Stable Feature Selection for EEG-based Emotion Recognition

Zirui Lan
Fraunhofer Singapore
Singapore
LanZ0001@e.ntu.edu.sg

Olga Sourina
Fraunhofer Singapore
Singapore
EOSourina@ntu.edu.sg

Lipo Wang
School of EEE
Nanyang Technological University
Singapore
ELPWang@ntu.edu.sg

Yisi Liu
Fraunhofer Singapore
Singapore
LiuYS@ntu.edu.sg

Reinhold Scherer
Institute of Neural Engineering
Graz University of Technology
Graz, Austria
Reinhold.Scherer@tugraz.at

Gernot R. Müller-Putz
Institute of Neural Engineering
Graz University of Technology
Graz, Austria
Gernot.Mueller@tugraz.at

***Abstract*—Affective brain-computer interface (aBCI) introduces personal affective factors into human-computer interactions, which could potentially enrich the user's experience during the interaction with a computer. However, affective neural patterns are volatile even within the same subject. To maintain satisfactory emotion recognition accuracy, the state-of-the-art aBCIs mainly tailor the classifier to the subject-of-interest and require frequent re-calibrations for the classifier. In this paper, we demonstrate that the recognition accuracy of aBCIs deteriorates when re-calibration is ruled out during the long-term usage for the same subject. Then, we propose a stable feature selection method to choose the most stable affective features, for mitigating the accuracy deterioration to a lesser extent and maximizing the aBCI performance in the long run. We validate our method on a dataset comprising six subjects' EEG data collected during two sessions per day for each subject for eight consecutive days.**

Keywords—Electroencephalography (EEG), stable feature, feature selection, emotion recognition, intra correlation coefficient (ICC)

I. Introduction

The state-of-the-art EEG-based emotion recognition algorithms tailor the classifier to each individual user, requiring a calibration session before the subject starts to use the system. However, due to the volatile affective neural patterns, frequent re-calibrations are needed during use to maintain satisfactory recognition accuracy. A great amount of the existing studies [1-11] investigate and report only the short-term performance of an affective BCI. In these studies, affective EEG data are collected within a short period of time, usually in one single day within one experiment session. K-fold cross-validations are carried out to assess the system performance. The recognition accuracy assessed in this way is over-optimistic and can hardly represent the system performance in the long run, especially when no re-calibration is allowed during use. On the other hand, there is little study of the long-term affective BCI performance. We devote this paper to the investigation of affective BCI performance over a long course of time. As the (re-)calibration process may be time-consuming, tedious and laborious, we are motivated to mitigate the burden of frequent re-calibrations on the user of interest. Ideally, a stable affective EEG feature should give consistent measurement of the same emotion on the same subject over a long course of time. We hypothesize that unstable features may worsen the recognition performance of the BCI in the long run. By using stable EEG features, recognition accuracy may be improved. We introduce an ANOVA-based method to quantify the stability score of the state-of-the-art affective EEG features. We then propose a stable feature selection method to choose the optimal set of stable features that maximize the recognition accuracy of the system in the long run.

This paper is organized as follows. Section II explains the methodologies. Section III documents the experiments. Section IV presents the results with discussions. Section V concludes this chapter.

II. Methods

A. Feature Extraction

1) Fractal Dimension

Let $\boldsymbol{x} \in \mathbb{R}^n$ denote a column vector of n EEG time series samples (raw signals) from one channel. Construct k new time series by re-sampling $\boldsymbol{x}$ as follows.

$$\boldsymbol{x}_k^m = \left[\boldsymbol{x}(m), \boldsymbol{x}(m+k), \dots, \boldsymbol{x}\left(m + \left\lfloor \frac{n-m}{k} \right\rfloor k\right)\right]^{\mathsf{T}}, m = 1, 2, \dots, k, \quad (1)$$

where $\lfloor \cdot \rfloor$ denotes the floor function, m the initial time series sample and k the interval. We compute the length of the curve for each new series as follows.

$$l_k^m = \frac{1}{k}\left\{\left(\sum_{i=1}^{\left\lfloor \frac{n-m}{k} \right\rfloor} |\boldsymbol{x}(m+ik) - \boldsymbol{x}(m+(i-1)k)|\right)\right\}\left(\frac{n-1}{\left\lfloor \frac{n-m}{k} \right\rfloor k}\right), \quad (2)$$

Let l_k denote the mean of l_k^m for $m = 1, 2, \dots k$, the fractal dimension of time series $\boldsymbol{x}$ is computed as [12]

978-1-5386-7315-7/18/$31.00 ©2018 IEEE
DOI 10.1109/CW.2018.00042

$$FD = -\lim_{k\to\infty}\frac{\log(l_k)}{\log(k)}, \tag{3}$$

Apparently, in numerical evaluation, it is not possible for k to be infinite. It has proven [13, 14] that the computed fractal value approximates the true, theoretical fractal value reasonably well given a reasonably large k. Based on the study in [14], $k =$ 32 yields a good balance between accuracy and computational resources required. In this study, we follow the same parameter setting.

2) Statistics

A set of six statistical features were adopted in [15] for EEG-based emotion recognition, which, in combination with the fractal dimension feature, have been demonstrated to improve the classification accuracy [15]. Six statistical features are computed as follows.

Mean of the raw signals:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}(i), \tag{4}$$

Standard deviation of the raw signals:

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}(i)-\mu_x)^2}, \tag{5}$$

Mean of the absolute values of the first order difference of the raw signals:

$$\delta_x = \frac{1}{n-1}\sum_{i=1}^{n-1}|\boldsymbol{x}(i+1)-\boldsymbol{x}(i)|, \tag{6}$$

Mean of the absolute values of the first order difference of the normalized signals:

$$\tilde{\delta}_x = \frac{1}{n-1}\sum_{i=1}^{n-1}|\tilde{\boldsymbol{x}}(i+1)-\tilde{\boldsymbol{x}}(i)| = \frac{\delta_x}{\sigma_x}, \tag{7}$$

Mean of the absolute values of the second order difference of the raw signals:

$$\gamma_x = \frac{1}{n-2}\sum_{i=1}^{n-2}|\boldsymbol{x}(i+2)-\boldsymbol{x}(\mathrm{i})|, \tag{8}$$

Mean of the absolute values of the second order difference of the normalized signals:

$$\tilde{\gamma}_x = \frac{1}{n-2}\sum_{i=1}^{n-2}|\tilde{\boldsymbol{x}}(i+2)-\tilde{\boldsymbol{x}}(i)| = \frac{\gamma_x}{\sigma_x}. \tag{9}$$

In (4)–(9), $\tilde{\boldsymbol{x}}$ denotes the normalized (zero mean, unit variance) signals, i.e., $\tilde{\boldsymbol{x}} = (\boldsymbol{x}-\mu_x)/\sigma_x$.

