

Corresponding an Audio-Processing Transformer to EEG Brain Activity Induced by Naturalistic Audio-Visual Video Stimuli

Hiroki Kurashige

School of Information Telecommunication Engineering,
Research & Information Center (TRIC),
Institute of Advanced Biosciences
Tokai University
Tokyo, Japan
h.kura00@gmail.com

Jun Kaneko

School of Information Telecommunication Engineering
Tokai University
Tokyo, Japan

Abstract— We attempted to relate EEG brain activities evoked by naturalistic audio-visual video stimuli to the outputs of an audio-processing Transformer induced by audio inputs extracted from the same stimuli. We found a good correspondence, especially in low-frequency brain activity. This is complementary to a previous study that showed good correspondence between high-frequency brain activity and a movie-processing Transformer. These suggest that combining audio- and movie-processing Transformers and using them as a brain simulator is promising. That is, by utilizing this, it should be possible to synthesize audio-visual video stimuli that can intervene in a variety of brain activities and functions.

Keywords— human brain; electroencephalogram; audio-visual stimuli; deep neural networks; audio-processing Transformer.

I. INTRODUCTION

In recent years, methods to induce specific brain activities using artificially designed sensory stimuli and thereby non-invasively treat neurological disorders have been investigated [1]–[5]. However, the repertoire of brain activities that can be induced by the methods is narrow. If a more diverse range of brain activities can be induced with artificially designed sensory stimuli, the treatment of a wider variety of diseases and the enhancement of brain function will be possible in a non-invasive and easy way. The final goal of the present study is to develop an optimal synthesis method for audiovisual stimuli that will enable such applications. In particular, we aim to do this by utilizing deep neural networks (DNNs).

The brain is the origin of DNNs, and many of the computational units and architectures of DNNs are inspired by those of the brain. Furthermore, there is a correspondence between the functions of the brain and DNNs. That is, when the same stimulus is presented to the brain and a DNN, it is possible to map the response of one to the response of the other by a simple transformation [6]–[8]. Such functional correspondences may enable the use of DNNs as brain simulators, enabling the evaluation of the effects of specific

sensory stimuli on the brain by presenting the stimuli on the DNNs and observing their responses.

This may lead to the development of new intervention methods utilizing DNN brain simulators in which one can design effective intervention stimuli through optimization of them to induce brain activity as required for therapeutic purposes (see [9]–[11]). Since this methodology is applicable to a wide variety of brain activities, it should be useful not only for clinical treatments but also for the enhancement of cognitive abilities in healthy humans. In particular, intervention in neural oscillations (or brain waves) in specific frequency bands is an important challenge. Brain waves in the range of a few Hz to over 100 Hz are known to be involved in various functions of healthy and diseased brains [12]–[14]. In fact, methods of intervening in such oscillatory activity using external stimuli such as magnetic or electrical stimulation to treat diseases and enhance brain function have been studied [15]–[17]. Therefore, interventions based on the use of DNNs as brain simulators also need to process such high-speed activities.

Since functional magnetic resonance imaging which is a method typically used in previous studies is unable to capture such activities, the authors have attempted to establish a correspondence between DNNs and brain activity measured using the electroencephalogram (EEG), which is capable of measuring at millisecond-order resolution [18]. Given that the goal is to design optimal sensory stimuli for intervention based on the correspondence between DNNs and the brain, it is appropriate to use multimodal stimuli that enable more efficient and diverse interventions. Therefore, in the authors' previous study, naturalistic audio-visual video stimuli were used [18]. There, we investigated the correspondence between EEG brain activity and silent movie-processing DNNs. As a result, a good correspondence was observed, particularly in the high-frequency bands such as beta and gamma oscillations. Although this result was neurophysiologically plausible, to achieve the goal of intervening in brain activity and functions using sensory stimuli, it is necessary to correlate DNNs with brain activity in the low-frequency range as well.

