Combining Audio and Brain Activity for Predicting Speech Quality

Ivan Halim Parmonangan[1], *Hiroki Tanaka*[1,2], *Sakriani Sakti*[1,2], *and Satoshi Nakamura*[1,2]

[1]Division of Information Science, Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

{ivan.halim_parmonangan.ia4, hiroki-tan, ssakti, s-nakamura}@is.naist.jp

Abstract

Since the perceived audio quality of the synthesized speech may determine a system's market success, quality evaluations are critical. Audio quality evaluations are usually done in either subjectively or objectively. Due to their costly and time-consuming nature, the subjective approaches have generally been replaced by the faster, more cost-efficient objective approaches. The primary downside of the objective approaches primarily is that they lack the human influence factors which are crucial for deriving the subjective perception of quality. However, it cannot be observed directly and manifested in individual brain activity. Thus, we combined predictions from single-subject electroencephalograph (EEG) information and audio features to improve the predictions of the overall quality of synthesized speech. Our result shows that by combining the results from both audio and EEG models, a very simple neural network can surpass the performance of the single-modal approach.

Index Terms: EEG, text-to-speech, quality prediction, late-integration

1. Introduction

One factor that may determine a TTS system's success is its perceived audio quality which could refer to overall impression, naturalness, intelligibility, and others. Synthesized audio quality is mainly evaluated by either subjective or objective approaches. The International Telecommunication Union recommends multidimensional subjective tests for quality evaluation, that involve user interviews, ratings, and surveys to identify the the ground truth from the end user's opinions of the audio [1]. However, it is impossible to calculate the mean opinion score of a general population using the opinion scores from only a single participant. Such subjective tests are time consuming and expensive, fueling the development of objective approaches such as [2, 3, 4] are being developed for estimating the quality of synthesized speech.

An objective approach replaces the human listener with a computer algorithm, which learns the mapping between several acoustic features to the previously recorded subjective ratings. Consequently, it lacks the critical human influence factors that motivate our perception of quality [5]. Moreover, the exact relationship between acoustic features and perceived quality remains unclear [6]. Therefore, even though the predicted quality is high, the actual quality might not meet human expectations. For this reason, human influence factors must be considered as well as acoustic features.

Human influence factors are manifested in our brains where the process of forming judgements and creating quality formation occurs [7]. Thus, probing neural activity might provide insight into the human quality judgement process [8]. However, just using EEG features for predictions remains difficult due to their low signal-to-noise ratio (SNR) [9, 10]. The target brain activity is often buried under multiple source of the 'artifacts' of environmental, physiological, and activity-specific noise of similar or greater amplitude.

Brain signals are not only have high variance across and within subjects [11], but they are also non-stationary. Their statistics may vary across time [12, 13, 14], resulting in poor generalization for machine learning models which trained on temporally-limited data. Nevertheless, it was claimed that EEG performed better for emotion detection than the other physiological signals [15].

In audio quality prediction tasks, a previous work [16] predicted individual overall impression, valence, and arousal by using partial least squares (PLS) on subject EEGs. Since this approach still used individual opinion scores instead of the average of the opinion scores or the mean opinion scores (MOS), it still requires a number of participants to calculate the general scores. To minimize the need for human subjects, another study [17] predicted the overall impression MOS of synthesized Japanese speech using only single-subject EEGs instead of individual opinion scores by support vector regression (SVR).

A fusion method between EEG and peripheral signals was proposed because it was more robust than single input type approaches [15]. Another study [18] proposed a fusion EEG and a galvanic skin response (GSR) for emotion recognition with a convolution neural network (CNN). The proposed model outperformed similar studies with a single input type on the same dataset. A similar study also reported that the emotion recognition performance was improved with EEG and eye-tracking data [19].

A study using multiple input sources has also been proposed [20]. They made a linear regression model using audio features, audio features with subjective rating variables, and audio features with neurophysiological features. This study argued that with linear regression, combined features outperformed audio only features. However, their study did not perform any prediction.