3) Spectral Band Power

Spectral band power, or simply "power", is one of the most extensively used features in EEG-related research [1, 3, 7, 9, 11]. In EEG study, there is common agreement on partitioning the EEG power spectrum into several sub-bands (though the frequency range may slightly differ from case to case): alpha band, theta band, beta band etc. In our study, the EEG power features from theta band (4 – 8 Hz), alpha band (8 – 12 Hz), and beta band (12 – 30 Hz) are computed.

The power features are obtained by first computing the Fourier Transform on the EEG signals. The discrete Fourier Transform transforms a time-series $\boldsymbol{x} = [\boldsymbol{x}(1), \boldsymbol{x}(2), \dots, \boldsymbol{x}(N)]^\top$ to another series $\boldsymbol{s} = [\boldsymbol{s}(1), \boldsymbol{s}(2), \dots, \boldsymbol{s}(N)]^\top$ in a frequency domain. $\boldsymbol{s}$ is computed as

$$\boldsymbol{s}(k) = \sum_{n=0}^{N-1}\boldsymbol{x}(n)e^{-\frac{j2\pi kn}{N}}, \tag{10}$$

where N is the number of sampling points. Then, the power spectrum density is computed as

$$\hat{\boldsymbol{s}}(k) = \frac{1}{N}|\boldsymbol{s}(k)|^2, \tag{11}$$

Lastly, the spectral band power features are computed by averaging the power spectrum density $\hat{\boldsymbol{s}}(k)$ over the targeted sub-band. E.g., the alpha band power is computed by averaging $\hat{\boldsymbol{s}}(k)$ over 8 – 12 Hz.

4) Higher Order Crossing

Higher Order Crossings (HOC) was proposed in [16] to capture the oscillatory pattern of EEG, and used in [15, 17-19] as features to recognize human emotion from EEG signals. The HOC is computed by first zero-meaning the time-series $\boldsymbol{x}$ as

$$\boldsymbol{z}(i) = \boldsymbol{x}(i) - \mu_x, \tag{12}$$

where $\boldsymbol{z}$ is the zero-meaned series of $\boldsymbol{x}$ and μ_x the mean of $\boldsymbol{x}$ computed as per (4). Then, a sequence of filter ∇ is successively applied to $\boldsymbol{z}$, where ∇ is the backward difference operator, $\nabla \equiv \boldsymbol{z}(i) - \boldsymbol{z}(i-1)$. Denote the *k*th-order filtered sequence of $\boldsymbol{z}$ as $\boldsymbol{\xi}_k(\boldsymbol{z})$, $\boldsymbol{\xi}_k(\boldsymbol{z})$ is obtained by iteratively applying ∇ on $\boldsymbol{z}$, as

$$\boldsymbol{\xi}_k(\boldsymbol{z}) = \nabla^{k-1}\boldsymbol{z}, \nabla^0\boldsymbol{z} = \boldsymbol{z}. \tag{13}$$

Then, as its name suggests, the feature consists in counting the number of zero-crossing, which is equivalent to the times of sign changes, in sequence $\boldsymbol{\xi}_k(\boldsymbol{z})$. We follow [15] and compute the HOC feature of order $k = 1, 2, 3, \dots, 36$.

5) Signal Energy

The signal energy is the sum of squared amplitude of the time-series signal [20], computed as

$$\varepsilon = \sum_i|\boldsymbol{x}(i)|^2. \tag{14}$$

6) Hjorth Feature

Hjorth [21] proposed three features of a time-series, which have been used as affective EEG features in [22, 23].

Activity:

$$a(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}(i)-\mu_x)^2, \tag{15}$$

where μ_x is the mean of $\boldsymbol{x}$ computed as per (4).

Mobility:

$$m(\boldsymbol{x}) = \sqrt{\frac{\mathrm{var}(\dot{x})}{\mathrm{var}(x)}}, \tag{16}$$

where $\dot{x}$ is the time derivative of the time-series $\boldsymbol{x}$, and var($\cdot$) is the variance operator.

Complexity:

$$c(\boldsymbol{x}) = \frac{m(\dot{x})}{m(x)}, \tag{17}$$

which is the mobility of the time derivative of $\boldsymbol{x}$ over the mobility of $\boldsymbol{x}$.

B. Feature Stability Measurement

The stability of feature parameters was quantified by the Intraclass Correlation Coefficient (ICC). ICC allows for the

TABLE I THE ANALYSIS OF VARIANCE TABLE.

Treatment (emotion)	**Measurement**				**Total**	**Average**
1	x_{11}	x_{12}	$\cdots$	x_{1k}	$x_{1\cdot}$	$\bar{x}_{1\cdot}$
2	x_{21}	x_{22}	$\cdots$	x_{2k}	$x_{2\cdot}$	$\bar{x}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$x_{n\cdot}$	$\bar{x}_{n\cdot}$
					$x_{\cdot\cdot}$	$\bar{x}_{\cdot\cdot}$

Source of variance	***Sum of squares***	***Degree of freedom***	***Mean square***
Between treatment	$SS_B = k\sum_{i=1}^{n}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$	$n-1$	$MS_B = SS_B/(n-1)$
Within treatment	$SS_W = SS_T - SS_B$	$nk-n$	$MS_W = SS_E/(nk-n)$
Total	$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x}_{\cdot\cdot})^2$	$nk-1$	

assessment of similarity in grouped data. It describes how well the data from the same group resemble each other. ICC was often used in EEG stability study [24, 25]. ICC is derived from a one-way ANOVA model and defined as [26]

$$\mathrm{ICC} = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}, \tag{18}$$

where MS_B, MS_W and k denote the mean square error between groups, the mean square error within group, and the number of samples in each group, respectively. A larger ICC value indicates higher similarity among group data. ICC tends to one when there is absolute agreement among the grouped data, i.e., $MS_W = 0$. A smaller ICC value suggests a lower similarity level. ICC value can drop below zero in the case when MS_W is larger than MS_B, accounting for dissimilarity among the grouped data.

