1, pp. 73–79, 2008.

[14] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[15] R. M. Mehmood and H. J. 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.

[16] A. Patil, C. Deshmukh, and A. Panat, "Feature extraction of eeg for emotion recognition using hjorth features and higher order crossings," in *2016 Conference on Advances in Signal Processing (CASP)*. IEEE, 2016, pp. 429–434.

[17] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, "Real-time eeg-based happiness detection system," *The Scientific World Journal*, vol. 2013, 2013.

[18] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[19] S. Narayanan and P. 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.

[20] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. 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.