Predicting Text-To-Speech Quality using Brain Activity

Rhenaldy
Computer Science Department
Bina Nusantara University
Jakarta 11480, Indonesia
rhenaldy@binus.ac.id

Ladysa Stella Karenza
Computer Science Department
Bina Nusantara University
Jakarta 11480, Indonesia
ladysa.karenza@binus.ac.id

Ivan Halim Parmonangan
Computer Science Department
Bina Nusantara University
Jakarta 11480, Indonesia
ivan.parmonangan@binus.edu

Felix Indra Kurniadi
Computer Science Department
Bina Nusantara University
Jakarta 11480, Indonesia
felix.kurniadi@binus.edu

***Abstract*—With the importance of audio quality in developing a text-to-speech systems, it is important to conduct audio quality evaluation. Various methods have been developed in order to conduct audio quality evaluation, which are done either subjectively or objectively. Subjective methods require a large amount of time and resources, while objective methods lack human influence factors, which are important to a user's perception of the experience offered by the audio quality. These human influence factors manifest inside an individual's brain in forms such as electroencephalograph (EEG). In this study, we performed audio quality predictions using EEG data, which resulted in the proposed model yielding lower error distribution compared to other methods.**

Keywords—text-to-speech, EEG, audio quality, SVR, MLP, Decision Tree Regressor

I. INTRODUCTION

Text-To-Speech (TTS) systems are widely implemented within various technologies, such as smartphones, desktops, and other electronic devices. Thus, it is crucial to develop a TTS system with the utmost quality to reach the product's market success. Factors such as audio quality and customer's overall satisfaction towards the audio quality can heavily determine the success of the TTS system. In order to develop a TTS system with favorable audio quality, the International Telecommunication Union (ITU) recommends the conduction of multi-dimensional subjective assessment through listening tests. These listening tests consist of scoring based on various scales or dimensions such as overall impression, voice pleasantness, and more [1].

In conducting such tests, it requires a large amount of time and resources considering the need of human volunteers participating in the tests. Such reasons urge the development of human replacement in estimating audio quality. Objective or instrumental models such as P.563 [2], ANIQUE+ [3], and HMMs [4] are example of previous efforts in developing an objective approach that is able to replace the need for human listener. The downside to these objective approaches is that they lack human influence factors (HIFs) [5], considering that an audio quality's predicted objective score may contradict with the expectation of the human listener.

Hence, subjective approaches that involves human listener tends to be favored more with tests conducted in various different forms or shape, such as subjective ratings [1, 6], speech signal analysis [7, 8], and the physiological responses of a human listener towards speech [9, 10, 11].

In conducting audio quality prediction tasks, a previous work [12] predicted the mean opinion score (MOS) of an audio using a Convolutional Neural Network (CNN) based architecture by combining both EEG and the audio itself. Independent training of audio and EEG models was also conducted, which resulted in significant reduction of the error distribution. Another study also [13] predicted audio quality using Partial Least Squares (PLS) on the subject's EEG records. The proposed model predicts the subjective scores, which includes the overall impression, valence, and arousal, by conducting tensor structured features.

Another study [14] also used a CNN-based architecture to decode human brain activity with grayscale images as the stimuli. The proposed model uses multichannel EEG time-series as the input data, while the results are compared with other feature extraction and prediction techniques. There are also previous studies [15] that uses a similair architecture in detecting Parkinson's disease. The proposed model uses a resting-state EEG, with the results showing that the proposed model is able to efficiently and accurately detect Parkinson's disease.

In this study, we conducted the following tasks:

1. Comparing the results between models with different preprocessing methods.
2. Training and testing the proposed models using non-augmented data.
3. Training the proposed models using augmented data, while testing was done using non-augmented data.
4. Utilizing the EEG features to predict mean opinion scores.

II. DATASET

The dataset used in this study is the same dataset used by previous work [12], the PhySyQX dataset. This dataset consists of subjective audio ratings, individual opinion scores of twelve subjective dimensions, and brain activity records which was processed using electroencephalography and functional near-infrared spectroscopy [10]. In performing the tasks in this study, we utilized the EEG spectograms and the overall opinion scores. This is due to the baseline method only using EEG features as the input, while the overall impression opinion score is used as the output.