C. Stable Feature Selection

A stable affective EEG feature should give consistent measurements of the same emotion on the same subject over the course of time, therefore there is the possibility to reduce the need of re-calibration by using the stable features. To this end, we propose a stable feature selection method based on ICC score ranking. The proposed method consists of three steps: ICC assessment, ICC score ranking, and iterative feature selection.

We assess the long-term stability of different EEG features with ICC. Let $\boldsymbol{X}$ be the matrix of feature parameters of a specific kind of feature, rows of $\boldsymbol{X}$ correspond to different emotions, and columns of $\boldsymbol{X}$ correspond to different repeated measurements over the course of time. Intuitively, we want the feature parameters to be consistent when measuring the same emotion repeatedly over the course of time. Therefore, we want the parameters within the same row to be similar to each other. Moreover, we want the parameters measuring different affective states to be discriminative, so that different affective states are distinguishable. Therefore, we want different rows to be dissimilar to each other. The ICC measurement takes both considerations into account. The ICC is computed as per (18), which is based on ANOVA. For clarity, we display $\boldsymbol{X}$ in the ANOVA table as in Table I. In Table I, we refer treatment to different emotions induced by specific affective stimuli. x_{ij} is the feature parameter of the *j*-th measurement of emotion i. $x_{i\cdot}$ is the sum of all measurements of emotion i, $x_{i\cdot} = \sum_{j=1}^{k} x_{ij}$. $\bar{x}_{i\cdot}$ is the average of all measurements of emotion i, $\bar{x}_{i\cdot} = (1/k)\sum_{j=1}^{k} x_{ij}$. $x_{\cdot\cdot}$ is the sum of all measurements over all emotions, $x_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$. $\bar{x}_{\cdot\cdot}$ is the average of all measurements over all emotions, $\bar{x}_{\cdot\cdot} = (1/nk)\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$.

We can obtain the stability score of each feature by computing the ICCs, thereafter, we rank the feature according to the stability score in descending order. Features with higher ICC are more stable over the course of time, and exhibit better discriminability among different emotions. Our proposed feature selection method consists in iteratively selecting the top stable features and validating the inter-session emotion recognition accuracy. The feature subset that yields the best accuracy is retained.

III. EXPERIMENTS

A. Data Collection

The stability of affective EEG features is of our interest of investigation. In contrast to existing affective EEG benchmark dataset such as the DEAP dataset [27], which includes a relatively large number of subjects but only one EEG recording session within one day for each subject, we designed and carried out an experiment to collect the affective EEG data from multiple sessions during the course of several days. This preliminary study included six subjects, five males and one female, aged 24 – 28. All subjects reported no history of mental diseases or head injuries. Two sessions were recorded per day for each subject for eight consecutive days, i.e., 16 sessions were recorded for each subject. An Emotiv EEG device [28] was used to record the EEG data at a sampling rate of 128Hz. Each session consisted of four trials, with each trial corresponding to one induced emotion, i.e., four emotions were elicited in one session, so totally each subject has 4×2×8 = 64 trials. There are standard affective stimuli libraries such as International Affective Picture System (IAPS) [29] and International Affective Digitized Sounds (IADS) [30]. In our study, the IADS was chosen for the experiment design as during the exposure of the subjects to the audio stimuli, the subjects can keep their eyes closed and hence avoid possible ocular movements which could contaminate the EEG signals. The emotion induction experiment protocol followed work [10]. Sound clips from the same category of the IADS were chosen and concatenated to make a 76-second audio file, with the first 16 seconds silent to calm the subject down. Four audio files were used as stimuli to evoke four different emotions, namely pleasant, happy, angry and frightened. During each session of the experiment only one subject was invited to the lab and was well-instructed about the protocol of the experiment. The subject wore the Emotiv EEG device and a pair of earphones with volume properly adjusted, and he/she was required to sit still with eyes closed and avoided muscle movements as much as possible to reduce possible artifacts from eyeballs movement, teeth clenching, neck movement etc. Following each trial, the subject was required to complete a self-assessment to describe his emotion (happy, frightened etc.). This self-assessment was used as a ground truth to assess that the subject has experienced the target emotion we wish to induce. The protocol of this emotion induction experiment is depicted in Fig. 1.

TABLE II REFERENCED STATE-OF-THE-ART AFFECTIVE EEG FEATURES

Feature (dimension, abbreviation)	Reference
6 statistics (30, STAT)	[9, 10, 15, 36, 37]
36 higher order crossings (180, HOC)	[15, 17-19]
Fractal dimension + 6 statistics + 36 higher order crossings (215, FD1)	[10, 15]
Fractal dimension + 6 statistics (35, FD2)	[10, 15]
3 Hjorth (15, HJORTH)	[21, 22]
Signal energy (5, SE)	[20]
Spectral power of δ, θ, α, and β bands (20, POW)	[1, 4, 9, 38]

B. Simulation 1: With Re-calibration

In this experiment, we simulate the recognition performance of an affective BCI where re-calibration of the system can be carried out each time before the subject uses the system. Specifically, we evaluate the within-session cross-validation recognition accuracy using the state-of-the-art affective EEG features referenced in Table II.

