Emotion classification using linear predictive features on wavelet-decomposed EEG data*

Luka Kraljević, *Student Member, IEEE,* Mladen Russo, *Member, IEEE*, and Marjan Sikora

***Abstract*— Emotions play a significant role in human communication and decision making. In order to bypass current limitations of human-computer interaction, more natural, trustworthy and nonverbal way of communication is needed. This required machines to be able to explain and perceive person's emotions. Our work is based on the concept that each emotional state can be placed on a two-dimensional plane with arousal and valence as the axes. We propose a new feature set based on using the linear predictive coefficients on wavelet-decomposed EEG signals. Emotion classification is then performed using support vector machine with Gaussian kernel. Proposed approach is evaluated on EEG signals from publicly available DEAP dataset and results show that our method is effective and outperforms some state of the art methods.**

I. Introduction

Considering the increased use of machines in everyday life, intelligent interaction between humans and machines has become a priority issue. Emotions play a significant role in human communication and decision making. In order to bypass current limitations of human-computer interaction (HCI) more natural, trustworthy and nonverbal way of communication is needed. Empowering computer systems with the ability to recognize and respond to human emotions automatically is the key to intelligent HCI. To properly react to a person's emotional state, machines have to be able to gather information about the current situation and also to be equipped with measures to explain and perceive emotions of this person.

Many approaches for estimating human emotions have been proposed in the past few decades. The conventional approaches focus on analysis of visual and auditory signals such as speech and facial expressions. However, these methods might be more prone to deception, because it is easy to fake facial expression or change tone of voice. The more reliable way is to use physiological signals which compared to visual and auditory signals are continues and are hard to conceal, such as Galvanic Skin Response (GSR), Electrocardiogram (ECG), Skin Temperature(ST) and Electroencephalogram (EEG). Therefore, physiological signals offer a great potential for unbiased emotion recognition.

There are several models of emotions such as six basic emotions used mostly in facial expression recognition [1]. A good model of emotions can also be characterized by two main dimensions called valence and arousal [2]. The concept is that each emotional state can be placed on a two-dimensional plane with arousal and valence as the axes. The valence emotion ranges from negative to positive, whereas the arousal emotion ranges from calm to excited. The valence-arousal dimensional model, represented in Fig. 1, is widely used in many research studies.

In this paper, in order to classify emotions in the valence-arousal space, we propose the use of linear predictive features on wavelet-decomposed EEG data. Emotion classification is performed using SVM (Support Vector Machine) with Gaussian kernel function.

II. Related Work

Using EEG for detecting emotional state is relatively new in comparison to audio-visual methods although there are many researches in the emotion prediction using EEG signals. Haag et al. [3] were able to achieve a recognition rate of 96.58% on arousal ratings (high - low) while valence ratings were classified correctly with 89.93%. The data for the experiment was gathered from a single subject on different days and different times of the day. For classification, they used a neural network. Y. P. Lin et al., [4] reported averaged classification accuracy of 82.29% ± 3.06% of 26 subjects. They identified 30 subject-independent features that were most relevant to emotional processing across subjects. SVM was used as a classifier. Chanel et al. [5] reported an average accuracy of 63% by using EEG time-frequency information like features and SVM as a classifier to characterize EEG signals into three emotional states. Picard et al. [6] were able to differ between eight emotion categories with an accuracy of 81% using EEG data, blood volume pressure, skin conductance, and respiration information. Data was recorded from one person during some weeks. Unfortunately, it is difficult to make real comparisons between these researches because they differ on several criterions. A higher number of the subjects makes research more significant. Accuracy interpretation may differ in terms of the model obtained for emotion identification is user-specific or not. The difference in trial duration is also one of the obstacles as well as different emotion model [1][2]. There is also variety of stimuli used for emotion elicitation such as self-eliciting, recalling, picture, sound and video. There already exist stimuli databases that have been designed for the purpose of emotion elicitation such as the International Affective Picture or Digitized Sound System (IAPS, IADS)[7,8].