In contrast to the previous works and ideas, we performed the following:

1. individually trained both EEG and audio models using CNN-based architecture from previous work using PLS;
2. used both audio and EEG features to do overall impression MOS prediction from previous study that only modeled using linear regression.

2. Dataset

This study used the PhySyQX dataset [21], which includes audio samples and brain activity records of 21 subjects and each subject's audio ratings. The dataset also contains a record of individual opinion scores of subjective dimension ratings for each presented audio stimulus. In this study, we used the audio, the

EEG data, and the overall impression opinion score. The baseline input used only EEG features, while the proposed method used both EEG and audio features.

2.1. Participants

Twenty-one healthy English speakers (eight females and 13 males) whose average age was 23.8 ($\pm$4.35) were recruited. The audio stimuli were presented by earphones at their individual volume preferences. The protocol was approved by the INRS Research Ethics Office after obtaining written, informed consent from every participant.

2.2. EEG Records

The EEG recording was done using Biosemi ActiveTwo system with 512 Hz sampling rate without online filtering and then down-sampled to 256 Hz. All channels of the raw EEG signals were referenced to 'Cz' followed by 0.5-50 Hz bandpass filtering using FIR filter in the EEGLAB [22]. Independent component analysis (ICA) was applied to remove eye blink artifacts and semi-automatic rejection of the noisy components was carried out using the ADJUST toolbox [23].

2.3. Speech Stimuli

The 44 speech stimuli, which were presented to each subject in the dataset, consisted of collected speech from four human speakers and seven commercially available TTS systems. From each human and the TTS system, four English sentences were recorded in durations that ranged from 13 to 22 seconds. These recordings were presented randomly to each subject. Eight copyrighted audio stimuli used in the EEG recording were excluded in the published dataset leaving only 36 audio stimuli available for our study. In summary, the published dataset has eight sample sentences. Four sentences are human spoken and synthesized totalling 32 audio samples. Another four sentences are generated by one TTS system.

2.4. Experiment Procedure

The experiment procedure followed the ITU-T P.85 [1] recommendations. The brain activity were recorded simultaneously as the participants listened to the audio stimuli. Each subject listened to 44 audio samples, however, the dataset only provides 36 out of the 44 used audio recordings because some files are copyrighted. We omitted the EEG records without audio pairs.

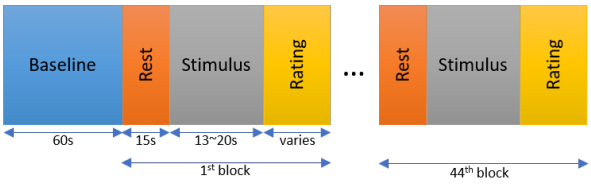

Figure 1: *Stimuli presentation to each subject. Each block consists of rest-stimuli-rating sequences. Baseline stimuli were presented at the beginning followed by 44 blocks of rest-stimulus-rating sequences.*

The experiment scenario in Figure 1 describes how the audio was presented to each subject during the subjective evaluation recording session. Prior to the data collection, subjects were presented with a sample speech that was followed by a series of rating questions to accustom them to the actual task. A 15-seconds rest period was provided before each presentation to allow the subject's brain activity and blood flow to return to the baseline levels. Following each audio stimuli, a randomized series of rating questions was asked to be answered in a continuous scale by the subjects. The rating questions consists of overall impression, naturalness, and others totalling twelve questions collected in continuous scale.

2.5. Mean Opinion Score

The overall impression opinion scores ranged from one-(bad) to five-(excellent) in continuous scale. This scale evaluates the overall quality of the synthesized signal [1]. To calculate the MOS, we averaged all participants' overall impression opinion scores of each audio sample. After collecting all audio MOS, they were normalized to have zero mean and unit variance.

