2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Decoding auditory attention from EEG using a convolutional neural network*

Winko W. An[1†], Alexander Pei[1†], Abigail L. Noyce[2], and Barbara Shinn-Cunningham[1,2]

Abstract— **Brain-computer interface (BCI) systems allow users to communicate directly with a device using their brain. BCI devices leveraging electroencephalography (EEG) signals as a means of communication typically use manual feature engineering on the data to perform decoding. This approach is time intensive, requires substantial domain knowledge, and does not translate well, even to similar tasks. To combat this issue, we designed a convolutional neural network (CNN) model to perform decoding on EEG data collected from an auditory attention paradigm. Our CNN model not only bypasses the need for manual feature engineering, but additionally improves decoding accuracy ($\sim$77%) and efficiency ($\sim$11 bits/min) compared to a support vector machine (SVM) baseline. The results demonstrate the potential for the use of CNN in auditory BCI designs.**

I. INTRODUCTION

Electroencephalography (EEG), a noninvasive, mobile, and low cost neuroimaging technique, has become a popular method for developing brain-computer interfaces (BCIs) [1]. Many successful BCI systems have been built around visual attention: when users are asked to focus on a particular visual object, their attentional state can be decoded from EEG signatures such as evoked responses and oscillations. Due to the strength and robustness of visual responses in EEG, these visual paradigms can achieve high decoding accuracy and transmission efficiency. For example, Lin et al. reported an average information transfer rate (ITR) of 20.26 bits/min in their BCI system built on visual event-related potentials (ERPs, [2]).

Visual BCIs require the deployment of visual attention, which is not always feasible in real-life scenarios like while walking or driving, or for users with visual impairment. An alternative solution using a different sensory modality is therefore desirable. Previous BCI studies have attempted to decode auditory attention from EEG signals. Kim et al. [3] used two streams of modulated signal with a constant-frequency carrier as the stimuli and decoded users' attention from their auditory steady-state response (ASSR). Kaongoen and Jo [4] developed a hybrid auditory BCI paradigm combining ASSR and ERP. Considering that these modulated signals are not particularly pleasant and can cause user fatigue, researchers have also explored using more user-friendly stimuli, such as drip-drop sounds [5], sequences of tones [6], and music [7], in their BCI design. However, the improved user-friendliness was achieved at the cost of system efficiency — none of these studies yielded an ITR over 3 bit/min. We recently reached a balance between these two goals [8]. We directed users' attention to spatialized human-voiced syllables, trained a support vector machine (SVM) with time-frequency measures of EEG, and achieved high decoding throughput ($\sim$10 bits/min).

One possible way to improve the results in [8] is to adopt a deep learning approach, such as a convolutional neural network (CNN). As opposed to conventional machine learning algorithms like SVM, CNNs do not depend on hand-crafted features for classification. Instead, it automatically learns kernel functions through training, which can help extract features that differentiate multiple classes. CNNs have been widely used in computer vision and more recently in general EEG studies [9], but have not been popularly used in auditory BCIs. In this study, we explored the efficacy of CNNs in decoding auditory attention by comparing with a SVM baseline. We also examined the correlation between CNN decoding and behavioral performance to find a possible cause of the observed individual differences in decoding results.

II. METHODS

A. Participants

Thirty adults with normal hearing (21.95 ± 4.95 years old, 15 female) were recruited for this study. The Institutional Review Board of Boston University approved this study. Participants were briefed and consented before partaking in this study, and were compensated for their time.

B. Experiment

Subjects sat in a soundproof booth while wearing a pair of insert earphones (ER1, Etymotic Research). The sound stimuli consisted of syllables /ba/, /da/ and /ga/ spoken by native English speakers with varying pitch. To spatialize the sound stimuli, the audio waveforms were convolved with head-related transfer functions provided by the Media Lab, MIT [10]. The simulated locations were center, 30° from the left (L30) or right (R30), or 90° from the left (L90) or right (R90, Fig. 1a), in the horizontal plane.

The trial began with a one-second visual cue (VC). The VC "Space" indicated that the subject should perform spatial attention, while the VC "Relax" required no attention from

*This work was supported by the Office of Naval Research (Project ID: N00014-18-1-2069).

†Contributed equally to this work

[1]Winko W. An (email: wenkanga@andrew.cmu.edu), Alexander Pei (email: apei2@andrew.cmu.edu) and Barbara Shinn-Cunningham (email: bgsc@andrew.cmu.edu) are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2]Abigail L. Noyce (email: anoyce@andrew.cmu.edu) and Barbara Shinn-Cunningham are with the Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