In 2012, Koelstra et al. [9] presented publicly available dataset (DEAP - dataset for emotion analysis using EEG,

* This work has been fully supported by the Croatian Science Foundation under the project UIP-2014-09-3875.

L. Kraljević, M Russo, M. Sikora are with the University of Split, FESB, Laboratory for Smart Environment Technologies, HR-21000 Split, Croatia (e-mails: {lkraljev, mrusso, sikora}@fesb.hr).

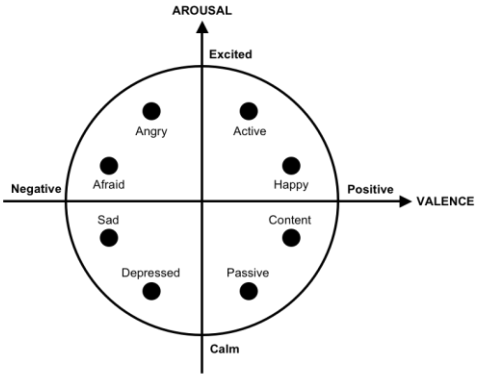

Figure 1. Valence-arousal dimensional model.

physiological and video signals) which is considered as de facto standard dataset nowadays. The authors showed that valence, arousal and liking assessment of emotion could be obtained with an accuracy of 57.6%, 62% and 55.4% for single trial classification. The classification was performed using the Naive Bayes classifier. Spectral Power Asymmetry (ASM) and Power Spectral Density (PSD) were used as input features. Wichakam and Vateekul [10] generated and tested many subsets of DEAP. The features extracted where PSD, wavelet transform, average band power and statistics values (max, min, average, standard deviation, and range). Experimental results showed high correct classification rates using only 10 channels. Accuracy for both arousal and valence where 64.9% and 67.3% for liking in single trial classification. Accuracy in subject independent test where 51.1% for valance, 52.9% for arousal and 67% for liking. Bahari et al. [11] applied Recurrence Quantification Analysis (RQA) to extract non-linear features and provide them as input to KNN (K-Nearest Neighbor) classifier. The authors reported the best accuracies when a random selection of feature vectors is used as feature selection with mean accuracies of 58.05%, 64.56%, and 67.42%, for valence, arousal and liking, respectively. The authors also reported results for case of user independent classification over all subjects as accuracy of 59.4%, 65% and 61.8% for three classes of valence, arousal and liking, although performing the test by leaving out only 25% of user's trials could hardly call fair. For evaluation authors used DEAP dataset. Jirayucharoensak et al. [12] performed experiments thresholding DEAP dataset in 3 classes for valence and 3 classes for arousal. The best average accuracies obtained for leave-one-subject-out classification were 53.42% ± 9.64% for valence and 52.03% ± 9.74% for arousal. Torres-Valencia et al. [13] propose the use of generative models as Hidden Markov Models (HMM) instead of feature extraction over DEAP dataset. They reported accuracy of 55%±4.5 for arousal and 58.75±3.8 for valence. The authors tested HMM using cross validation by setting the 80% of the signals as training and leaving the other 20% as test. Statistical validation is obtained by repeating procedure 10 times.

II. Proposed Approach

In order to obtain results which can be compared to others, in this paper we also use EEG data from the DEAP dataset.

We propose a new feature set based on using the linear predictive coefficients (LPC) on wavelet-decomposed EEG signals. Emotion classification performance is evaluated using SVM with Gaussian kernel.

A. DEAP Dataset

DEAP is an open source affective computing database containing multi-modal emotion data. The EEG and peripheral physiological signals of 32 subjects were recorded as each subject watched 40 preselected music videos. After watching each music video, the subjects reported affective states with self-assessment manikins (SAM) [14] of their levels of arousal, valence, dominance, and liking. All the rating scores are in the range from 1 to 9. EEG was recorded with a sampling frequency of 512Hz using 32 electrodes placed according to the international 10-20 system [15] as shown in Fig 2.