A. Participants

A number of eight female and thirteen male fluent English speakers (average age of 23.8 (±4.35) years) with no records of hearing disorders were recruited. The speech stimuli were presented through insert earphones to the twenty-one participants at their own preferred volume levels. The protocol was approved by the INRS Research Ethics Office, while the participants consented to releasing the data online with the data de-identified and were compensated for their time in participating.

B. EEG Records

The EEG Recording was done by fitting the participants with a compatible fNIRS-EEG cap, which has 62 EEG electrodes. In this study, we only use the raw EEG data, which are then processed using the Biosemi ActiveTwo system with the sampling rate of 512 Hz without online filtering, followed by down-sampling at 256 Hz. The raw EEG data was then pre-processed using EEGLAB [16], referenced as 'Cz' followed by band-pass filtering ranging from 0.5 to 50Hz using the FIR filter provided by EEGLAB. Then, independent component analysis (ICA) was applied to the EEG data in order to remove eye blink artifacts, while ADJUST toolbox [17] was implemented for semi-automatic rejection of noisy components.

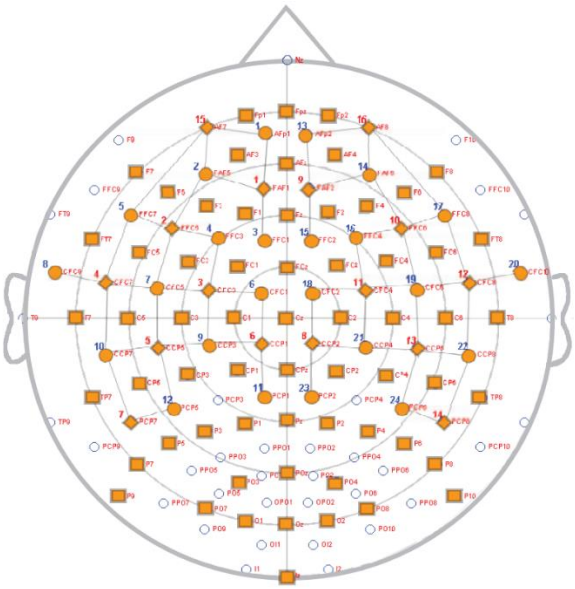

Figure 1: *Topology of the fNIRS-EEG cap, with the EEG electrodes represented by rectangles. The circle and diamonds respectively represent the fNIRS detectors and sources [10].*

C. Scores

The dataset also provided the overall impression opinion scores that ranges from one to five, with one as bad to five as excellent. However, we will be using mean opinion scores (MOS) to train and test our models rather than using opinion scores (OS). This is due to the findings of a previous work that showed training using MOS yields a lower error distribution per subject rather than OS [12]. Hence, we will be training and testing our models using MOS.

III. Methods

In this study, we used Convolutional Neural Network as the baseline that only used the EEG features and compare it to the proposed model. The study was done in a subject-dependent manner, by training the EEG models for each subject. Evaluation was done using the root-mean-square error (RMSE) metric, which is formulated as below with N as the number of test samples, $\hat{y}_i$ as the i-th predicted value of the test data, and y_i as the actual value.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad (1)$$

A. Baseline Model

The baseline used in this study is the result of a previous work [12] that uses a Convolutional Neural Network (CNN) based architecture trained using the same dataset and preprocessing method. The baseline model predicted both individual opinion scores and mean opinion scores. The previous study trained three different models, audio only, EEG only, and a fusion of audio and EEG. In this study, we will be comparing the results of the CNN-based model that only utilized the EEG features.

B. Fast Fourier transform

In this study, we will be extracting raw EEG signal and augmented raw EEG signal. We augmented each raw signal 30 times using Gaussian noise.

In order to match the process in the previous study [12], the spectrogram of each channel was extracted using fast Fourier-transform (FFT) with a 1-second window and an 0.5 second overlap, followed by normalization with the highest frequency limited to 45 Hz. This was done for each EEG sample.

C. Peak Picking Technique

Alongside fast Fourier transform, this study also implemented peak picking technique toward the raw EEG signal. Peak picking was implemented with a minimal horizontal distance of 18 in samples between neighboring peaks.