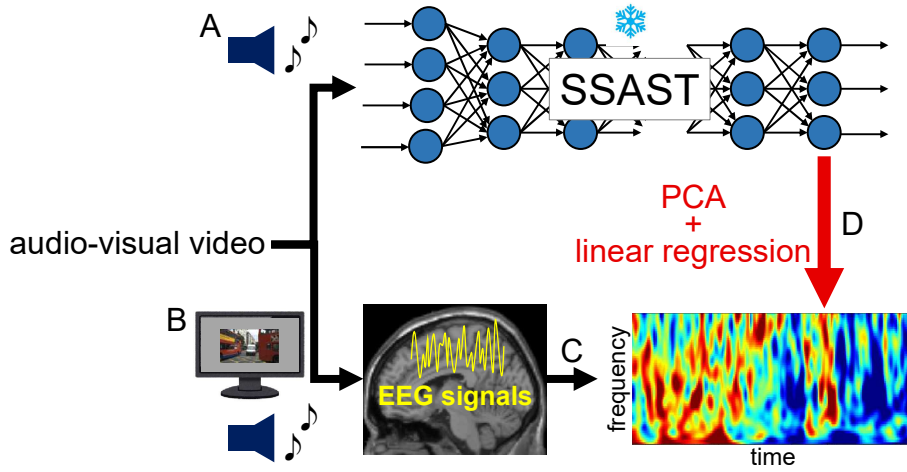


Figure 1: Schematic image of the experimental and analysis procedure in this study. (A) Naturalistic audio stimuli extracted from the audio-visual videos are presented to an audio-processing DNN, SSAST, that is pre-trained with naturalistic audio datasets using a self-supervised learning framework. (B) The same audio-visual videos are presented to human participants and their evoked brain activity is measured using EEG. Note that visual stimuli are also presented here, not just audio stimuli. (C) The power spectral density is calculated from the EEG data. (D) The power spectral densities are predicted by linear regression from the SSAST outputs that are dimensionality-reduced using principal component analysis (PCA).

In our experiment of this previous study, audio-visual stimuli were used, but only the visual modality was analyzed as we mentioned above. However, if the DNN for audio processing shows a good correspondence with brain activity in the low-frequency range, the intervention in brain activity and functions using audio-visual stimuli will be more promising. Therefore, here, we investigated the correspondence between the responses of an audio spectrogram Transformer and EEG brain activity responses to a variety of naturalistic audio-visual video stimuli.

II. RELATED WORK

In previous studies that have investigated the correspondence between EEG brain activity and DNNs, visual stimuli such as images and videos have mainly been used. In [19], an experimental paradigm called rapid serial visual presentation that presents images at high speed was used to record EEG responses from 10 participants in 16,740 image conditions. The same images were also input into DNNs to obtain their responses. Based on this data, a mapping was determined that corresponded between the DNN responses and EEG brain activity. This mapping was also able to predict the brain activity of new participants. In [20], using the projection of EEG responses evoked by image stimuli into the shared subspace as the CLIP embedding, zero-shot decoding and image reconstruction from brain activity were demonstrated.

However, the correspondence between EEG and DNN responses to auditory stimuli is a largely unexplored field. Furthermore, even if functional magnetic resonance imaging studies are included, most existing studies on the correspondence between auditory brain activity and DNNs focus only on speech or music (e.g. [21], also see [22]), and the correspondence in response to more general auditory stimuli is largely unknown. Therefore, in the present study, we used various naturalistic auditory stimuli to induce DNN responses and attempted to correspond them to EEG brain activity.

III. MATERIALS AND METHODS

The dataset used in this study is the same as that used in a previous study conducted by the authors [18], which investigated the correspondence between video-processing DNN and EEG brain activity. In that study, the analysis focused on visual processing, but in this study, the analysis focused on audio processing. Therefore, the experimental conditions and preprocessing of the EEG data are the same as in the previous study, so we will explain them briefly here. For details, please refer to [18].

A. Participants

The study involved five healthy young male participants. Their ages ranged from 19 to 22 years. All had normal or corrected-to-normal vision and normal audition. They also had no history of psychiatric or neurological disorders. All participants were native Japanese speakers. None had any experience of learning Spanish or Russian, and therefore were unable to receive linguistic meaning from the sounds of these languages. This study was approved by the institutional ethics committee of Tokai University (approval number: 21107) and was carried out following the Declaration of Helsinki. Written informed consent was obtained from each subject.