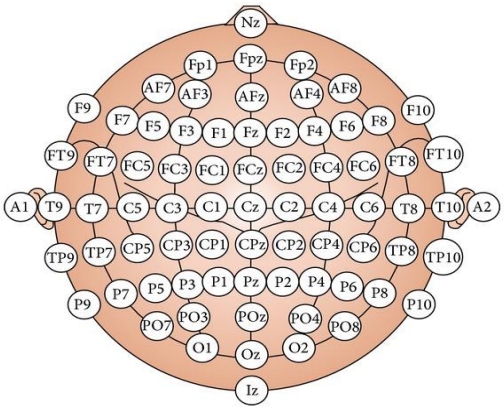

Figure 2. International 10-20 system of electrode placed.

There is also publicly available pre-processed version of DEEP dataset which was used as benchmark data in this research. This version of data is well suited to those who doesn't have necessary measuring equipment or to those who just wish to test a classification or regression technique and compare performance with other research who used this dataset. The full set of signals stored in DEAP database are the EEG, EMG, EOG, GSR, Temperature, Respiration pattern, plethysmography (HR), audio and video signals. In this work, we are interested in EEG based emotion classification and we are only using signals from the EEG subset. According to DEAP authors [9], row EEG data where first down-sampled to 128 Hz for computational efficiency. Bandpass frequency filter (4-45 Hz) where applied. EOG artefacts were removed by Blind Source Separation (BSS). EEG data was averaged to the common reference and segmented into 60 seconds trials and a 3 second pre-trial for base line removal. Pre-processed data comes as a matrix with dimensions (40*32*8064) (trial*channel *data) together with 1280*4 class labels.

C. Binary Class Conversion

Physiological responses in DEAP dataset are labelled as arousal and valence in scale from 1 to 9. In this work, those scales have been thresholded to form binary classification problem. Emotions rated above five have been classified as high otherwise as low, thus forming valance-arousal space (LALV, HALV, LAHV, HAHV). Class distribution of each emotion state is shown in Table I.

TABLE I. CLASS DISTRIBUTIONS

	Positive	**Negative**
Valence	724	556
Arousal	754	526

Table I. data clearly shows how unbalanced the classes are, and this should be taken into account when presenting classification performance.

B. Feature Extraction

Feature Extraction plays an important role in obtaining useful information from the signal. A key to successful feature extraction is to choose the features or information which are the most important for classification.

It is impractical to use all spectral samples as features so the spectrum is usually divided into bands (typically alpha, beta, gamma, delta, theta) and the mean value (or some other statistical property) for each band is used in the feature set.

Linear prediction is a well known and widely used method in speech processing. It can be used for spectral envelope estimation and our idea is to represent the spectrum in each frequency band (alpha, beta...) by several LPC coefficients. In this way, the spectral shape is preserved in more or less detail depending on the number of coefficients.

In the first step, we use Discrete Wavelet Transform (DWT) to decompose EEG signal into 5 frequency bands, as shown in Table II. We choose *db3* as mother wavelet. In contrast to Fourier transform which is known to lose the time information of a signal when converting the signal into frequency domain, DWT offers simultaneous localization in time and frequency.

TABLE II. EEG SIGNAL DECOMPOSITION

Frequency band	**Frequency range (Hz)**	**Frequency bandwidth (Hz)**	**Decomposition level**
Delta	0 - 4	4	D5
Theta	4 - 8	4	D4
Alpha	8 - 16	8	D3
Beta	16 - 32	16	D2
Gamma	32 - 64	32	D1

The main idea behind LP is that signal from one channel can be estimated based on a linear combination of the previous samples:

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-1) \qquad (1)$$

where $\hat{x}(n)$ is the predicted signal , $x(n-1)$ the previous observed sample, a_i the prediction coefficients and p is the order of prediction. The prediction error $e(n)$ can be computed by the difference between actual signal $x(n-1)$ and the predicted signal $\hat{x}(n)$ which is given by

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^{p} a_i x(n-1) \quad (2)$$

The primary objective of LP analysis is to compute the LP coefficients a_i which minimized the prediction error $e(n)$. The LPCs are obtained using Levinson-Durbin recursive algorithm.