After the implementation of peak picking technique, each processed EEG signal is resampled into 256 samples, followed by standardization of data through centering and scaling.

D. Model Construction

The models implemented in this research consist of Support Vector Regression (SVR), Multi-layer Perceptron Regressor (MLP), and Decision Tree Regressor. Our SVR model used the radial basis function (rbf) kernel with the regularization of 5. Our MLP model consists of two hidden layer with the size of 128 neurons. While our Decision Tree Regressor uses the default parameter values set by scikit-learn.

Two EEG models of each subject were trained independently using nested cross-validation by dividing the dataset into 4 different sets of data. From the four different sets, we tested every possible combination of train-test data with the train data using 3 sets and the remaining one set as the test data. It resulted in 4 different models per subject, with the RMSE of each model averaged.

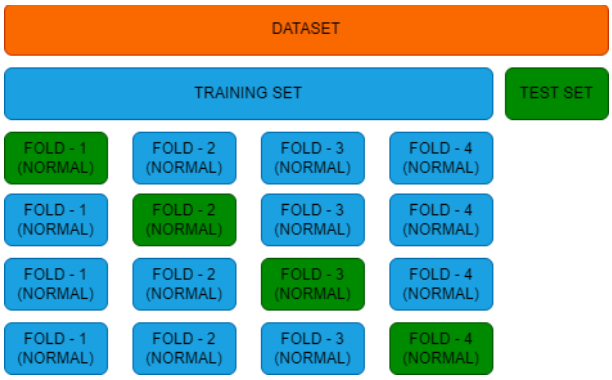

Figure 2: *Cross validation for models trained using non-augmented data.*

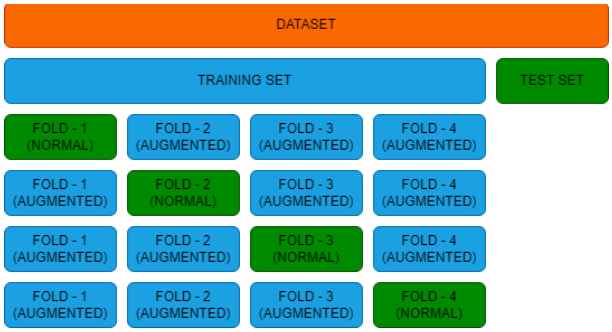

Figure 3: *Cross validation for models trained using augmented data.*

IV. Results

When comparing the average result of a model, it can often tell which model performed better. But, it often hinder the significance of a model. Thus, Wilcoxon signed-rank test (α=0.01,T=42,N=21) will be conducted in order to determine the significance of each model's performance compared to other models.

Table 1 shows a side-by-side comparison of the MOS prediction RMSE results per subject from models using both fast Fourier transform and peak picking technique as the data preprocessing method, with non-augmented data as training and testing data.

Sbj.	fast Fourier transform			Peak Picking Technique		
	SVR	MLP	DTR	SVR	MLP	DTR
1	0.985	8.791	1.312	0.963	3.258	1.413
2	0.970	6.515	1.498	0.958	4.012	1.372
3	0.979	6.656	1.255	0.925	2.855	1.695
4	1.001	10.098	1.300	0.961	3.111	1.230
5	0.955	8.879	1.582	1.036	4.589	1.300
6	0.978	9.741	1.435	1.030	3.618	1.306
7	0.924	6.848	1.414	0.987	2.678	1.319
8	0.945	9.617	1.318	1.001	4.491	1.572
9	0.999	8.287	1.327	0.980	3.731	1.438
10	0.981	5.175	1.279	0.999	4.394	1.354
11	1.000	7.694	1.387	0.962	3.842	1.389
12	0.947	8.874	1.326	0.912	4.724	1.421
13	0.985	4.589	1.474	1.021	3.588	1.578
14	0.989	8.668	1.177	1.019	2.858	1.347
15	0.979	5.047	1.436	1.037	3.768	1.339
16	1.043	11.12	1.554	0.986	4.802	1.632
17	0.985	7.751	1.284	0.948	3.624	1.343
18	0.975	7.588	1.300	1.009	2.903	1.312
19	0.985	9.040	1.345	0.994	5.212	1.355
20	0.971	5.697	1.391	0.915	1.445	0.951
21	1.000	13.295	1.204	0.997	2.681	1.343
avg.	0.980	8.094	1.362	0.983	3.628	1.381
std.	0.024	2.078	0.105	0.037	0.88	0.151