B. Experimental Procedures and Stimuli

An experiment was conducted to measure EEG responses to naturalistic audio-visual video stimuli. The experiment consisted of three days in total. 472 10-second short clips, three Spanish and one Russian documentary were presented. In the experiment using short clips, participants pressed a keyboard button to answer whether the clip they had just watched was to their liking or not, in the time between clips. This was a task to maintain participants' concentration. On the other hand, in the experiment using documentaries, participants were only required to watch them, and there were no additional tasks.

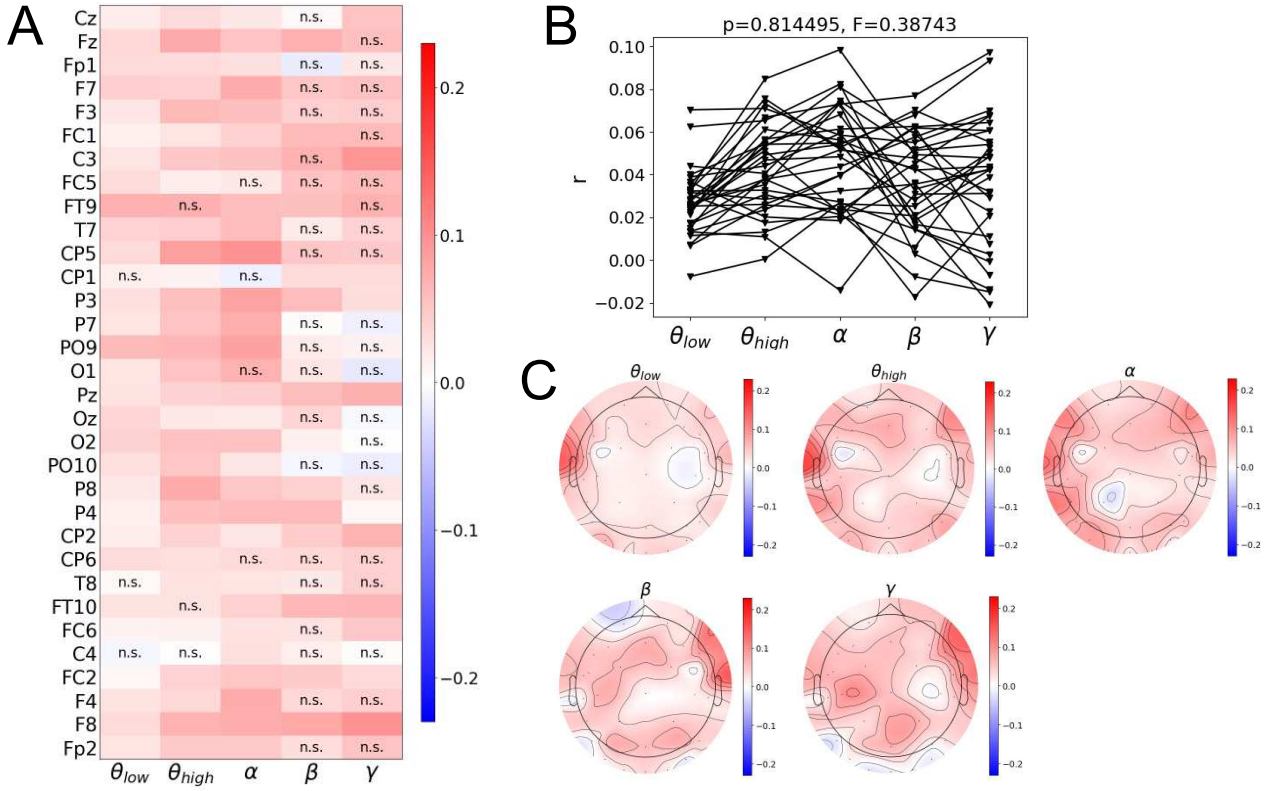


Figure 2: Prediction of brain activity induced by the documentary video stimuli from SSAST outputs. (A) Correlation coefficients between experimental and predicted values are shown using a color code. The horizontal axis represents frequency bands, and the vertical axis represents channels. 'n.s.' indicates locations where no statistically significant positive correlation was observed, and other locations indicate statistically significant positive correlations. (B) Frequency band dependence of the correlation coefficient between experimental and predicted values. The vertical axis shows the average correlation coefficient for all participants. Each line corresponds to an EEG channel. (C) Topographic plot of the correlation coefficient between experimental and predicted values. The color code indicates the correlation coefficient values.