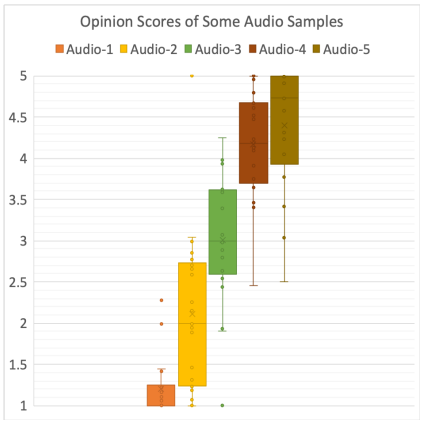

Figure 2: *Overall impression opinion scores of four audio samples. Individual opinion scores have high variance, whereas the mean opinion scores generalizes the opinion scores from the whole population.*

Since the study aimed to predict the generalized rather than the personalized opinion of the overall impression, we used MOS to reduce the individuality in this study. Thus, the EEG models trained in this study are supposed to learn the pattern between the individual EEG data with the MOS. Figure 2 shows how individual opinion scores of some audio samples are distributed among participants.

3. Predicting Mean Opinion Scores

We used partial least square (PLS) regression as the baseline using only the EEG features and tested the proposed CNN-based EEG and audio combined model to predict the overall impression MOS. The tests were done in subject-dependent manner and evaluated using root-mean-square error (RMSE).

3.1. Baseline Partial Least Square Model

The baseline of our study follows a previous work by [16] with the same individual number of components and feature extraction method. In addition to predicting the individual opinion scores, our baseline also predicted the MOS. The baseline in our study used the same train-validation-test method used in our proposed model unlike the leave-one-out cross validation method used in previous work [16].

The features for the baseline input were the channel-based power spectrum density, channel-pair based phase spectrum

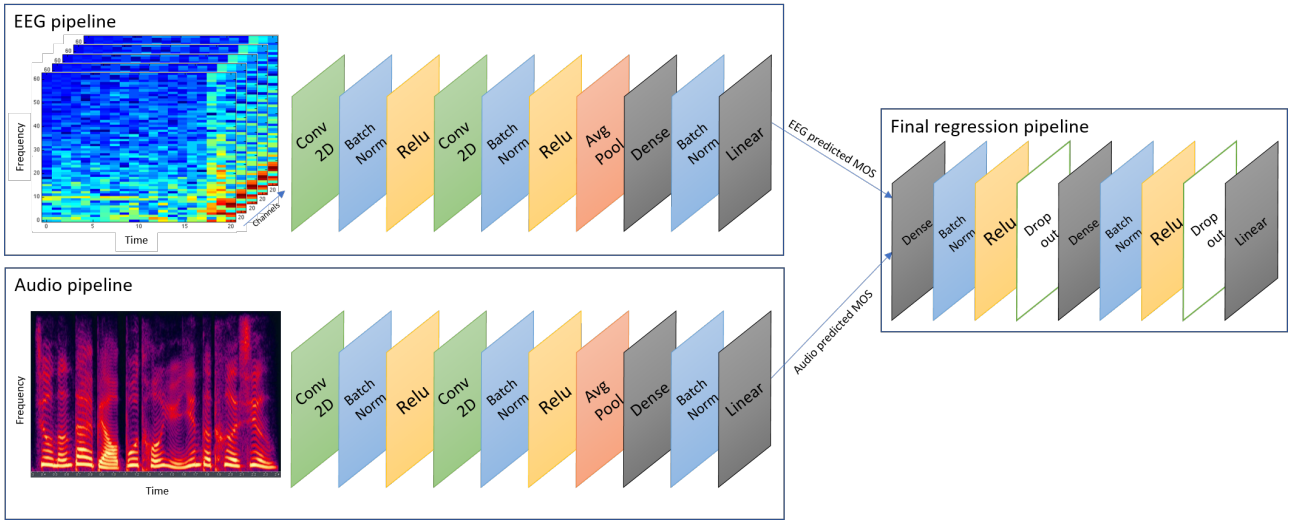

Figure 3: *Simplified illustration of the model architecture: EEG pipeline and audio pipelines were trained independently. Final regression pipeline used the paired outputs from the above pipelines to predict the final MOS.*

density, and channel-pair-based power spectrum density as used in the previous work [16]. Our study also preserved the vector structure used in the previous study. However, the baseline in our study excluded some EEG samples without audio pairs whilst the previous work used all the available EEG samples.