We base the simulation on the EEG data we collected in Section III.A. Each EEG trial lasts for 76 seconds. We discard both ends of the EEG trial and retain the middle part of the EEG trial for the subsequent processing, based on the assumption that emotions are better elicited in the middle of the trial. The division of the EEG trial is illustrated in Fig. 2. EEG features are extracted out of the valid segments of the EEG trials on a sliding-windowed basis. The final feature vector is a concatenation of the feature vectors from channel AF3, F7, FC5, T7, and F4, which were justified in [14] to be the top five discriminative channels concerning emotion recognition. The width of the window is 4-second, and the step of the move is 1-second, as was used in [14]. Thus, each valid segment yields 7 samples.

In this within-session cross-validation evaluation, the training data and test data are from the EEG trials within the same session. As the time gap between the acquisition of training and test data is minimal, the evaluation can approximate the performance of the BCI where calibration is carried out shortly before use. We use one valid segment as the training data and the other as the test data, and repeat the process until each segment has served as the test data for once. The per-session recognition accuracy is averaged across all possible runs. In this very case, the evaluation is repeated twice per session, which is referred to as a two-fold cross validation. As we recognize four emotions in each session, the training data comprise 7×4 = 28 samples for four emotions, totally. Likewise, the test data consist of 28 samples for four emotions. We adopt the Logistic Regression (LR) [31] classifier. The simulation is implemented in MATLAB R2017a, where we use the MATLAB built-in toolbox of the LR classifier with the default hyperparameters. The evaluation is carried out for each of the subjects on a session-by-session basis. The mean classification accuracy over 16 sessions and the standard deviations are displayed in Table III.

C. Simulation 2: Without Re-calibration

In this experiment, we simulate the recognition performance where no re-calibration is allowed during the long-term use of the BCI. We evaluate the inter-session leave-one-session-out cross-validation accuracy of the system for this purpose. Recall that in our dataset, we have 16 recording sessions per subject throughout the course of eight days. In this evaluation, we reserve one session as the calibration session whose EEG data are used to train the classifier, and pool together the data from the remaining 15 sessions as test data. We repeat the evaluation until each session has served as calibration session for once. In this very case, the process will be repeated 16 times per subject, and the reported recognition accuracy is the mean accuracy of 16 runs. This evaluation is to simulate the system performance in the long run, since there is a longer time gap between the training session and testing sessions—up to eight days. We adopt the features referenced in Table II in this simulation, in the same sliding-windowed manner as in Section III.B. We use only the valid segment 1 (see Fig. 2) of each EEG trial and reserve the valid segment 2 for the testing purpose in Simulation 3 introduced in the following section. The sliding-windowed feature extraction yields 7 samples per valid segment. The training data consist of 7×4 = 28 samples for four emotions recorded in the same session. The test data comprise 7×4×15 = 420 samples pooled together from the remaining 15 sessions. The mean classification accuracy over 16 runs and the standard deviations are displayed in Table IV.

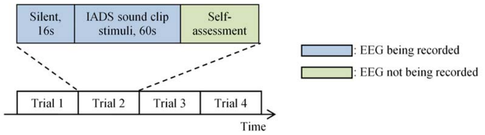

Fig. 1 Protocol of emotion induction experiment.

Discarded	Valid segment 1	Valid segment 2	Discarded
1 – 20 sec	21 – 30 sec	31 – 40 sec	41 – end

Fig. 2 Division of the EEG trial. EEG data at both ends are discarded. The middle part is retained and divided into two valid segments of the same length. Only valid segments are used for the subsequent processing.

D. Simulation 3: Stable Feature Selection

In this experiment, we validate the effect of our proposed stable feature selection algorithm based on the simulation of emotion recognition where no re-calibration is allowed during the long-term use of the BCI. This simulation is similar to simulation 2, with the focus on the comparison between the state-of-the-art feature set and the stable feature set we propose.

We propose to find the stable features on a subject-dependent basis. The subject-dependent evaluation intends to find subject-specific stable features for each subject. We quantify the long-term feature stability by computing the ICC scores on the training set consisting of the valid segment 1 (see Fig. 2) from all available trials (16 trials per subject), rank the feature according to the stability scores, and retain the optimal subset of features pertinent to the subject in question that maximizes the recognition accuracy when iteratively evaluating the inter-session leave-one-session-out cross-validation accuracy using the top n stable features. The results are shown in Table V and Fig. 3. After we find the stable features, we evaluate the performance of the stable features on the test set comprising the valid segment 2 (see Fig. 2) from all available trials. The recognition performance on the test set is shown in Table VI.

TABLE III FOUR-EMOTION RECOGNITION ACCURACY OF SIMULATION 1, MEAN ACCURACY (%) ± STANDARD DEVIATION (%)

Feature	Subject					
	1	*2*	*3*	*4*	*5*	*6*
STAT	56.81 ± 10.52	44.75 ± 16.66	43.64 ± 13.89	71.43 ± 14.32	47.92 ± 15.44	73.88 ± 15.29
HOC	32.25 ± 10.50	30.25 ± 10.05	28.46 ± 10.24	43.53 ± 12.20	28.37 ± 10.95	36.61 ± 12.29
FD1	43.08 ± 13.98	37.39 ± 12.58	33.59 ± 8.12	58.59 ± 13.40	39.58 ± 12.05	54.58 ± 11.03
FD2	**57.14** ± 9.93	**46.88** ± 17.25	**45.76** ± 13.01	72.54 ± 14.49	**48.91** ± 15.42	**76.23** ± 15.51
HJORTH	53.24 ± 11.81	46.65 ± 14.30	41.41 ± 14.39	**72.77** ± 17.82	47.92 ± 15.67	72.54 ± 18.78
SE	45.54 ± 15.95	40.63 ± 12.67	41.96 ± 17.57	59.49 ± 16.23	41.96 ± 18.90	62.83 ± 20.02
POW	48.66 ± 12.21	46.88 ± 17.72	36.05 ± 14.70	69.20 ± 15.83	42.26 ± 18.03	62.72 ± 16.00
Upp Chan Lvl	42.79	42.80	42.79	39.36	42.70	42.79

