2022 10th International Winter Conference on Brain-Computer Interface (BCI) | 978-1-6654-1337-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/BCI53720.2022.9735124

EEG-Transformer: Self-attention from Transformer Architecture for Decoding EEG of Imagined Speech

Young-Eun Lee
Dept. Brain and Cognitive Engineering
Korea University
Seoul, Republic of Korea
ye_lee@korea.ac.kr

Seo-Hyun Lee
Dept. Brain and Cognitive Engineering
Korea University
Seoul, Republic of Korea
seohyunlee@korea.ac.kr

***Abstract*—Transformers are groundbreaking architectures that have changed a flow of deep learning, and many high-performance models are developing based on transformer architectures. Transformers implemented only with attention with encoder-decoder structure following seq2seq without using RNN, but had better performance than RNN. Herein, we investigate the decoding technique for electroencephalography (EEG) composed of self-attention module from transformer architecture during imagined speech and overt speech. We performed classification of nine subjects using convolutional neural network based on EEGNet that captures temporal-spectral-spatial features from EEG of imagined speech and overt speech. Furthermore, we applied the self-attention module to decoding EEG to improve the performance and lower the number of parameters. Our results demonstrate the possibility of decoding brain activities of imagined speech and overt speech using attention modules. Also, only single channel EEG or ear-EEG can be used to decode the imagined speech for practical BCIs.**

Keywords*—transformer, attention module, brain-computer interface, imagined speech

I. Introduction

Brain-computer interfaces (BCIs) are one of the most important consideration for communication systems in real life. Many researchers have studied BCI to recognize human cognitive state or intention based on brain signals such as electroencephalography (EEG) to recognize the crucial features from the brain activity. [1]–[4]. To enhance the performance of decoding EEG signals, preprocessing technology is also important to get a high quality signals with higher accuracy of decoding and lower signal-to-noise ratio [5]–[8]. Moreover, the decoding technologies including feature extraction and classification have improved significantly in recent years [9]–[12].

This work was partly supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00451, Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning; No. 2015-0-00185, Development of Intelligent Pattern Recognition Softwares for Ambulatory Brain Computer Interface; No.2021-0-02068, Artificial Intelligence Innovation Hub).

Recognizing brain activities during speech or imagined speech has recently attracted a lot of attention and is developing [13], [14]. In particular,imagined speech is evaluated as an advanced technology for brain signal-based communication systems [15]–[17]. Imagined speech refers to the internal pronunciation of speech only by imagination without auditory output or pronunciation [18]. Recent studies have shown some features and potentials of imagined speech decoding [16], [19], but fundamental neural properties and their practical use remain to be investigated. Therefore, research on the decoding of imagined speech requires the development of brain signal decoding techniques for more accurate and practical BCI [20], [21].

Several deep learning techniques have been published to decode EEG brain signals, which are architectural designs that considers the characteristics of brain signal characteristics [22], [23]. It was often used to decode human intention using motor imagery or event-related potential, and have shown superior performance than the conventional machine learning methods such as linear discriminant analysis and support vector machine [11], [24], [25]. Recently, there are several attempts to find optimal features of EEG by deep neural networks based on the three main features of EEG, temporal, spectral, and spatial features [26], [27]. In addition, EEG-based speaker identification studies also have actively applied machine learning or deep learning techniques [28], [29]. Deep learning may be effective in capturing prominent features from brain signals to verify individual characteristics.

Transformer [30] is a model from Google's 2017 paper "Attention is all you need," and is implemented only with attention, while following the encoder-decoder, the existing structure of seq2seq. This model does not use RNN, and even though the encoder-decoder structure is designed, it also has better performance than RNN. This is the basis for famous language models such as GPT-3 and DALL-E, and tools such as the Hugging Face Transformers library have made it easy for machine learning engineers to solve a wide range of NLP tasks and have since promoted numerous innovations in NLP and other fields [31]–[34]. Transformer's attention