3.2. EEG Feature Extraction

This study only used pre-processed EEGs from the dataset. Since each participant only had 36 samples with the available audio set, each raw signal was augmented 30 times with Gaussian noise. The EEGs were not augmented just for the baseline method.

Data augmentation increases the accuracy and the stability by offering better generalization on new data [24]. Some studies even augmented the signal with artifacts such as eye blinks, muscle activity, and white Gaussian noise to improve the robustness [25, 26, 27].

For each sample, 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. To match the feature used in the previous research [16], we also limited the highest frequency used to 45 Hz.

The pre-processed EEGs were inputted into the EEG pipeline as shown in Figure 3. The input dimension of the EEG pipeline was 45 frequency points, 60 time points, and 64 channels.

3.3. Audio Feature Extraction

The mel-spectrograms were extracted from each audio samples using the librosa package [28] on python. Before the feature extraction, each audio signal was zero-padded at the end to match the maximum length of the recording (around 25 seconds) without any augmentation. Normalization was also applied after the mel-spectrogram extraction.

The pre-processed audio were inputted into the audio pipeline which pictured in Figure 3. The input dimension of the audio pipeline was 80 mel-bands, 1092 time points, and one channel.

3.4. Combined Results from Audio and EEG Features

The EEG model and audio models were basically constructed under the same simple architecture as shown in EEG and audio pipeline in Figure 3. Our convolution layer had 32 filters. The kernel design of the convolution layer was adopted from a previous work [18] that used a 3*2 sized kernel with a 2*1 stride. Each convolution layer was followed by batch normalization, and this study used rectified linear unit (ReLU) as the activation function. After a 2*2 average pooling layer, we applied a 256-unit dense layer followed with batch normalization and the final linear output.

Following the EEG and audio pipelines, a simple, two-hidden-layer neural network with one linear output was implemented for the final regression pipeline. We used the outputs from audio and EEG pipelines as the input of the final regression pipeline, which was constructed with two simple dense layers, each of which consisted of four units. Each dense layer was followed by batch normalization, ReLU activation function, and a 0.2 rate dropout to prevent overfitting. Finally, a linear function was applied to calculate the final output.

3.5. Model Evaluation

The audio and EEG models of each subject were trained independently using nested cross-validation with validation and test sets. We trained both the EEG and audio models with all the possible training, validation, and test combinations.

The cross validation method was done as follows. First, from the four available sets, we selected the combination of two sets without repetition to be used as the train set for six training set combinations. Second, for each generated combination, the remaining two sets were used as validation and test sets. Since two sets remained and only one was used for testing, the training was done twice with the same training set. But the validation and test sets were switched in the next iteration, resulting in 12 models per subject. During the training, the best model of each cross-validation combination was saved based on the lowest validation RMSE.

Twelve result sets, each containing the predicted training, validation, and test scores from both the audio and EEG best

models, were fed into the final neural network predictor. No cross validation was done in the final regression since the data came from each cross validation model combination. Finally, we averaged RMSE of each model.

4. Results

The variance in individual labelling might hinder the generalization process in machine learning even if it is trained using individual EEG signals with their own opinion scores. Although the Wilcoxon signed-rank test result showed no significant difference ($\alpha = 0.01$, $T = 42$, $N = 21$, $W = 104$, $W > T$), Figure 4 shows that training with opinion scores yielded a higher error among participants in average ($\bar{X}_{PLS\_MOS} = 1.156 \pm 0.312$) than using the mean opinion scores ($\bar{X}_{PLS\_OS} = 1.122 \pm 0.275$).