TABLE IV FOUR-EMOTION RECOGNITION ACCURACY OF SIMULATION 2, MEAN ACCURACY (%) ± STANDARD DEVIATION (%)

Feature	Subject					
	1	*2*	*3*	*4*	*5*	*6*
STAT	**37.95** ± 5.01	24.79 ± 1.77	25.61 ± 1.65	39.49 ± 6.95	27.00 ± 3.98	30.39 ± 6.24
HOC	26.55 ± 4.27	24.78 ± 2.72	25.51 ± 2.63	28.68 ± 4.01	25.68 ± 2.78	27.01 ± 3.05
FD1	28.93 ± 3.98	24.52 ± 2.27	25.13 ± 2.83	33.68 ± 5.58	25.82 ± 3.01	28.45 ± 3.67
FD2	37.38 ± 6.05	25.25 ± 2.68	25.16 ± 2.62	**39.70** ± 7.10	27.52 ± 3.88	29.61 ± 6.25
HJORTH	31.77 ± 6.05	25.85 ± 3.33	27.05 ± 3.84	35.19 ± 8.13	26.32 ± 3.96	28.18 ± 4.82
SE	28.07 ± 2.83	25.80 ± 3.04	26.99 ± 2.79	38.35 ± 5.97	**27.96** ± 4.37	28.53 ± 3.84
POW	30.49 ± 4.30	**28.41** ± 4.25	**28.01** ± 3.55	39.42 ± 6.44	27.63 ± 4.53	**31.49** ± 6.94
Upp Chan Lvl	29.33	29.09	28.83	28.30	27.75	28.85

TABLE V FOUR-EMOTION RECOGNITION ACCURACY OF SIMULATION 3 USING THE TOP *N* STABLE FEATURES. MEAN ACCURACY (%) ± STANDARD DEVIATION (%) (# OF STABLE FEATURES)

Feature	Subject					
	1	*2*	*3*	*4*	*5*	*6*
Our Selected Stable Feature	41.55 ± 4.31 (2)	30.24 ± 5.14 (7)	33.87 ± 3.55 (5)	45.22 ± 4.57 (1)	30.68 ± 3.43 (42)	33.63 ± 7.99 (34)

IV. RESULTS AND DISCUSSIONS

A. Simulation 1: With Re-calibration

Table III shows the mean accuracy ± standard deviation per subject based on the 2-fold cross-validation evaluation, which simulates the use case where re-calibration is allowed each time before a subject uses the BCI. The recognition accuracies vary between subjects and features, ranging from 28.37 % (Subject 5, HOC) to 76.23 % (Subject 6, FD2). HOC is found to be inferior to other referenced features on all subjects. The best performing feature varies between subjects. For subject 1, 2, 3, 5, and 6, referenced feature set FD2 yield better recognition accuracy than other referenced features in most cases. For subject 2, FD2, POW and HJORTH features give similar performance, outperforming other referenced features. For subject 4, STAT, FD2 and HJORTH features yield comparable results, being better than other referenced features. In general, FD2 performs well on all subjects in this simulation, which may suggest that FD2 is good for the use case where re-calibration is allowed from time to time.

For a four-class classification task, the theoretical chance level of random guess is 25.00 %. However, it is known that the real chance level is dependent on the classifier as well as the number of test samples. For an infinite number of test samples, the real chance level approaches the theoretical value. For a finite number of test samples, the real chance level is computed based on repeated simulations of classifying samples with randomized class label, as is suggested in [32, 33]. We carry out such simulation and present also in Table III the upper bound of the 95 % confidence interval of the simulated chance level for the best performing feature (in bold) for each classifier. Results show that the best-performing features yield recognition accuracy higher than the upper bound of the chance level. We assert that the best-performing features perform significantly better than chance level at a 5 % significance level.

B. Simulation 2: Without Re-calibration

Table IV shows the mean accuracy ± standard deviation per subject based on inter-session leave-one-session-out cross-validation evaluation, which simulates the long-term recognition performance of the BCI when no re-calibration is permitted during use. Notable accuracy drop can be observed, compared to when re-calibration is allowed at each new session. This experiment establishes that intra-subject variance of affective feature parameters does exist and does have a negative impact on the recognition performance, though the severity varies from subject to subject. For subject 2 and 3, the recognition performance is severely affected by the variance—the best recognition performance has dropped and fallen within the 95 % confidence interval of the simulated chance level. We therefore assert that subject 2 and 3 are performing at random guess level. For subject 1, 4 and 6, the best performance remains significantly better than the chance level at 5 % significance level, which seems to suffer from the variance problem to a lesser extent. Subject 5 gives mediocre performance. We loosely categorize subject 1, 4, and 6 as good performer, subject 5 as moderate performer and subject 2 and 3 as weak performer.

C. Simulation 3: Stable Feature Selection

To improve the long-term recognition accuracy, we propose to use stable features to mitigate the intra-subject variance of the