Feature vectors are generated from each of 32 EEG channels including Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, and O2. The number of LPC coefficients should be chosen so the reconstructed spectrum can be as different as possible for different emotional states and as similar as possible for the same emotional states. We experimented with several numbers and obtained the best results when the number of coefficients was 5.

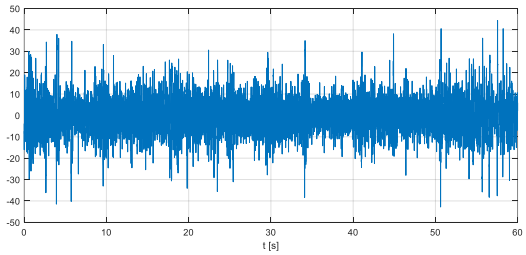

Figure 3. An example of EEG signal from DEAP dataset.

Fig. 3. shows an example of EEG signal from DEAP dataset (one trial) and Fig. 4. shows its original spectrum (top plot), spectrum decomposed in five frequency bands by DWT (middle plot) and spectral envelope represented by 5 LPC coefficients (bottom plot).

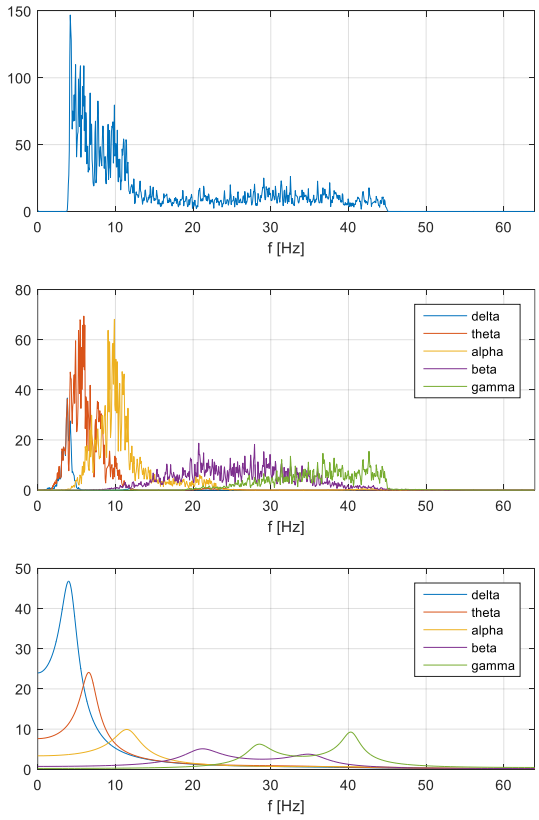

Figure 4. Sample EEG signal decomposition into frequency bands and LPC reconstructed spectrum.

Since the LP order of prediction was 5, total number of features was 800 (5 LPCs x 5 frequency bands x 32 channels).

C. Feature Reduction

Features selection is an important step in pattern classification since there is a difficulty in measuring classification information in all features. It can not only improve learning efficiency, but can also improve prediction performance. The objective of feature selection is to extract a subset of features by removing redundant features. In this paper we used F-score, a simple technique which measures the discrimination between different EEG patterns. F-score of the *i*-th feature is defined as in [14]:

$$F(i) = \frac{(\bar{X}_i^{(+)} - \bar{X}_i)^2 + (\bar{X}_i^{(-)} - \bar{X}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{X}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{X}_i^{(-)})^2} \quad (3)$$

where $\bar{X}_i$, $\bar{X}_i^{(+)}$and $\bar{X}_i^{(-)}$ are average *i*th feature of the whole, positive and negative data set. $x_{k,i}^{(+)}$ is *i*th feature of *k*th positive instance and $x_{k,i}^{(-)}$ is *i*th feature of *k*th negative instance. Number of positive and negative instances are n_+ and n_-, respectively. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets.