978-1-6654-1337-4/22/$31.00 ©2022 IEEE

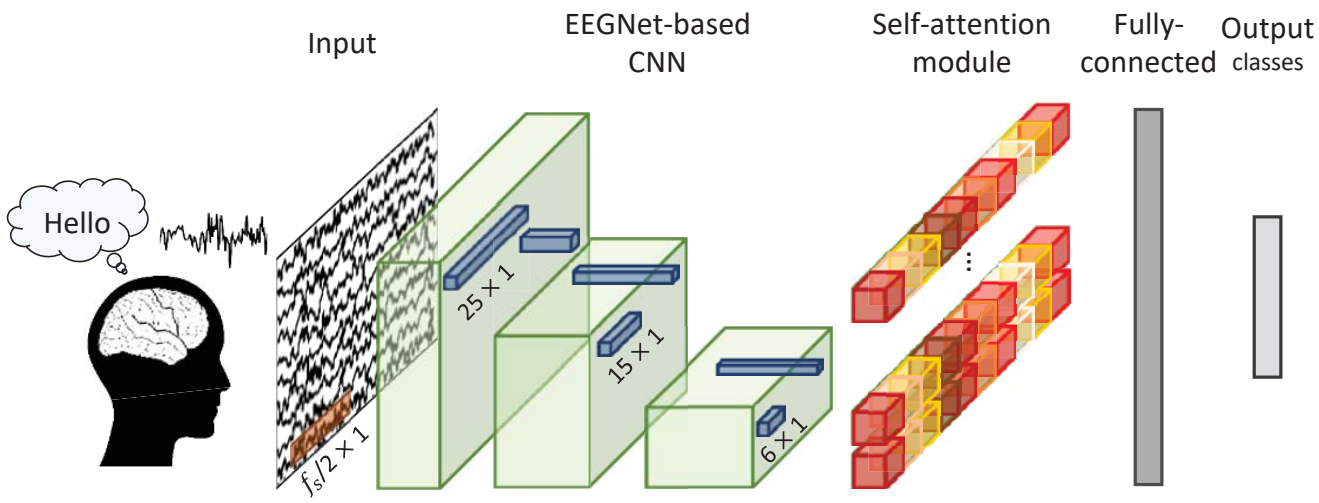

Fig. 1. Total frameworks in this study. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable classification token to the sequence.

was created to overcome the limitations of RNN, which was slow in computation due to difficulties in parallel processing. Transformers do not need to process data sequentially like RNNs. In addition, this processing method is possible because it allows much more parallelization than RNN.

Recently, there have been several attempts to commercialize BCI technology [35], [36]. For example, portable and non-hair EEGs were frequently investigated to improve the applicability of BCI in real life, and endogenous paradigms such as motor imagery and imagined speech are used rather than exogenous paradigm such as event-related potential and steady-state visual evoked potential, which needs external devices to give stimuli [37]. In particular, the ear-EEG composed of electrodes disposed inside or around the ear has many advantages over the existing scalp-EEG in terms of stability and portability. In addition, since the Broca-Wernicke region, which is mainly analyzed during overt speech or imagined speech, is distributed close to the left ear channels, there is a possibility that only a small number of channels can be used to recognize the user's intention [13], [16].

II. Materials and Methods

A. Data Description

The experimental protocol followed the previous works [16], [20]. Nine subjects (three males; age 25.00 ± 2.96) participated in the study. The study was approved by the Korea University Institutional Review Board [KUIRB-2019-0143-01] and was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all subjects.

We recorded EEG signals from scalp during overt speech and imagined speech. After recording two seconds of resting state, two more seconds of voice audio for each word/phrase were provided, followed by consecutive trials of imagined speech or overt speech [16], [38]. During the experiment in which each block was repeated the imagined or overt speech four times, only the first trial of each block was used to match the number of trials with different experimental conditions. Each participant conducted a random experiment 25 times for every 12 words, and a total of 300 trials for each condition. There are 13 classification outputs, consisting of 12 words (ambulance, clock, hello, help me, light, pain, stop, thank you, toilet, TV, water, and yes) and resting state.