Table 1: *The MOS prediction RMSE per subject, with the average and standard deviation of each model using both fast Fourier transform and peak picking technique*

While table 2 shows a side-by-side comparison of the MOS prediction RMSE results per subject from models using both non-augmented data and augmented data, with fast Fourier transform as the data preprocessing method.

Sbj.	Non-augmented			Augmented		
	SVR	MLP	DTR	SVR	MLP	DTR
1	0.985	8.791	1.312	0.985	1.005	1.427
2	0.970	6.515	1.498	0.970	0.999	1.489
3	0.979	6.656	1.255	0.979	1.030	1.210
4	1.001	10.098	1.300	1.001	1.026	1.506
5	0.955	8.879	1.582	0.955	1.060	1.412
6	0.978	9.741	1.435	0.978	1.017	1.458
7	0.924	6.848	1.414	0.924	0.986	1.450
8	0.945	9.617	1.318	0.945	0.969	1.335
9	0.999	8.287	1.327	0.999	1.089	1.412
10	0.981	5.175	1.279	0.981	1.022	1.092
11	1.000	7.694	1.387	1.000	1.038	1.332
12	0.947	8.874	1.326	0.947	1.059	1.543
13	0.985	4.589	1.474	0.985	1.184	1.281
14	0.989	8.668	1.177	0.989	1.004	1.221
15	0.979	5.047	1.436	0.979	1.104	1.402
16	1.043	11.12	1.554	1.043	1.153	1.381
17	0.985	7.751	1.284	0.985	1.043	1.319
18	0.975	7.588	1.300	0.975	0.995	1.245
19	0.985	9.040	1.345	0.985	0.987	1.199
20	0.971	5.697	1.391	0.971	1.015	1.329
21	1.000	13.295	1.204	1.000	1.034	1.255
avg.	0.980	8.094	1.362	0.980	1.039	1.348
std.	0.024	2.078	0.105	0.024	0.053	0.114

Table 2: *The MOS prediction RMSE per subject, with the average and standard deviation of each model using both non-augmented and augmented data.*

A. Comparison of Preprocessing Methods

The prediction results of each method trained with data preprocessed using both fast Fourier transform and peak picking technique were averaged and compared. Both SVR-based model $(\bar{x}_{SVR(FE)} = 0.98 \pm 0.024)$ and decision tree regressor $(\bar{x}_{DTR(FE)} = 1.362 \pm 0.105)$ performed better with fast Fourier transform as the data preprocessing method, although there was no significant difference in both SVR $(W = 109, W > T)$ and Decision Tree Regressor $(W = 91, W > T)$ performances when the data was preprocessed using fast Fourier transform.

Meanwhile, multi-layer perceptron regressor $(\bar{x}_{MLP(PPT)} = 3.628 \pm 0.88)$ performed significantly better when the data was preprocessed using peak picking technique. $(W = 0, W < T)$

B. Model Training using Non-Augmented Data

The prediction results of each method trained using non-augmented EEG data were averaged and compared. The average RMSE of the SVR-based model performed significantly better $(\bar{x}_{SVR} = 0.98 \pm 0.024)$ compared to the other proposed method as shown by the result of the Wilcoxon signed-rank test. Meanwhile, Decision Tree Regressor performed significantly better $(\bar{x}_{DTR} = 1.362 \pm 0.105)$ compared to multi-layer perceptron regressor $(\bar{x}_{MLP} = 8.094 \pm 2.078)$.

C. *Model Training using Augmented Data*

The prediction results of each method trained using augmented EEG data were averaged and compared. The average RMSE of the SVR-based model performed significantly better ($\bar{x}_{SVR} = 0.98 \pm 0.024$) compared to the other proposed method as shown by the result of the Wilcoxon signed-rank test. Meanwhile, multi-layer perceptron regressor performed significantly better ($\bar{x}_{MLP} = 1.039 \pm 0.053$) compared to Decision Tree Regressor ($\bar{x}_{DTR} = 1.348 \pm 0.114$). It was also found that our SVR-based model also performed better than the baseline CNN-based model, although there was no significant difference in performance ($W = 100, W > T$).