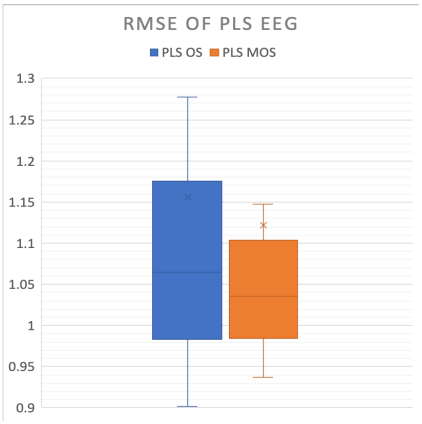

Figure 4: *Error distribution per subject of both mean opinion scores and opinion scores based baselines. Two outliers whose values exceeded the maximum RMSE from each category were excluded from the plot.*

4.1. Independently Trained EEG and Audio Model

The prediction results of audio only features were not comparable to the results of using EEG features because the EEG prediction was done per subject while there was no subject in audio only MOS prediction. Nevertheless, the average RMSE of the CNN-based MOS predictions using audio features was lower ($\bar{X}_{audio} = 0.862$) than using EEG spectrogram ($\bar{X}_{EEG} = 0.984 \pm 0.037$). However, the proposed model achieved lower average RMSE ($\bar{X}_{audio+EEG} = 0.732 \pm 0.017$) than using only audio features.

4.2. Combining The EEG and Audio Prediction Results

The EEG results are shown in Table 1. Each table row shows the test set RMSE of each subject. Using the Wilcoxon signed-rank test ($\alpha = 0.01$, $T = 42$, $N = 21$), the CNN-based EEG model results were significantly better than the baseline EEG model results ($W = 27$, $W < T$). Our proposed late-fusion model results were also significantly better than the CNN-based EEG model ($W = 0$, $W < T$).

5. Discussion

Our proposed model was trained and tested using individual EEG records following [29] that reports emotion regulation is subject-dependent. The proposed model predicted MOS rather than individual opinion scores.

Table 1: *MOS prediction RMSE per subject. The proposed method produced significantly lower RMSE than the baseline.*

Sbj.	**PLS**$_{eeg}$	**CNN**$_{eeg}$	**CNN**$_{aud.+eeg}$	**CNN**$_{aud.}$
1	1.102	1.084	0.752	0.862
2	0.990	0.974	0.767	-
3	0.948	1.019	0.737	-
4	0.997	1.010	0.719	-
5	1.007	0.947	0.750	-
6	1.143	0.962	0.745	-
7	0.965	0.927	0.751	-
8	1.147	0.962	0.717	-
9	1.104	0.998	0.751	-
10	2.088	0.942	0.733	-
11	0.978	0.976	0.742	-
12	1.036	0.983	0.732	-
13	1.018	0.994	0.722	-
14	1.032	0.994	0.731	-
15	1.086	1.003	0.710	-
16	1.785	1.005	0.735	-
17	1.068	0.981	0.729	-
18	1.075	1.010	0.742	-
19	0.971	1.034	0.709	-
20	0.937	0.927	0.714	-
21	1.091	0.936	0.694	-
avg.	1.122	0.984	0.732	0.862
std.	0.275	0.037	0.017	-

The baseline PLS results showed that using MOS rather than opinion scores as the label also minimized the error although it was not significant. Training and testing with just one subject might yield better models since they have less data variance to generalize. However, the number of samples was greatly reduced. Therefore, creating a subject-independent model is an interesting future work.

Audio-only features performed better than individual EEGs under the same model architecture. This finding agrees with another study [20] that concluded that audio features are more informative than EEG features to predict subjective quality ratings.

