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

$$x_k^m = [x(m), x(m+k), ..., x(m+\frac{n-m}{k}k)], m = 1, 2, ..., k,$$

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.

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

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

DOI: 10.1109/CW.2018.00042
\[ FD = - \lim_{k \to \infty} \frac{\log(s_k)}{\log(k)} \]  

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:**

\[ \mu_x = \frac{1}{n} \sum_{i=1}^{n} x(i), \]  

(4)

**Standard deviation of the raw signals:**

\[ \sigma_x = \frac{1}{\sqrt{n-1}} \sum_{i=1}^{n} (x(i) - \mu_x)^2, \]  

(5)

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

\[ \delta_x = \frac{1}{n-1} \sum_{i=1}^{n-1} \left| x(i+1) - x(i) \right|, \]  

(6)

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

\[ \overline{\delta}_x = \frac{1}{n-2} \sum_{i=1}^{n-2} \left| \overline{x}(i+2) - \overline{x}(i) \right| = \frac{\delta_x}{\sigma_x}, \]  

(7)

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

\[ \gamma_x = \frac{1}{n-2} \sum_{i=1}^{n-2} \left| x(i+2) - x(i) \right|, \]  

(8)

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

\[ \overline{\gamma}_x = \frac{1}{n-2} \sum_{i=1}^{n-2} \left| \overline{x}(i+2) - \overline{x}(i) \right| = \frac{\gamma_x}{\sigma_x}, \]  

(9)

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

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 \( x = [x(1), x(2), ..., x(N)]^T \) to another series \( s = [s(1), s(2), ..., s(N)]^T \) in a frequency domain. \( s \) is computed as

\[ s(k) = \sum_{n=0}^{N-1} x(n)e^{-i\frac{2\pi kn}{N}}, \]  

(10)

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

\[ \hat{s}(k) = \frac{1}{N}|s(k)|^2, \]  

(11)

Lastly, the spectral band power features are computed by averaging the power spectrum density \( \hat{s}(k) \) over the targeted sub-band. E.g., the alpha band power is computed by averaging \( \hat{s}(k) \) 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 \( x \) as

\[ z(i) = x(i) - \mu_x, \]  

(12)

where \( z \) is the zero-meaned series of \( x \) and \( \mu_x \) the mean of \( x \) computed as per (4). Then, a sequence of filter \( \nabla \) is successively applied to \( z \), where \( \nabla \) is the backward difference operator, \( \nabla z(i) = z(i) - z(i-1) \). Denote the \( k \)-th order filtered sequence of \( z \) as \( \xi_k(z) \), \( \xi_k(z) \) is obtained by iteratively applying \( \nabla \) on \( z \), as

\[ \xi_k(z) = \nabla^{k-1}z, \nabla^0z = z. \]  

(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 \( \xi_k(z) \). We follow [15] and compute the HOC feature of order \( k = 1, 2, 3, ..., 36 \).

5) **Signal Energy**

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

\[ e = \sum_i |x(i)|^2. \]  

(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:**

\[ a(x) = \frac{1}{n} \sum_{i=1}^{n} (x(i) - \mu_x)^2; \]  

(15)

where \( \mu_x \) is the mean of \( x \) computed as per (4).

**Mobility:**

\[ m(x) = \sqrt{\frac{\text{var}(x)}{\text{var}(x)^2}} \]  

(16)

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

**Complexity:**

\[ c(x) = \frac{m(x)}{m(x)} \]  

(17)

which is the mobility of the time derivative of \( x \) over the mobility of \( x \).

B. **Feature Stability Measurement**

The stability of feature parameters was quantified by the Intraclass Correlation Coefficient (ICC). ICC allows for the
ICC value can drop below zero in the case when which is based on ANOVA. For clarity, we display considerations into account. The ICC is computed as per (18), dissimilar to each other. The ICC measurement takes both distinguishable. Therefore, we want different rows to be states to be discriminative, so that different affective states are parameters within the same row to be similar to each other. parameters to be consistent when measuring the same emotion over the course of time. Intuitively, we want the feature emotions, the sum of all measurements of emotion the average of all measurements of emotion

\[
\begin{align*}
\text{Within treatment} & : SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 \\
\text{Between treatment} & : SS_B = SS_T - SS_W \\
\text{Total} & : SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x})^2
\end{align*}
\]

where \( MS_W \) 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 \( X \) be the matrix of feature parameters of a specific kind of feature, rows of \( X \) correspond to different emotions, and columns of \( X \) 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 \( X \) 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 \( i \). \( x_i \) is the sum of all measurements of emotion \( i \), \( x_i = \sum_{j=1}^{k} x_{ij} \). \( \bar{x}_i \) is the average of all measurements of emotion \( i \), \( \bar{x}_i = (1/k) \sum_{j=1}^{k} x_{ij} \). \( x \) is the sum of all measurements over all emotions, \( \bar{x} = (1/nk) \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} \).