The four documentary videos with Spanish or Russian audio were downloaded from the video-sharing site Vimeo (<https://vimeo.com/>), and used as they were. The length of the videos was 18min 5sec (Spanish), 21min 21sec (Spanish), 16min 53sec (Spanish), and 23min 41sec (Russian). The themes of the videos were 'royal guard,' 'conflicts between a company and local people over development,' 'issues of environmental conservation and fisheries,' and 'war,' respectively. The 472 short clips were made by combining 118 silent 10-second videos extracted from 118 videos downloaded from the NHK Creative Library website (<https://www.nhk.or.jp/archives/creative/en/>) with 118 audios downloaded from the same site in random combinations. The videos were selected from the categories of 'bird,' 'culture,' 'earth,' 'fish,' 'food,' 'mammal,' 'nature,' 'plant,' 'vehicle,' and 'VFX'. Note that the movie and audio in each short clip are unrelated, due to such a way the stimuli were created. This is in contrast to the documentary stimuli, where the two are matched.

C. EEG data acquisitions and preprocessing

EEG recordings were conducted using a 32-channel Flex Gel Sensor Kit (Emotiv Inc., USA) from participants under the above-mentioned stimulus presentation. The electrodes were placed in the following positions according to the International

10-20 system: 'Cz', 'Fz', 'Fp1', 'F7', 'F3', 'FC1', 'C3', 'FC5', 'FT9', 'T7', 'CP5', 'P3', 'PO9', 'O1', 'Pz', 'Oz', 'O2', 'PO10', 'P8', 'P4', 'CP2', 'CP6', 'T8', 'FT10', 'FC6', 'C4', 'FC2', 'F4', 'F8', 'Fp2'. The reference electrodes were placed on both earlobes.

The preprocessing was carried out using EEGLab [23] on Matlab (MathWorks, U.S.), and the following steps were performed. First, a band-pass filter was applied to the data between 1 and 40 Hz using a Hamming windowed sinc FIR filter. Then, independent component analysis (ICA) was used to identify and remove artifacts from eye movements and blinks. After that, a re-referencing process was applied, which refers to the average value of the electrodes. Then, time-frequency analysis was performed using the wavelet method to calculate the time-series of the power spectral density. Finally, those time-series of the power spectral density of each EEG channel was normalized to a z-score with an average of 0 and a variance of 1 for each session.

D. Audio processing using SSAST

In this study, we performed a correspondence between a DNN named Self-Supervised Audio Spectrogram Transformer (SSAST) [24] and EEG brain activities. SSAST is a vision Transformer-based model that gives an embedding representation to audio spectrogram inputs. The input is

divided into 16×16 patches, and after linear projection, position embedding is added. This is input into a Transformer Encoder with 768 embedding dimensions, 12 layers, and 12 heads. Pre-training was done in a self-supervised learning framework. This consisted of discriminating and reconstructing masked input patches. In this study, we used a pre-trained model named SSAST-Base-Patch-400. This model has been trained on two datasets, AudioSet [25] and LibriSpeech [26].

By moving the time window in 0.1-second increments on the audio used in the EEG experiment, we extracted 3-second sub-clips. In the case of short clips, the period from 1 second to 8 seconds after the beginning of the clip was targeted for this sub-clipping. This resulted in 41 sub-clips being created from each short clip, giving a total of 19,352 sub-clips. In the case of the documentaries, the sub-clips were created from 1 second after the beginning to the end. This resulted in 10,804, 12,763, 10,090, and 14,170 sub-clips being created from each file. We then input the spectrograms extracted from these sub-clips into SSAST to obtain outputs that would be used as the embedded representations of the auditory stimuli.