By combining the predicted MOS from the independently trained EEG and audio model using a late-fusion approach, our prediction performance surpassed both the EEG-only and audio-only models. Investigating what part or features of the EEG signals enables the improvement of the prediction might be productive in the future.

6. Conclusion

This study predicted the mean opinion score of the audio's overall quality by combining single subject EEG with the audio sample. The experimental result showed that by combining results from both independently trained audio and EEG models, the error could be reduced significantly.

7. Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR19A5 and JSPS KAKENHI Grant Number JP17H06101 and JP17K00237, Japan.

8. 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] D. Kim and A. Tarraf, "Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007.

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

[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. Mayo, R. A. Clark, and S. King, "Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311 – 326, 2011.

[7] J.-N. Voigt-Antons, S. Arndt, R. Schleicher, and S. Möller, *Brain Activity Correlates of Quality of Experience*, 03 2014, pp. 109–119.

[8] J.-N. Voigt-Antons, *EEG Frequency Band Power Changes Evoked by Listening to Audiobooks with Varying Quality Profiles*, 02 2015, pp. 73–80.

[9] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The prep pipeline: standardized preprocessing for large-scale eeg analysis," *Frontiers in Neuroinformatics*, vol. 9, p. 16, 2015.

[10] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort, "Autoreject: Automated artifact rejection for meg and eeg data," *NeuroImage*, vol. 159, pp. 417 – 429, 2017.

[11] S. Saha and M. Baumert, "Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review," *Frontiers in Computational Neuroscience*, vol. 13, 12 2019.

[12] S. Cole and B. Voytek, "Cycle-by-cycle analysis of neural oscillations," *Journal of Neurophysiology*, vol. 122, no. 2, pp. 849–861, 2019, pMID: 31268801. [Online]. Available: https://doi.org/10.1152/jn.00273.2019

[13] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hämäläinen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations," *NeuroImage*, vol. 70, pp. 410 – 422, 2013.

[14] Y.-W. Shen and Y.-P. Lin, "Challenge for affective brain-computer interfaces: Non-stationary spatio-spectral eeg oscillations of emotional responses," *Frontiers in Human Neuroscience*, vol. 13, p. 366, 2019.

[15] Z. Khalili and M. H. Moradi, "Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of eeg," in *2009 International Joint Conference on Neural Networks*, 2009, pp. 1571–1575.

[16] 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.

[17] I. H. Parmonangan, H. Tanaka, S. Sakti, S. Takamichi, and S. Nakamura, "Speech Quality Evaluation of Synthesized Japanese Speech Using EEG," in *Proc. Interspeech 2019*, 2019, pp. 1228–1232.

[18] Y.-H. Kwon, S.-B. Shin, and S.-D. Kim, "Electroencephalography based fusion two-dimensional (2d)-convolution neural networks (cnn) model for emotion recognition system," *Sensors*, vol. 18, no. 5, p. 1383, Apr 2018.

[19] W. Zheng, B. Dong, and B. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 5040–5043.

[20] 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, p. 5, 2016.

[21] 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.

[22] 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.

[23] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "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.

[24] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, vol. abs/1611.03530, 2016.

[25] S. Jirayucharoensak, S. Pan-ngum, and P. Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *TheScientificWorldJournal*, vol. 2014, p. 627892, 09 2014.

[26] Z. Yin and J. Zhang, "Cross-session classification of mental workload levels using eeg and an adaptive deep learning model," *Biomedical Signal Processing and Control*, vol. 33, pp. 30 – 47, 2017.

[27] ——, "Cross-subject recognition of operator functional states via eeg and switching deep belief networks with adaptive weights," *Neurocomputing*, vol. 260, pp. 349 – 366, 2017.

[28] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, and F. Zalkow, "librosa/librosa: 0.7.2," 2020.

[29] J. Gross and O. John, "Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being," *Journal of personality and social psychology*, vol. 85, pp. 348–62, 09 2003.