We used F-score to sort features by their discrimination power (larger the F-score, greater discrimination power). A key to obtain optimal number of features lies in defining optimal threshold (retain features above threshold). This is achieved by iteratively conducting classification by changing threshold i.e., removing a subset of features based on F-score rank.

D. Classification.

Support Vector Machine (SVM) is a learning algorithm for pattern classification and regression. SVM is known to have good generalization properties and to be insensitive to overtraining and the curse of dimensionality e.g. number of features exceeds the number of training examples. The basic idea of SVM is finding the optimal hyperplane where the expected classification error of test samples is minimized. Optimal allocation of the hyperplane between the boundary of the two classes helps to achieve the maximum discrimination thus increasing the generalization capability.

In the proposed approach, SVM with Gaussian kernel function is employed as the classifier to perform two class classification, such as high and low, for emotional dimensions of valence and arousal. To correctly evaluate classification performance, test data must never be used in the training dataset, so we divided the complete dataset into two categories: training data and test data. Test data was randomly selected as 10% of each user's trials (4 trials from each user for test, 36 trials from each user for training). Considering 32 users in total, there were 128 trials in the test set and 1152 trials in the training set. Training was performed using the standard 10-fold cross-validation protocol in an effort to minimize the potential of overfitting. The whole process was automatically repeated and results were averaged.

We also performed a Leave One Subject Out (LOSO) experiment. In this case, training dataset was a composition of all input features from 31 subjects while testing was performed using data samples of the one remaining subject. Training was performed using the standard 10-fold cross-validation protocol. The experiment was then repeated for all subjects.

III. Results and Performance Comparison

In this section, we present the performance of the proposed methods using DEAP database. For performance comparison, Accuracy is a common measure used in literature. Given the number of True Positive - TP, number of True Negative - TN, number of false positive - FP and number of False Negative - FN, accuracy can be calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (4)$$

Table III summarizes the comparison of accuracies between the proposed method and other methods. Results show that the proposed methods performs well in comparison with other state of the art methods.

TABLE III. Accuracy Comparison for Subject Dependant Tests

	Valence	**Arousal**
Koelstra [9]	57.6	62.00
Wichakam [10]	64.9	64.9
Bahari [11]	58.05	64.56
Torres-Valencia [13]	58.75	55.00
Proposed method	62.0	62.67

Results obtained in the second experiment (LOSO) are quite lower and all authors do not present them. Table IV summarizes the comparison of accuracies between the proposed method and other methods.

TABLE IV. Accuracy Comparison for Subject Independent Tests

	Valence	**Arousal**
Jirayucharoensak [12]	53.42	52.03
Wichakam [10]	51.1	52.9
Proposed method	52.84	56.87

Results show that the proposed methods outperforms the method in [10], and compared to [12], our method is a bit lower for valence classification but higher for arousal classification.

As a measure of performance, accuracy can be misleading when there is a large class imbalance. Another measure of test accuracy is F_1 score which takes the class balance into account (see Table I.). It considers both precision and recall of the test to compute the score. F_1 score is defined as:

$$F_1 = 2 \times \frac{PR \times RC}{PR+RC} \quad (5)$$

where PR stands for precision and RC stands for recall. Precision is the number of True Positives divided by the number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV). Recall is the number of True

Positives divided by the number of True Positives and the number of False Negatives, e.g. it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.

In the proposed approach, F_1 score obtained for subject dependent tests was 68,45% and 70.97% for valence and arousal, respectively. F_1 score was also calculated in [9], were the authors obtained 56.3% and 58.3% for valence and arousal, respectively. Result show that our method clearly outperforms the method in [9]. We also calculated the F_1 score for subject independent tests (LOSO experiment) and obtained 60.71% and 65.6% for valence and arousal, respectively.