E. Prediction of the EEG activity from the SSAST responses

Principal component analysis (PCA) was applied to these embedded representations, and the dimensions were reduced so that the cumulative proportion of variance was 85% and 95% for documentaries and short clips, respectively. Each principal component that was adopted was normalized to a z-score with a mean of 0 and a variance of 1. The EEG power spectral density of the interval corresponding to the last 1.5 seconds of each audio sub-clip was averaged. Furthermore, this was divided into the low theta (4-6 Hz), high theta (6-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-40 Hz) bands, and averaged within each band. From here on, we will consider these as brain activity responses to the sub-clip. Multiple regression analysis was performed using the reduced-dimensionality embedded representations as explanatory variables and the brain activity responses as dependent variables. For the documentary videos, sub-clips extracted from the second video were used as test data, and for the short clips, sub-clips extracted from 15 randomly selected videos were used as test data. The others were used for training. Training for the regressor was conducted for each participant.

IV. RESULTS

We investigated the correspondence between brain activity and the audio-processing DNN (SSAST) using EEG responses to two types of audio-visual videos. The main difference between them is that the audio and movie are matched in the documentary videos, while they are unrelated in the short clip videos.

A. Documentary Video-evoked EEG activity

We input audio sub-clips extracted from documentary videos into the SSAST and obtained their embedded representations as their outputs. Using these embedded representations, we predicted the EEG power spectrum densities recorded from participants under the same video presentation using multiple regression. We evaluated the

prediction performance using a one-sided test based on the correlation analysis of the predicted and actual EEG power spectrum densities for the test data. The statistics obtained for each participant were integrated using the Stouffer's Z-score method [27] and tested. Multiple comparison correction was performed using the Benjamini-Hochberg method [28].

As Figure 2A shows, statistically significant predictability was observed especially in the low frequency bands (theta to alpha) for the wide range of electrodes. On the other hand, in the high-frequency bands, the predictability was not significant for the majority of electrodes. When comparing the regions where the predictability was significant and those where it was not, the average correlation coefficients were not necessarily lower in the latter. This is because the variance of the correlation coefficients between participants was larger in the latter. In other words, the predictability of responses in the low-frequency range was to some extent consistent among participants, but in the high-frequency range it was rather inconsistent. Thus, the frequency dependence of the results of the test is due to differences in variance rather than differences in means, which, not surprisingly, could not be captured by a two-way repeated measures ANOVA (Figure 2B and Table I). Also, no dependence of predictability on electrode locations was observed (Figure 2C).

B. Short Clip-evoked EEG activity

The same analysis as above was performed for the audio sub-clips extracted from the short clips (Figure 3). Unlike the documentary case, there was no clear frequency band dependence in the result of the test of predictability. On the other hand, a dependence on electrode position was observed. In particular, predictability was lower in the occipital to parietal lobes. This electrode position dependence was also confirmed by a two-way repeated measures ANOVA (Table II). In addition, comparing Figures 2A and 3A, it appears that the predictability in the case of short clips is less biased towards the low frequency band than that in the case of documentaries. Rather, it seems that the predictability is lower in the low frequency band. We will discuss these points in the 'Discussion' section.

TABLE I. RESULT OF ANOVA (DOCUMENTARIES)

	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Channel	31/124	0.00577	0.384	0.999
Band	4/16	0.00989	0.387	0.814
Channel x Band	124/496	0.00195	0.842	0.877

TABLE II. RESULT OF ANOVA (SHORT CLIPS)

	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Channel	31/124	0.0152	1.564	0.0451
Band	4/16	0.00662	0.110	0.977
Channel x Band	124/496	0.00246	0.706	0.990