B. EEG Preprocessing

The EEG signal was down-sampled at 250 Hz and divided into 2 seconds from the start of each trial. The preprocessing of EEG signal was performed with a 5th Butterworth filter in the high-gamma band of 30–120 Hz, and baseline was corrected by subtracting the average of 500 ms before the start of each trial. We selected the channels located in the Broca and Wernicke's areas (AF3, F3, F5, FC3, FC5, T7, C5, TP7, CP5, and P5). For removing the artifacts of EOG and EMG from muscle activity around mouth, we conducted artifact removal methods using independent component analysis with references from EOG and EMG. All data processing procedures were performed in Python and Matlab using OpenBMI Toolbox [39], BBCI Toolbox [40], and EEGLAB [41].

C. Architecture

The proposed classification framework consists of convolution layers and separable convolution layers for extracting time-spectral-spatial information, as shown in Fig. 1. Given the input as raw signals (C × T), classification output is set to 13 classes. The kernel size of the first layer is set in relation to the sampling frequency of the data for performing a temporal convolution that imitates the band-pass filter [26]. Since support vector machine (SVM) classifier has been reported to be robust in decoding the imagined speech

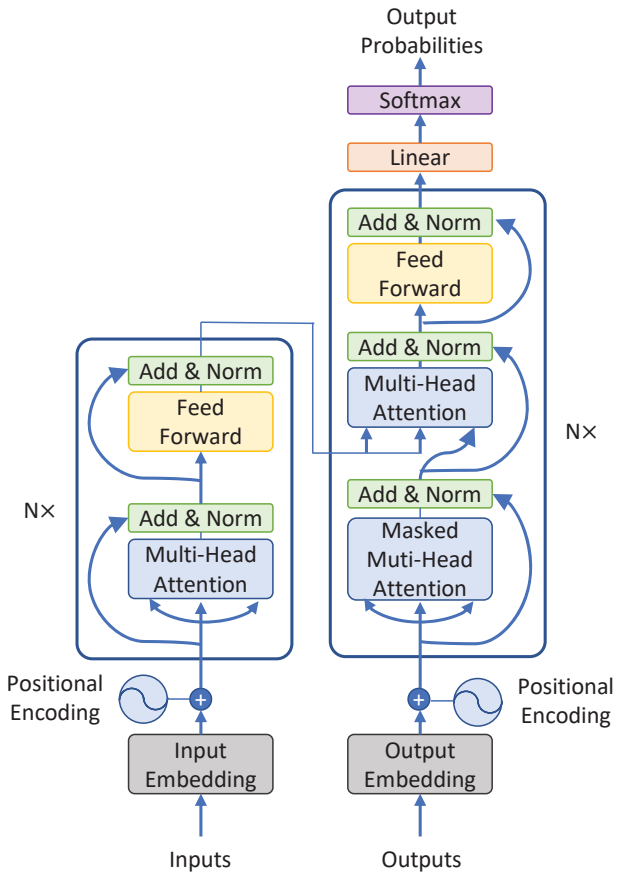

Fig. 2. Transformer architecture. Self-attention and the feed-forward networks are followed by dropout and bypassed by a residual connection with subsequent layer normalization. This figure is inspired from the Transformer paper [30].

[16], [19], we used the squared hinge loss for training that functions similar to the margin loss of SVM. The evaluation was conducted through 5-fold cross-validation and 1000 epoch training for each condition. The probability of chance level of this experiment was 11.11% because the number of samples of each subject were the same for each state condition.

The self-attention module was shown is Fig. 1 and Fig. 2 as well. The attention module serves for mapping a set of query and key-value pairs, where the output is calculated as the weighted sum of the values, where the weights assigned to each value are calculated by the compatibility function of the query and its key. The multi-head attention can jointly focus on information in different representation subspaces at different positions, resulting in an average calculation with one attention.

D. Statistical Analysis

Statistical analysis were performed to verify the results of classification. Kruskal-Wallis non-parametric one-way analysis of variance (ANOVA) were performed to compare the classification performance of imagined speech and overt speech. Post-hoc analysis was conducted with non-parametric permutation-based t-test. The Kruskal-Wallis test was also performed on classification performance using a single channel EEG to estimate the significance of the selected channel. In addition, a paired *t*-test was performed to identify significant connectivity changes in Broca and Wernicke's area during imagined speech and resting states.