**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 \( 4 \times 2 \times 8 = 64 \) 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

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

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 $7 \times 4 = 28$ samples for four emotions, totally. Likewise, the test data consist of 28 samples for each emotion. 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 $7 \times 4 = 28$ samples for four emotions recorded in the same session. The test data comprise $7 \times 4 \times 15 = 420$ samples pooled together from the remaining 15 sessions. The mean classification accuracy over 16 runs and the standard deviations are displayed in Table IV.

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.

DOI: 10.1109/CW.2018.00042
TABLE III FOUR-EMOTION RECOGNITION ACCURACY OF SIMULATION 1, MEAN ACCURACY (%) ± STANDARD DEVIATION (%)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAT</td>
<td>56.81 ± 10.52</td>
<td>44.75 ± 16.66</td>
<td>43.64 ± 13.89</td>
<td>71.43 ± 14.32</td>
<td>47.92 ± 15.44</td>
<td>73.88 ± 15.29</td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>32.25 ± 10.50</td>
<td>30.25 ± 10.05</td>
<td>28.46 ± 10.24</td>
<td>43.53 ± 12.20</td>
<td>28.37 ± 10.95</td>
<td>36.61 ± 12.29</td>
<td></td>
</tr>
<tr>
<td>FD1</td>
<td>43.08 ± 13.98</td>
<td>37.39 ± 12.58</td>
<td>33.59 ± 8.12</td>
<td>58.59 ± 13.40</td>
<td>39.58 ± 12.05</td>
<td>54.58 ± 11.03</td>
<td></td>
</tr>
<tr>
<td>FD2</td>
<td>57.14 ± 9.93</td>
<td>46.88 ± 17.25</td>
<td>45.76 ± 13.01</td>
<td>72.54 ± 14.49</td>
<td>48.91 ± 15.42</td>
<td>76.23 ± 15.51</td>
<td></td>
</tr>
<tr>
<td>HJORTH</td>
<td>53.24 ± 11.81</td>
<td>46.65 ± 14.30</td>
<td>41.41 ± 14.39</td>
<td>77.27 ± 17.82</td>
<td>47.92 ± 15.67</td>
<td>72.54 ± 18.78</td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>45.54 ± 15.95</td>
<td>40.63 ± 12.67</td>
<td>41.96 ± 17.57</td>
<td>59.49 ± 16.23</td>
<td>41.96 ± 18.90</td>
<td>62.83 ± 20.02</td>
<td></td>
</tr>
<tr>
<td>POW</td>
<td>48.66 ± 12.21</td>
<td>46.88 ± 17.72</td>
<td>36.05 ± 14.70</td>
<td>69.20 ± 15.83</td>
<td>42.26 ± 18.03</td>
<td>62.72 ± 16.00</td>
<td></td>
</tr>
<tr>
<td>Upp Chan Lvl</td>
<td>42.79</td>
<td>42.80</td>
<td>42.79</td>
<td>39.36</td>
<td>42.70</td>
<td>42.79</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Selected Stable Feature</td>
<td>41.55 ± 4.31 (2)</td>
<td>30.24 ± 5.14 (7)</td>
<td>33.87 ± 3.55 (5)</td>
<td>45.22 ± 4.57 (1)</td>
<td>30.68 ± 3.43 (42)</td>
<td>33.63 ± 7.99 (34)</td>
<td></td>
</tr>
</tbody>
</table>

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
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, $1 \leq n \leq 255$. 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. Static features top the stability ranking, together with Hjorth features and some HOCs.
VI. The results are principally consistent with the findings based on the mean accuracy of 1.92%, and 1.62%, for subject 1, 3, 4, 5, and 6, respectively. Performing referenced features by 2.54%, 0.23%, 3.12%, 2.99%, and 5.22%, respectively.

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. 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 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 peaks at a small subset of stable features, and as more features are selected by our proposed method. We demonstrate 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 deterioration 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 deterioration 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 deterioration when the BCI operates without re-calibration.
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