V. Discussions

Our proposed method were that we trained and tested the model using both fast Fourier transform and peak picking technique separately as preprocessing methods, as well as training and testing the model using both non-augmented data and augmented data separately, which then predicts the mean opinion score (MOS) of an audio sample. From the results, it was shown that most models such as SVR dan Decision Tree Regressor performed better when implementing fast Fourier transform as the data preprocessing method. While peak picking technique enables the multi-layer perceptron regressor model to perform significantly better compared to data preprocessed using fast Fourier transform, other models such as SVR and Decision Tree Regressor averaged significantly better when compared to multi-layer perceptron regressor using both fast Fourier transform and peak picking technique.

The difference in results between fast Fourier transform and peak picking technique can be seen from the shape of the data after preprocessing. The implementation of fast Fourier transform was done to extract spectrograms, which were used as training and testing data for the proposed models. The data within a spectrogram are represented with minimal information loss. Hence, spectrograms are able to provide the model with accurate information, compared to the peak properties of an EEG signal which often lack accuracy within the represented data.

It was also shown that most machine learning method performed better when training with augmented data and testing with non-augmented data than training and testing with non-augmented data. Although, it was found that there was no difference in performance between SVR-based model that used non-augmented data and SVR-based model that used augmented data.

Even with the availability of the augmented data, the SVR-based model still did not yield a lower error than training using non-augmented data. This concludes that the SVR-based model is able to linearly separate the EEG features given from the dataset. However, the SVR-based model is limited in yielding lower error distribution, given the fact that training with non-augmented data and augmented data yielded the same error distribution per subject.

The SVR-based model performed better compared to other models due to the characteristics of the EEG data and the characteristics of SVR itself. Most individuals' opinion regarding the quality of a Text-To-Speech (TTS) system tends to be close or almost the same, which can be categorized as linearly separable. It can also be seen through the figure below, which shows the overall opinion scores of five audio samples from five different TTS systems.

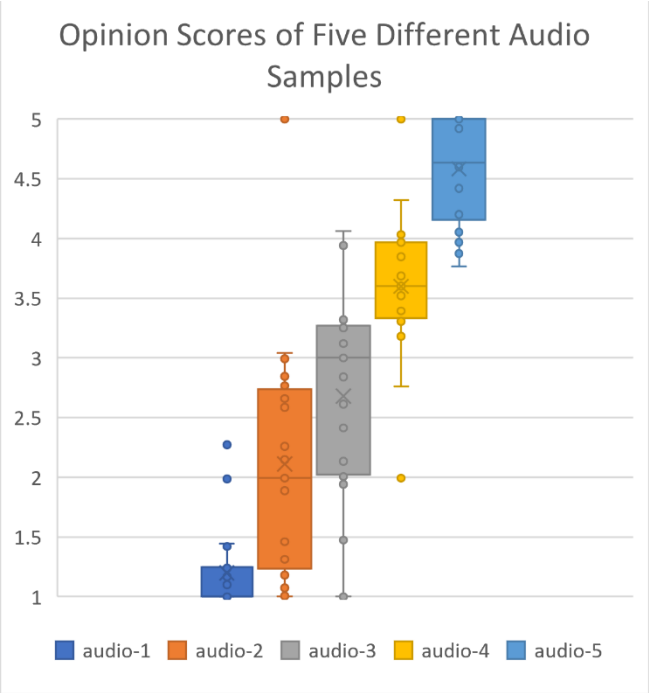

Figure 4: *Opinion scores of five audio samples from different TTS systems.*

The characteristics of SVR that enabled the model to outperform other models are its ability to reduce worst-case errors between actual values and predicted values and its ability to learn high-dimensional data without suffering performance loss.