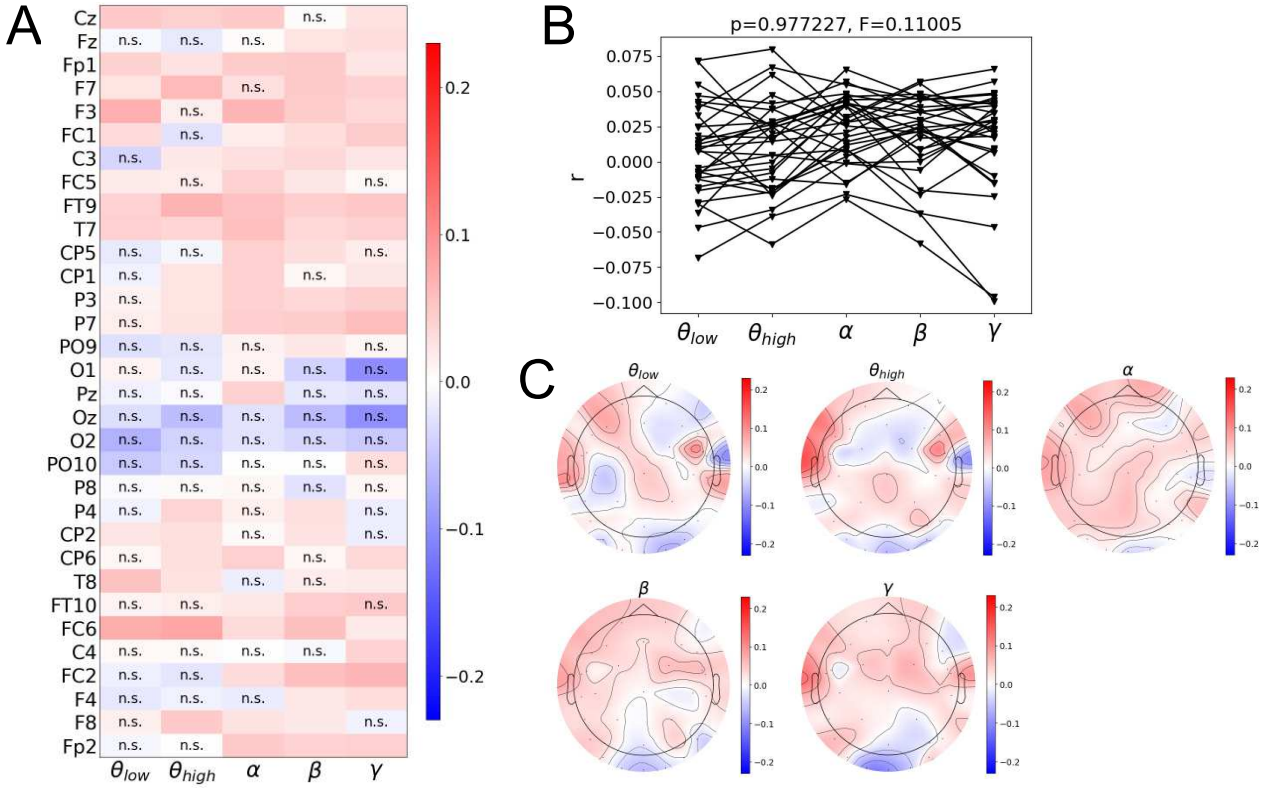


Figure 3: Prediction of brain activity induced by the short clip stimuli from SSAST outputs. (A) Correlation coefficients between experimental and predicted values are shown using a color code. (B) Frequency band dependence of the correlation coefficient between the experimental and predicted values. The vertical axis shows the average correlation coefficient for all participants. Each line corresponds to an EEG channel. (C) Topographic plot of the correlation coefficient between the experimental and predicted values. The color code indicates the correlation coefficient values. The notation used is the same as in Figure 2.

V. DISCUSSION

In this study, we aimed to model EEG brain activity induced by naturalistic audio-visual video stimuli using DNNs. In our previous study, we showed that movie-processing DNNs can capture high-frequency EEG activity well [18]. This result was neurophysiologically plausible, but considering the final goal of using DNNs as brain simulators to synthesize optimal sensory stimuli for intervening in brain activities and functions, it is important to be able to capture low-frequency EEG activity as well. Therefore, in this study, we analyzed the prediction of EEG activity from the embedded representation obtained from SSAST, which is one of the audio-processing DNNs. As a result, it was found that low-frequency EEG activity could be captured well, especially in the case of the documentary stimuli. Therefore, it was suggested that by integrating audio and movie-processing DNNs and using them as a brain simulator, it would be possible to handle a wide range of frequencies in a complementary manner. Additionally, it was considered that documentary stimuli is probably more promising than short clips for intervention.