III. Results and Discussion

A. Decoding Performance

We developed the frameworks of decoding the speech-related EEG signals of 13 classes in two conditions of imagined speech and overt speech. The performance of imagined speech and overt speech was compared. The average accuracy of overt speech was 49.5% for 13 classes, including 9 subjects' performances. The EEG signal during overt speech can contain more significant representation in brain activities. As we conducted preprocessing to remove artifacts related to EOG and EMG around mouth, the EEG signal only contains the brain activity to intent to move mouth and tongue to speak out the pronunciation for each word. The average accuracy of imagined speech was 35.07% for 13 classes, including 9 subjects' performances. The EEG signal during imagined speech includes only brain activities rather than EMG since they did not move their muscle. Therefore, the performance of imagined speech normally inferior than it of overt speech. However, the difference between overt speech and imagined speech was significantly different($p < 0.05$), but not so huge while overt speech was expected to show superior performance.

B. Attention Module

We showed that deep learning model with self-attention module could show reasonable performance. The advantages of the self-Attention module are that it can reduce the total computational complexity per layer, that it can parallelize the computational volume to some extent, and that the path length of long-term dependency in the networks is short.

IV. Conclusion

In this study, we proposed attention module based on transformer architecture to decode imagined speech in EEG. As practical BCIs require a robust system and simple hardware usable in the real-world, we show that the proposed method improved the BCI performance. The results of recognizing speech from human intention had reasonable performance although we used only few channels. And we compared overt speech and imagined speech in terms of performance and statistical analysis. The EEG of overt speech showed superior performance than imagined speech, which was significant different, but not that huge than we expected. Therefore, technology of decoding imagined speech with attention module had potential to use as a real-world communication system. In the future, we developed the architecture that performed with higher performance for imagined speech. Moreover, parameter optimization of self-attention module can increase the performance as well.

REFERENCES

[1] Y. Zhang, H. Zhang, X. Chen, S.-W. Lee, and D. Shen, "Hybrid high-order functional connectivity networks using resting-state functional mri for mild cognitive impairment diagnosis," *Sci. rep.*, vol. 7, no. 1, pp. 1–15, 2017.

[2] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional cnn-bilstm network using eeg signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, 2020.

[3] Y. Zhang, H. Zhang, X. Chen, M. Liu, X. Zhu, S.-W. Lee, and D. Shen, "Strength and similarity guided group-level brain functional network construction for mci diagnosis," *Pattern Recognit.*, vol. 88, pp. 421–430, 2019.

[4] D.-O. Won, H.-J. Hwang, D.-M. Kim, K.-R. Müller, and S.-W. Lee, "Motion-based rapid serial visual presentation for gaze-independent brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 334–343, Aug. 2017.

[5] T. Castermans, M. Duvinage, M. Petieau, T. Hoellinger, C. De Saedeleer, K. Seetharaman, A. Bengoetxea, G. Cheron, and T. Dutoit, "Optimizing the performances of a P300-based brain-computer interface in ambulatory conditions," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 1, no. 4, pp. 566–577, Dec. 2011.

[6] Y.-E. Lee, N.-S. Kwak, and S.-W. Lee, "A real-time movement artifact removal method for ambulatory brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2020.

[7] N.-S. Kwak, K.-R. Müller, and S.-W. Lee, "A lower limb exoskeleton control system based on steady state visual evoked potentials," *J. Neural Eng.*, vol. 12, no. 5, p. 056009, Aug. 2015.

[8] Y.-E. Lee, M. Lee, and S.-W. Lee, "Reconstructing erp signals using generative adversarial networks for mobile brain-machine interface," *arXiv preprint arXiv:2005.08430*, 2020.

[9] N.-S. Kwak, K.-R. Müller, and S.-W. Lee, "A convolutional neural network for steady state visual evoked potential classification under ambulatory environment," *PloS One*, vol. 12, no. 2, p. e0172578, Feb. 2017.

[10] A. D. Nordin, W. D. Hairston, and D. P. Ferris, "Dual-electrode motion artifact cancellation for mobile electroencephalography," *J. Neural Eng.*, vol. 15, no. 5, p. 056024, Aug. 2018.

[11] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, 2019.