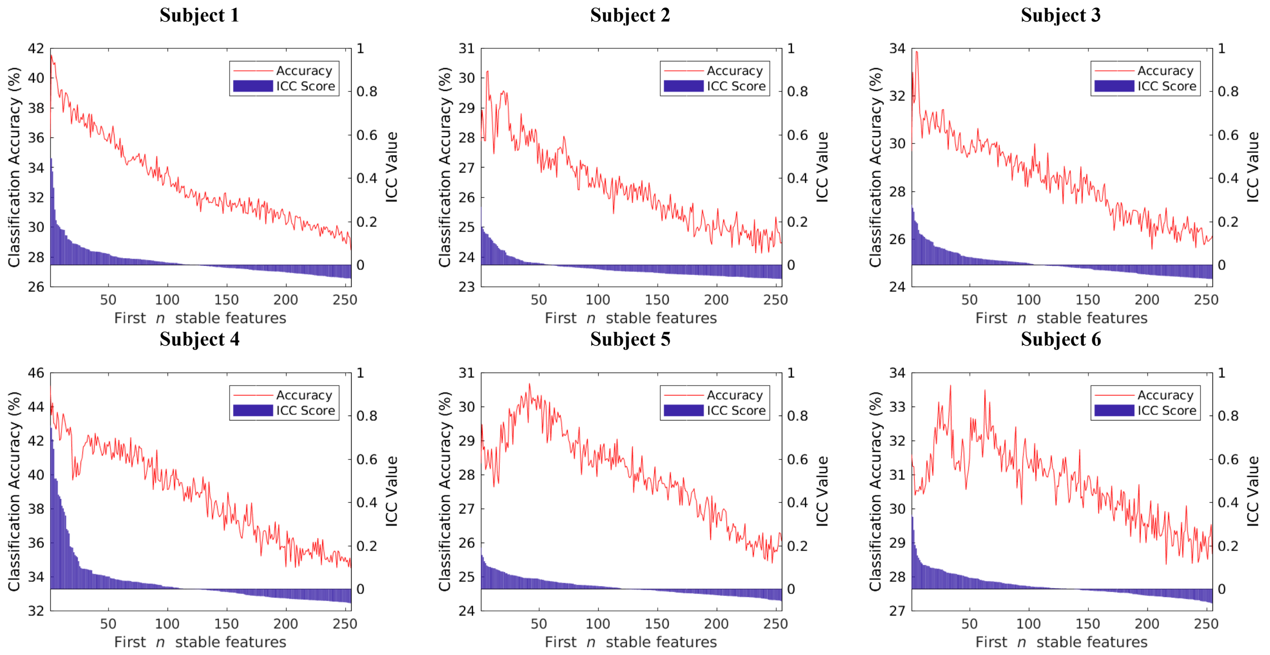

Fig. 3 ICC scores of each feature and the inter-session leave-one-session-out cross-validation accuracy using the top n stable features, $1 \leq n \leq 255$. The features are ranked by the ICC score in descending order.

affective feature parameters. Ideally, stable feature should give consistent measurement of the same affective state over the course of time, therefore there is the possibility to mitigate the variance among repeated sessions on different days. We propose a feature selection method that consists in quantifying the long-term stability of features with ICC model, ranking the features according to stability scores and iteratively selecting the topmost stable feature for inclusion into the stable feature subset. We propose to find the subject-dependent stable features.

Fig. 3 presents the results of subject-dependent stable feature selection. The bar plot in Fig. 3 indicates the stability score given in ICC values. The higher the stability score, the less variance the feature exhibits. The stability scores are ranked in descending order. Table VII shows the ranking of the top 10 most stable features and their respective ICC scores. As we can see, the feature stability varies from subject to subject. For subject 1 and 4, the stability scores of the topmost stable features are notably higher than that of the other subjects. Generally, we observe that only a fraction of the features carries positive stability scores. For those with negative stability score, it suggests that the variance of the feature parameters over the course of time is even larger than the variance of the feature parameters between different emotions. Intuitively, these unstable features contribute to the deterioration of long-term recognition performance.

The curves superimposed on the bar plots indicate the inter-session leave-one-session-out cross-validation accuracy for classifying four emotions using only the first n stable features, with n varying from 1 to 255. As we can see, the curves exhibit similar trend among all subjects. The accuracy peaks at a small subset of stable features, then deteriorates when more and more unstable features are included into the feature subset being examined as n increases. For subject 2, 3, 4, 5, and 6, we can clearly see that the accuracy quickly deteriorates as features that carry negative stability scores are included into the feature subset being examined. This experiment shows the advantage of stable features over unstable features when the long-term performance is the utmost concern, and establishes the effectiveness of our proposed feature selection method. The peak recognition accuracy (peak of the accuracy curves in Fig. 3) and the number of stable features needed to achieve the peak performance is given in Table V. Comparing Table V with Table IV, we can see that stable features selected by our algorithm have outperformed nearly all referenced features. Comparing our features to the best-performing referenced features in Table IV (bold values), our features improve the accuracy by 3.60 %, 1.83 %, 5.86 %, 5.52 %, 2.72 %, and 2.14 %, for subject 1, 2, 3, 4, 5, and 6, respectively. Moreover, our selected features have a smaller dimension than the referenced state-of-the-art features, mitigating the burden of classifier training.

In addition, we observe that ICC value is in direct correlation with the long-term recognition performance, which validates our hypothesis that using stable features improves the accuracy. As can be seen from Fig. 3 (and also Table VII), the stability scores of the top stable features for subject 1 and subject 4 are notably higher than that for the other subjects. The long-term recognition performance of selected stable features of subject 1 and subject 4 are also notably higher than that of the other subjects. Generally, the higher the stability score, the better the recognition accuracy.

Looking at the subject-dependent feature ranking in Table VII, we can see that the feature ranking exhibits similar pattern among subject 1, 4, and 6. Statistic features top the stability ranking, together with Hjorth features and some HOCs.

TABLE VI COMPARISON OF INTER-SESSION LEAVE-ONE-SESSION-OUT CROSS-VALIDATION ACCURACY ON THE TEST DATA BETWEEN USING REFERENCED STATE-OF-THE-ART FEATURE SET AND STABLE FEATURE SET SELECTED BY OUR PROPOSED ALGORITHM. MEAN ACCURACY (%) ± STANDARD DEVIATION (%).

Feature	Subject					
	1	*2*	*3*	*4*	*5*	*6*
STAT	36.79 ± 6.04	26.80 ± 3.87	26.88 ± 3.97	38.68 ± 5.92	28.38 ± 4.06	31.29 ± 7.76
HOC	28.68 ± 3.11	24.51 ± 2.84	25.55 ± 3.87	28.62 ± 3.74	25.90 ± 2.67	27.23 ± 4.30
FD1	30.92 ± 3.58	24.64 ± 3.56	25.95 ± 4.43	35.51 ± 5.57	26.41 ± 2.87	29.99 ± 5.22
FD2	35.61 ± 5.47	26.44 ± 4.22	27.50 ± 3.57	40.54 ± 5.89	27.47 ± 3.49	31.82 ± 7.93
HJORTH	31.65 ± 5.86	26.62 ± 2.80	26.82 ± 3.15	38.47 ± 5.85	26.76 ± 2.84	29.64 ± 3.78
SE	26.28 ± 3.97	26.61 ± 5.40	26.64 ± 2.93	36.98 ± 8.46	28.89 ± 3.40	27.49 ± 5.36
POW	33.41 ± 7.11	**27.95** ± 3.66	28.04 ± 3.14	38.85 ± 8.02	27.65 ± 3.94	31.92 ± 7.68
Ours	**39.33** ± 6.13	26.52 ± 4.23	**28.27** ± 3.72	**43.66** ± 6.09	**30.81** ± 5.11	**33.54** ± 6.93

TABLE VII FEATURE RANKING OF THE TOP 10 STABLE FEATURES AND THEIR RESPECTIVE ICC SCORES.