The most notable difference in obtained results between the two types of video stimuli (documentaries and short clips) was the dependence of the predictability of brain activity on the electrode location. A significant dependence was only shown in the case of short clips. In particular, the predictability was

lower for electrodes across the occipital and parietal lobes. This is considered to be because the auditory and visual stimuli were matched in the case of documentaries, but were unrelated in the case of short clips. In neurophysiology, it is well known that the occipital lobe is strongly specialized for visual processing [29]. Therefore, it is not surprising that the responses in the occipital lobe could not be predicted from the audio stimuli, as was the case with the short clips. On the other hand, in the case of the documentary, the auditory and visual stimuli were correlated in terms of their content, and it seems likely that the responses in the occipital lobe were predictable.

Furthermore, differences were also observed in the relationship between prediction performance and frequency band for the two types of video stimuli. In short, the predictability of brain activity was higher in the low frequency band for documentaries and in the high frequency band for short clips. This may be explained by differences in attention and vigilance. In the condition where short clips with different themes are presented one after another in a short period of time, it is more likely to maintain attention and vigilance. In addition, in the case of short clips, participants were asked to press a key during the interval between videos, and this could also contribute to maintaining attention and vigilance. In general, when attention and vigilance are high, high-frequency band oscillations become dominant, and when they are low, low-frequency band oscillations become dominant [30]-[32]. It is possible that this is reflected in the results.

VI. CONCLUSION

In this study, to establish DNNs as a brain simulator, we attempted to correspond EEG brain activity during viewing audio-visual videos to the responses of SSAST to audio inputs extracted from them. The results showed that, especially for documentary stimuli, EEG activity in the low-frequency band can be corresponded well. This is complementary to the results in the authors' previous study. By integrating these, it is suggested that it will be possible to design optimal sensory stimuli to intervene in brain activities and functions utilizing DNNs as a brain simulator.

ACKNOWLEDGMENT

This research was partially supported by the Leading Initiative for the Excellent Young Researchers (MEXT, Japan), the Grant-in-Aid for Scientific Research (C) (21K07264) (JSPS, Japan), and the JNNS30 Commemorative Research Grant (Japanese Neural Network Society, Japan). The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