VI. Conclusion

This study predicted the mean opinion score (MOS) of an audio using electroencephalography (EEG) data. This study also compared the results of using different preprocessing methods, as well as the results of using non-augmented data with augmented data, while determining the significance of models that use different machine learning methods. The Wilcoxon signed-rank test results showed that fast Fourier transform as the data preprocessing method enabled most models to perform better. It also showed that Support Vector Regression performed significantly better among all of the other proposed methods that was trained using non-augmented data and augmented data. The results also showed that the Support Vector Regression (SVR) based model performed better than the baseline Convolutional Neural Network based model, although there was no significant difference.

References

[1] ITU-T, "P. 85. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices," International Telecommunication Union, CH-Genf, 1994.

[2] ITU-T, "P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," ITU-T. Rec., Tech. Rep., 2004.

[3] D. Kim and A. Tarraf, "ANIQUE+: A New American National Standard for Non-intrusive Estimation of Narrow-band Speech Quality: Research Articles," Bell Lab. Tech. J., vol. 12, no. 1, pp. 221–236, May 2007. [Online]. Available: http://dx.doi.org/10.1002/bltj.v12:1.

[4] T. H. Falk and S. Moller, “Towards signal-based instrumental quality diagnosis for text-to-speech systems,” IEEE Signal Processing Letters, vol. 15, pp. 781–784, 2008.

[5] S. M. Patrick Le Callet and A. Perkis, “Qualinet white paper on definitions of quality of experience,” European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, 03 2013.

[6] C. Benoît, M. Grice, and V. Hazan, “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences,” Speech Communication, vol. 18, no. 4. Elsevier BV, pp. 381–392, Jun. 1996. doi: 10.1016/0167-6393(96)00026-x.

[7] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, “Quality prediction of synthesized speech based on perceptual quality dimensions,” Speech Communication, vol. 66. Elsevier BV, pp. 17–35, Feb. 2015. doi: 10.1016/j.specom.2014.06.003.

[8] A. Mariniak, “A global framework for the assessment of synthetic speech without subjects,” Proceedings of the third European Conference on Speech Communication and Technology. 1993.

[9] J.-N. Antons, R. Schleicher, S. Arndt, S. Moller, A. K. Porbadnigk, and G. Curio, “Analyzing Speech Quality Perception Using Electroencephalography,” IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6. Institute of Electrical and Electronics Engineers (IEEE), pp. 721–731, Oct. 2012. doi: 10.1109/jstsp.2012.2191936.

[10] R. Gupta, H. J. Banville, and T. H. Falk, “Physyqx: A database for physiological evaluation of synthesised speech quality-of-experience,” in 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1–5.

[11] R. Gupta, K. Laghari, H. Banville, and T. H. Falk, “Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling,” Human-centric Computing and Information Sciences, vol. 6, no. 1. Springer Science and Business Media LLC, May 05, 2016. doi: 10.1186/s13673-016-0062-5.

[12] I. H. Parmonangan, H. Tanaka, S. Sakti, and S. Nakamura, “Combining Audio and Brain Activity for Predicting Speech Quality,” Interspeech 2020. ISCA, Oct. 25, 2020. doi: 10.21437/interspeech.2020-1559.

[13] H. Maki, S. Sakti, H. Tanaka, and S. Nakamura, “Quality prediction of synthesized speech based on tensor structured eeg signals,” PLOS ONE, vol. 13, no. 6, pp. 1–13, 06 2018.

[14] R. Zafar, S. C. Dass, and A. S. Malik, “Electroencephalogram-based decoding cognitive states using convolutional neural network and likelihood ratio based score fusion,” PLOS ONE, vol. 12, no. 5. Public Library of Science (PLoS), p. e0178410, May 30, 2017. doi: 10.1371/journal.pone.0178410.

[15] M. Shaban and A. W. Amara, “Resting-state electroencephalography based deep-learning for the detection of Parkinson’s disease,” PLOS ONE, vol. 17, no. 2. Public Library of Science (PLoS), p. e0263159, Feb. 24, 2022. doi: 10.1371/journal.pone.0263159.

[16] A. Delorme and S. Makeig, “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” Journal of Neuroscience Methods, vol. 134, no. 1, pp. 9–21, 2004.

[17] A. Mognon et al., “ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features,” Psychophysiology, vol. 48, no. 2, pp. 229–240, 2011.