In order to improve learning efficiency, and possibly improve prediction performance, we also experimented with feature reduction methods. We used F-score measure for discrimination between different EEG patterns – eq. (3). We didn't observe an improvement in performance but we were able to reduce the number of features from 800 to 200 with practically no loss in accuracy performance – valence 61.99%, arousal 62.14%, F_1 score – valence 68.03%, arousal 70.52.

Feature reduction by F-score was also compared with PCA (Principal Component Analysis) feature reduction. For the 95% explained variance criterion, only 60 features were kept and obtained accuracy score was 60.59% for valence and 62.45% for arousal, and F_1 score was 67.16% for valence and 70.48% for arousal. Compared to F-score, results were quite similar - accuracy score was 60.09% for valence and 61.21% for arousal, and F_1 score was 67.05% for valence and 69.51% for arousal.

IV. Conclusion

In this work, we proposed an automatic emotion classification system based on EEG data. A new feature set was proposed based on using the linear predictive coefficients on wavelet-decomposed EEG signals. Support vector machine learning algorithm was applied to obtain the optimal boundary between the classes, and results were analyzed in terms of accuracy and F_1 scores. Experimental results on DEAP database show that the proposed approach is effective and mostly outperforms other state of the art methods.

Although the results are good compared to others, they are generally quite low and much improvement is required in order to use this technology in practice. In future, we will continue our research in direction of developing new and improved features.

References

[1] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 712–717, Oct. 1987.

[2] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[3] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system." Lecture Notes in *Computer Science*, vol. 3068, pp. 33–48, 2004.

[4] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, J.-H. Chen, "EEG-Based Emotion Recognition in Music Listening" in *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798-1806, July 2010.

[5] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *Int. J. Human-Comput. Stud.*, vol. 67, no. 8, pp. 607–627, Aug. 2009.

[6] 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, pp. 1175 – 1191, Oct 2001.

[7] M. M. Bradley, P. J. Lang, and B. N. Cuthbert, International Affective Picture System (IAPS): Digitized Photographs, Instruction Manual and Affective Ratings, University of Florida, Gainesville, Fla, USA, 2005.

[8] M. M. Bradley and P. J. Lang, The International Affective Digitized Sounds (IADS-2): Affective Ratings of Sounds and Instruction Manual, University of Florida, Gainesville, Fla, USA, 2nd edition, 2007.

[9] S. Koelstra, C. Muhl, M. Soleymani et al., "DEAP: a database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[10] I. Wichakam, P. Vateekul, "An evaluation of feature extraction in eeg-based emotion prediction with support vector machines", *the International Conference on Computer Science and Software Engineering (JCSSE)*, pp. 106-110, 2014.

[11] F. Bahari, A. Janghorbani, "Eeg-based emotion recognition using recurrence plot analysis and k nearest neighbor classifier", *Proceedings of 20th Iranian Conference on Biomedical Engineering (ICBME 2013)*, pp. 228-233, 2013.

[12] S. Jirayucharoensak, S. Pan-Ngum, P. Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaption", *The Scientific World Journal*, vol. 2014, pp. 1-10, 2014.

[13] Torres-Valencia et al., "Comparative analysis of physiological signals and electroencephalogram (eeg) for multimodal emotion recognition using generative models," in *IEEE Symposium on Image, Signal Processing and Artificial Vision*, 2014, pp. 1–5.

[14] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential." *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[15] F. Sharbrough, G. E.Chatrian, R. P. Lesser, H. Luders, M.Nuwer, and T. W. Picton, "American Electroencephalographic Society guidelines for standard electrode position nomenclature," *Journal of Clinical Neurophysiology*, vol. 8, no. 2, pp. 200–202, 1991.