- [1] H. F. Iaccarino, A. C. Singer, A. J. Martorell, et al., "Gamma frequency entrainment attenuates amyloid load and modifies microglia," *Nature*, vol. 540, no. 7632, pp. 230–235, Dec. 2016.
- [2] A. J. Martorell, A. L. Paulson, H. J. Suk, et al., "Multi-sensory gamma stimulation ameliorates Alzheimer's-associated pathology and improves cognition," *Cell*, vol. 177, no. 2, pp. 256–271, Apr. 2019.
- [3] C. Adaikkan, S. J. Middleton, A. Marco, et al., "Gamma entrainment binds higher-order brain regions and offers neuroprotection," *Neuron*, vol. 102, no. 5, pp. 929–943, Jun. 2019.
- [4] T. Kim, B. T. James, M. C. Kahn, et al., "Gamma entrainment using audiovisual stimuli alleviates chemobrain pathology and cognitive impairment induced by chemotherapy in mice," *Sci. Transl. Med.*, vol. 16, no. 737, eadf4601, Mar. 2024.
- [5] T. Black, B. W. Jenkins, R. B. Laprairie, and J. G. Howland, "Therapeutic potential of gamma entrainment using sensory stimulation for cognitive symptoms associated with schizophrenia," *Neurosci. Biobehav. Rev.*, vol. 161, 105681, Jun. 2014.
- [6] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 23, pp. 8619–8624, Jun. 2014.
- [7] U. Güçlü, and M. A. J. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *J. Neurosci.*, vol. 35, no. 27, pp. 10005–10014, Jul. 2015.
- [8] G. W. Lindsay, "Convolutional neural networks as a model of the visual system: past, present, and future," *J. Cogn. Neurosci.*, vol. 33, no. 10, pp. 2017–2031, Sep. 2021.
- [9] P. Bashivan, K. Kar, and J. J. DiCarlo, "Neural population control via deep image synthesis," *Science*, vol. 364, no. 6439, eaav9436, May 2019.
- [10] E. Y. Walker et al., "Inception loops discover what excites neurons most using deep predictive models," *Nat. Neurosci.*, vol. 22, no. 12, pp. 2060–2065, Dec. 2019.
- [11] G. Tuckute et al., "Driving and suppressing the human language network using large language models," *Nat. Hum. Behav.*, vol. 8, pp. 544–561, Jan. 2024.
- [12] E. Başar, and B. Güntekin, "A review of brain oscillations in cognitive disorders and the role of neurotransmitters," *Brain Res.*, vol. 1235, pp. 172–193, Oct. 2008.
- [13] E. Başar, "Brain oscillations in neuropsychiatric disease," *Dialogues Clin. Neurosci.*, vol. 15, no. 3, pp. 291–300, Sep. 2013.
- [14] D. H. Mathalon, and V. S. Sohal, "Neural oscillations and synchrony in brain dysfunction and neuropsychiatric disorders: it's about time," *JAMA Psychiatry*, vol. 72, no. 8, pp. 840–844, Aug. 2015.
- [15] S. M. McClintock et al., "Consensus recommendations for the clinical application of repetitive transcranial magnetic stimulation (rTMS) in the treatment of depression," *J. Clin. Psychiatry*, vol. 79, no. 1, 16cs10905, Jan./Feb. 2018.
- [16] O. Elyamany, G. Leicht, C. S. Herrmann, and C. Mulert, "Transcranial alternating current stimulation (tACS): from basic mechanisms towards first applications in psychiatry," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 271, pp. 135–156, Feb. 2021.
- [17] M. Wischniewski, I. Alekseichuk, and A. Opitz, "Neurocognitive, physiological, and biophysical effects of transcranial alternating current stimulation," *Trends Cog. Sci.*, vol. 27, no. 2, pp. 189–205, Feb. 2023.
- [18] H. Kurashige, and J. Kaneko, "Correspondence between the video-learning deep neural networks and EEG brain activity during naturalistic video viewing," in *7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS)*, vol. 7, pp. 200–207, 2022.
- [19] A. T. Gifford, K. Dwivedi, G. Roig, and R. M. Cichy, "A large and rich EEG dataset for modeling human visual object recognition," *Neuroimage*, vol. 264, 119754, Dec. 2022.
- [20] D. Li, C. Wei, S. Li, J. Zou, H. Qin, and Q. Liu, "Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion," *arXiv:2403.07721*, 2024.
- [21] J. Millet, C. Caucheteux, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J. R. King, "Toward a realistic model of speech processing in the brain with self-supervised learning," *Adv. Neural Inf. Process Syst.*, vol. 35, pp. 33428–33443, Dec. 2022.
- [22] G. Tuckute, J. Feather, D. Boebinger, J. H. McDermott, "Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions," *Plos. Biol.*, vol. 21, no. 12, e3002366, Dec. 2023.
- [23] 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.
- [24] Y. Gong, C. I. Lai, Y. A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10699–10709, 2022.
- [25] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, and R. C. Moore, "Audio Set: An ontology and human-labeled dataset for audio events," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780, 2017.
- [26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.
- [27] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr, *The American soldier: Adjustment during army life. (studies in social psychology in world war ii)*, vol. 1. Princeton, NJ: Princeton Univ. Press, 1949.
- [28] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995.
- [29] K. Grill-Spector, and R. Malach, "The human visual cortex," *Annu. Rev. Neurosci.*, vol. 27, pp. 649–677, Jul. 2004.
- [30] B. S. Oken, M. C. Salinsky, and S. M. Elsas, "Vigilance, alertness, or sustained attention: physiological basis and measurement," *Clin. Neurophysiol.*, vol. 117, no. 9, pp. 1885–1901, Sep. 2006.
- [31] A. Martel, S. Dähne, and B. Blankertz, "EEG predictors of covert vigilant attention," *J. Neural Eng.*, vol. 11, no. 3, 035009, May. 2014.
- [32] T. M. Curley, L. Borghetti, and M. B. Morris, "Gamma Power as an Index of Sustained Attention in Simulated Vigilance Tasks," *Top. Cogn. Sci.*, vol. 16, no. 1, pp. 113–128, Jan. 2024.