Rank	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6	
	Feature	*Score*	*Feature*	*Score*	*Feature*	*Score*	*Feature*	*Score*	*Feature*	*Score*	*Feature*	*Score*
1	hoc1_T7	0.4921	beta_F4	0.2671	hoc9_FC5	0.2771	stat5_T7	0.7548	hoc2_FC5	0.1909	stat3_T7	0.3430
2	stat3_T7	0.4913	hoc31_T7	0.1751	alpha_F7	0.2630	stat3_T7	0.7443	hoc32_AF3	0.1570	stat5_T7	0.3331
3	stat5_T7	0.4302	hoc33_T7	0.1651	hoc10_FC5	0.2445	beta_T7	0.6914	hoc28_AF3	0.1469	beta_T7	0.2726
4	hoc2_T7	0.3557	hoc34_T7	0.1523	hoc11_FC5	0.2040	stat2_T7	0.6473	hoc29_AF3	0.1279	hoc1_T7	0.2030
5	mblty_T7	0.2547	hoc32_T7	0.1450	hoc3_T7	0.1973	se_T7	0.5098	hoc30_AF3	0.1194	stat2_T7	0.1872
6	stat4_T7	0.2057	hoc2_F4	0.1428	hoc4_T7	0.1911	actvt_T7	0.5098	mblty_F4	0.1086	mblty_T7	0.1544
7	cpxty_T7	0.1874	beta_F7	0.1413	hoc8_FC5	0.1607	hoc1_T7	0.4992	hoc8_F4	0.1048	stat4_T7	0.1438
8	hoc13_AF3	0.1815	beta_AF3	0.1269	mblty_F4	0.1439	hoc5_T7	0.4353	hoc19_AF3	0.1015	hoc25_T7	0.1334
9	hoc29_AF3	0.1749	hoc2_FC5	0.1249	alpha_AF3	0.1398	alpha_T7	0.4272	hoc33_AF3	0.1011	stat6_T7	0.1233
10	stat6_T7	0.1652	hoc35_T7	0.1173	stat4_F4	0.1385	hoc4_T7	0.4161	hoc34_F4	0.0980	hoc26_T7	0.1173

However, for subject 2, 3 and 5, different ranking patterns are observed. HOCs are found to be more stable, mixed with some power features and Hjorth features. Interestingly, HOC features have been frequently selected given their relatively high stability scores, despite their mediocre performance in Simulation 1 in Table III. It may suggest that HOC features exhibit good stability and are suitable for the use case where the long-term recognition performance shall be put into consideration. However, it is not the optimal features if re-calibration is allowed before using the BCI from time to time.

D. Comparison on the Test Data

We further examine the performance of the stable features on unseen test data comprising Segment 2 (see Fig. 2) of all available trials. To simulate the long-term recognition performance, the same inter-session leave-one-session-out cross-validation evaluation scheme is applied. The stable feature set remains the same as was found on the training data. The recognition accuracy using our proposed stable features as well as the referenced state-of-the-art features is presented in Table VI. The results are principally consistent with the findings based on training data set. Our stable features outperform the best-performing referenced features by 2.54 %, 0.23 %, 3.12 %, 1.92 %, and 1.62 %, for subject 1, 3, 4, 5, and 6, respectively.

E. Limitation

In this study, we have proposed and validated a stable feature selection method for EEG-based emotion recognition on a dataset comprising six subjects. Further studies are needed to conclude the performance on a larger dataset. We have taken a subject-dependent approach to finding the subject-specific stable features. Compared to our previous studies [34, 35] where we had taken a subject-independent approach, subject-specific stable features are found to be more effective. However, since the effective stable feature set is subject-dependent, to find which requires ample labeled affective EEG data recorded over a long course of time. The acquisition of such data may post a burden to the subjects. Although the stable features perform relatively better than the referenced state-of-the-art in the long run, the absolute recognition accuracy is still admittedly low. It remains an open question as to how we can effectively mitigate or even eliminate the need of frequent re-calibrations of the BCI.

V. CONCLUSION

An EEG-based affective BCI needs frequent re-calibrations as the affective neural patterns are volatile over the course of time even for the same subject, and intra-subject variance exist in the affective feature parameters. In this paper, we propose a stable feature selection method to select the optimal feature set that maximize the recognition accuracy for the long run of an affective BCI. The proposed method consists in modeling the feature stability by ICC, feature ranking and iterative selection of stable features. We hypothesize that unstable features contribute to the accuracy deterioration when the BCI operates without re-calibration over the course of time, and by using stable features, the recognition accuracy can be improved. We carry out extensive comparison between our stable features and the state-of-the-art features. In Simulation 1, we show the recognition accuracy of an affective BCI using the state-of-the-art features, where the BCI is allowed to be re-calibrated from time to time. In Simulation 2, we simulate the long-term usage of an affective BCI and establish that accuracy deterioration will occur when the BCI operates without re-calibration. In Simulation 3, we analyze the performance of stable features selected by our proposed method. We demonstrate the accuracy trajectory when we iteratively include features into the selected feature subset. Experimental results show that recognition accuracy peaks at a small subset of stable features, and as more unstable features are included, the recognition accuracy quickly deteriorates. The experiment results validate our hypothesis. Comparisons between our stable features and the referenced state-of-the-art features show that our stable features yield better accuracy than the best-performing referenced features by 1.83 % – 5.85 % on the training set, and by 0.23 % – 2.54 % on the test set.

Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[1] K. Ishino and M. Hagiwara, "A feeling estimation system using a simple electroencephalograph," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, vol. 5, pp. 4204-4209.

[2] K. Schaaff, "EEG-based Emotion Recognition," Diplomarbeit am Institut fur Algorithmen und Kognitive Systeme, Universitat Karlsruhe (TH), 2008.

[3] Y. P. Lin, C. H. Wang, T. L. Wu, S. K. Jeng, and J. H. Chen, "EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 489-492.

[4] K. Schaaff and T. Schultz, "Towards an EEG-based emotion recognizer for humanoid robots," in *IEEE International Workshop on Robot and Human Interactive Communication*, 2009, pp. 792-796.

[5] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies,* vol. 67, no. 8, pp. 607-627, Aug 2009.

[6] M. Li and B. Lu, "Emotion classification based on gamma-band EEG," in *IEEE International Conference on Engineering in Medicine and Biology Society*, 2009, pp. 1223-1226.

[7] Y. P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering,* vol. 57, no. 7, pp. 1798-1806, 2010.

[8] M. Murugappan, R. Nagarajan, and S. Yaacob, "Combining spatial filtering and wavelet transform for classifying human emotions using EEG Signals," *Journal of Medical and Biological Engineering,* vol. 31, no. 1, pp. 45-51, 2011.

[9] X.-W. Wang, D. Nie, and B.-L. Lu, "EEG-based emotion recognition using frequency domain features and support vector machines," in *Neural Information Processing*, 2011, pp. 734-743: Springer.

[10] Y. Liu and O. Sourina, "EEG Databases for Emotion Recognition," in *2013 Internation Conference on Cyberworlds*, Yokohama, 2013, pp. 302-309.

[11] M. Kwon, J.-S. Kang, and M. Lee, "Emotion classification in movie clips based on 3D fuzzy GIST and EEG signal analysis," in *International Winter Workshop on Brain-Computer Interface (BCI)*, 2013, pp. 67-68.

[12] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena,* vol. 31, no. 2, pp. 277-283, 1988.

[13] Y. Liu and O. Sourina, "Real-Time Fractal-Based Valence Level Recognition from EEG," *Transactions on Computational Science XVIII,* vol. 7848, pp. 101-120, 2013.

[14] Y. Liu, "EEG-based Emotion Recognition for Real-time Applications," Ph.D. Thesis, Nanyang Technological University, 2014.

[15] Y. Liu and O. Sourina, "Real-Time Subject-Dependent EEG-Based Emotion Recognition Algorithm," in *Transactions on Computational Science XXIII*: Springer, 2014, pp. 199-223.

[16] B. Kedem and E. Slud, "Time series discrimination by higher order crossings," *The Annals of Statistics,* pp. 786-794, 1982.

[17] P. C. Petrantonakis and L. J. Hadjileontiadis, "Adaptive Emotional Information Retrieval From EEG Signals in the Time-Frequency Domain," *IEEE Transactions on Signal Processing,* vol. 60, no. 5, pp. 2604-2616, 2012.

[18] P. C. Petrantonakis and L. J. Hadjileontiadis, "A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition," *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, no. 5, pp. 737-746, 2011.

[19] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine,* vol. 14, no. 2, pp. 186-197, 2010.

[20] F. Feradov and T. Ganchev, "Ranking of EEG time-domain features on the negative emotions recognition task," *Annual Journal of Electronics,* vol. 9, pp. 26-29, 2015.

[21] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and clinical neurophysiology,* vol. 29, no. 3, pp. 306-310, 1970.

[22] K. Ansari-Asl, G. Chanel, and T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood," in *15th European Signal Processing Conference*, 2007, pp. 1241-1245: IEEE.

[23] R. Horlings, D. Datcu, and L. J. Rothkrantz, "Emotion recognition using brain activity," in *9th international conference on computer systems and technologies and workshop for PhD students in computing*, 2008, pp. II. 1-6: ACM.

[24] J. J. Allen, H. L. Urry, S. K. Hitt, and J. A. Coan, "The stability of resting frontal electroencephalographic asymmetry in depression," *Psychophysiology,* vol. 41, no. 2, pp. 269-280, 2004.

[25] S. Gudmundsson, T. P. Runarsson, S. Sigurdsson, G. Eiriksdottir, and K. Johnsen, "Reliability of quantitative EEG features," *Clinical Neurophysiology,* vol. 118, no. 10, pp. 2162-2171, 2007.

[26] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological methods,* vol. 1, no. 1, pp. 30-46, 1996.

[27] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Transactions on Affective Computing,* vol. 3, no. 1, pp. 18-31, 2012.

[28] Emotiv. Available: http://www.emotiv.com

[29] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8," University of Florida, 2008.

[30] M. M. Bradley and P. J. Lang, "The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual," University of Florida, 2007.

[31] F. C. Pampel, *Logistic regression: A primer*. Sage Publications, 2000.

[32] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random: a closer look on BCI results," *International Journal of Bioelectromagnetism,* vol. 10, no. 1, pp. 52-55, 2008.

[33] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods,* vol. 250, pp. 126-136, 2015.

[34] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Stability of features in real-time EEG-based emotion recognition algorithm," in *2014 International Conference on Cyberworlds*, 2014, pp. 137-144.

[35] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Real-time EEG-based emotion monitoring using stable features," *The Visual Computer,* vol. 32, no. 3, pp. 347-358, 2016.

[36] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence,* vol. 23, no. 10, pp. 1175-1191, 2001.

[37] K. Takahashi and A. Tsukaguchi, "Remarks on emotion recognition from multi-modal bio-potential signals," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, vol. 2, pp. 1654-1659.

[38] Y. Liu and O. Sourina, "EEG-based Dominance Level Recognition for Emotion-enabled Interaction," in *IEEE International Conference on Multimedia and Expo*, Melbourne, 2012, pp. 1039-1044.