[12] Y.-E. Lee and M. Lee, "Decoding visual responses based on deep neural networks with ear-EEG signals," in *Int. Winter Conf. Brain-Computer Interface (BCI)*, Jeongseon, Republic of Korea, Feb. 2020, pp. 1–6.

[13] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[14] J.-M. Schoffelen, A. Hultén, N. Lam, A. F. Marquand, J. Uddén, and P. Hagoort, "Frequency-specific directed interactions in the human brain network for language," *Proc. Natl. Acad. Sci. (PNAS)*, vol. 114, no. 30, pp. 8083–8088, 2017.

[15] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun. 2002.

[16] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2020.

[17] M.-H. Lee, J. Williamson, D.-O. Won, S. Fazli, and S.-W. Lee, "A high performance spelling system based on EEG-EOG signals with visual feedback," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 7, pp. 1443–1459, 2018.

[18] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2257–2271, 2017.

[19] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using riemannian manifold features," *J. Neural Eng.*, vol. 15, no. 1, p. 016002, 2017.

[20] S.-H. Lee, M. Lee, and S.-W. Lee, "EEG representations of spatial and temporal features in imagined speech and overt speech," in *Proc. Asian Conf. Pattern Recognit. (ACPR)*, 2019, pp. 387–400.

[21] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[22] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, Aug. 2017.

[23] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056013, Jul. 2018.

[24] H.-I. Suk and S.-W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Trans. PAMI*, vol. 35, no. 2, pp. 286–299, Feb. 2012.

[25] Z. Gao, W. Dang, M. Liu, W. Guo, K. Ma, and G. Chen, "Classification of EEG signals on VEP-based BCI systems with broad learning," *IEEE Trans. Syst. Man Cybern.: Syst.*, 2020.

[26] N. Waytowich, V. J. Lawhern, J. O. Garcia, J. Cummings, J. Faller, P. Sajda, and J. M. Vettel, "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," *J. Neural Eng.*, vol. 15, no. 6, p. 066031, 2018.

[27] M. H. Bhatti, J. Khan, M. U. G. Khan, R. Iqbal, M. Aloqaily, Y. Jararweh, and B. Gupta, "Soft computing-based EEG classification by optimal feature selection and neural networks," *IEEE Trans. Industr. Inform.*, vol. 15, no. 10, pp. 5747–5754, 2019.

[28] L. A. Moctezuma, A. A. Torres-García, L. Villaseñor-Pineda, and M. Carrillo, "Subjects identification using EEG-recorded imagined speech," *Expert Syst. Appl.*, vol. 118, pp. 201–208, 2019.

[29] D. Dash, P. Ferrari, and J. Wang, "Spatial and spectral fingerprint in the brain: Speaker identification from single trial MEG signals." in *Proc. Interspeech*, 2019, pp. 1203–1207.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural inf. processing syst. (NIPS)*, California, USA, 2017, pp. 5998–6008.

[31] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, virtual, 2020, pp. 10 076–10 085.

[32] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Int. conf. machine learning (ICML)*. California, USA: PMLR, 2019, pp. 7354–7363.

[33] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *Int. Conf. on Machine Learning*. Stockholm, Sweden: PMLR, 2018, pp. 4055–4064.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[35] L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *J. Neural Eng.*, vol. 14, no. 3, p. 036020, 2017.

[36] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Sci. Rep.*, vol. 5, p. 16743, Nov. 2015.

[37] Y.-E. Lee, G.-H. Shin, M. Lee, and S.-W. Lee, "Mobile BCI dataset of scalp- and ear-EEGs with ERP and SSVEP paradigms while standing, walking, and running," *Sci. Data*, pp. 1–12, 2021.

[38] S.-H. Lee, M. Lee, J.-H. Jeong, and S.-W. Lee, "Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery," in *Conf. Proc. IEEE. Int. Conf. Syst. Man Cybern. (SMC)*. IEEE, 2019, pp. 4409–4414.

[39] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, p. giz002, Jan. 2019.

[40] R. Krepki, B. Blankertz, G. Curio, and K.-R. Müller, "The Berlin Brain-Computer Interface (BBCI)–towards a new communication channel for online control in gaming applications," *Multimed. Tools. Appl.*, vol. 33, no. 1, pp. 73–90, Feb. 2007.

[41] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.