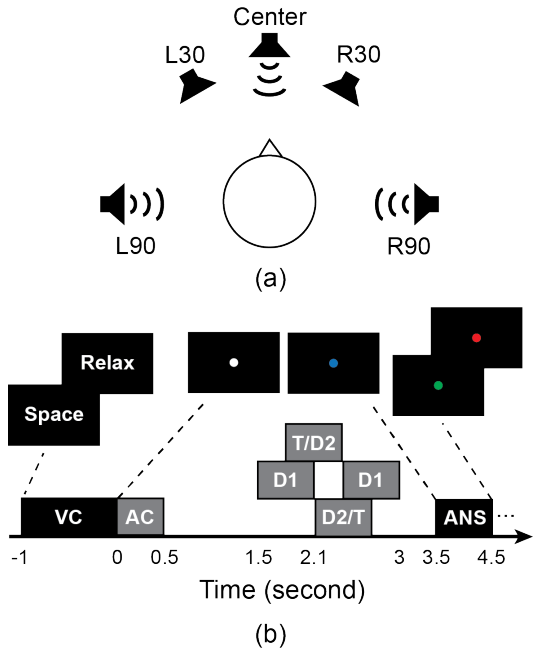

Fig. 1. (a) (Adapted from [8] with the authors' permission) Spoken syllables were spatialized to center, 30° left (L30), or right (R30), and 90° left (L90), or right (R90), always in the horizontal plane. This figure shows one possible scenario where sounds come from L90, R90 and center. (b) Illustration of the events within a trial. A visual cue (VC) was followed by an auditory cue (AC). A 4-syllable mixture was played 1 second after the AC. Participants were asked to respond when the fixation dot turned blue. A green or red dot at the end of the trial provided feedback.

the subject (a third condition, "Talker", is not reported here.) After the VC ended, a 500 ms auditory cue (AC) was played. For "Space" attention trials, the AC was a spatialized /a/ syllable, coming from either L90 or R90. In the "Relax" trials, the /a/ syllable came from the midline. 1000 ms after the AC, a 4-syllable mixture consisting of permuted syllables /ba/, /da/, and /ga/ was played. Each syllable was 600 ms in duration and had 300 ms delays between each subsequent syllable onset. The first and last syllables were distractors (D1), which came from the center. The second and third syllables were either another distractor (D2) or the target syllable (T). The target came from the same location as the AC, while D2 came from a location different than the target. For the "Space" trials, subjects were required to report the target using a key press ("1" for /ba/, "2" for /da/, and "3" for /ga/). During "Relax" trials, subjects were asked to passively listen and report a random syllable. Visual feedback was provided indicating correct responses.

In total, subjects completed 756 trials for the entire experiment, which lasted for approximately 2 hours. Each trial had variations in location and pitch of the talkers, ordering of the syllables, and attention type; only a subset of the overall data (i.e., trials that required spatial or no attention) was used in this study. We collapsed all spatial attention trials into one condition (288 trials), and all no-attention trials into another condition (252 trials) to perform binary classification. In both conditions, the exact same stimuli were presented; the only difference between the conditions was the instruction for the task.

C. EEG processing

EEG was collected using a 64-channel Biosemi system sampled at 2048 Hz. Raw EEG data were bandpass filtered (0.1 – 50 Hz), and were then downsampled to 256 Hz. As opposed to the previous study, which used independent component analysis (ICA) for artifact removal, we used artifact subspace removal (ASR) to remove artifacts because ASR is more feasible during real-time BCI decoding [11].

D. Feature extraction for support vector machine

Based on prior knowledge about the neural signatures of spatial attention [12], we used both time and frequency representations of the EEG data for decoding. The data for each trial were cropped to contain only the time window from 1.5 to 2.7 seconds after the AC, which we expected to contain the critical neural signatures of interest while the subject is actively performing attention. Continuous wavelet transforms (CWT) were used to generate a spectro-temporal representation of the time-series data. A Morlet wavelet with $\omega_0 = 6$ was used as the wavelet base. Normalization was done to have unit total energy at all scales [13]. The CWT coefficients were then collapsed into five distinct frequency bands that are known to contain signatures of cognitive processes: delta (2 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 14 Hz), beta (14 – 30 Hz) and gamma (30 – 50 Hz). This process yielded a multidimensional time-series for each channel consisting of the channel voltage and the magnitude of the wavelet coefficients in each frequency band. To reduce the data dimensionality and computational demands, data were binned into 100 ms windows. The resulting time-series matrix across channels and features was flattened to produce a single vector. A support vector machine (SVM) with a linear kernel was used to decode this data vector for spatial attention conditions vs. no attention conditions. We performed 10-fold cross-validation to generate training and test sets. The decoding accuracy was averaged across the 10 folds. This process was repeated 20 times, for a total of 200 trained models. Each subject was trained and tested independently from other subjects.

E. Convolutional neural network

Instead of manual feature engineering, the preprocessed EEG time-series in the same 1.5 – 2.7 second time-window was input into a CNN. Our architecture consisted of only three convolutional layers to avoid overfitting, given that our data set is relatively small. Average pooling layers were interwoven between convolutional layers to further reduce the dimensionality of the input. The resulting output of the convolutional layers was flattened and fed through fully connected layers followed by rectified linear unit (ReLU) layers to get a single prediction of the binary class. We additionally used dropout layers to assist with overfitting. Details about the layers can be seen in Fig. 2.

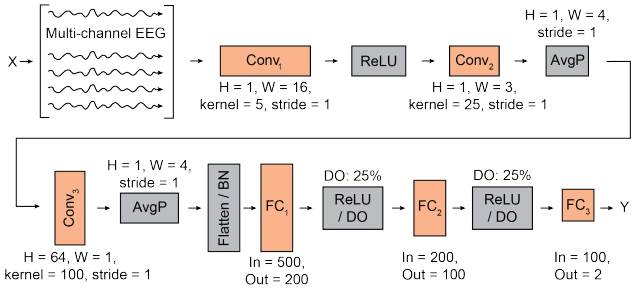

Fig. 2. The CNN architecture used in this study.
Conv – convolutional layer; H – height of kernal; W – width of kernal; ReLU – rectified linear unit; AvgP – average pooling layer; BN – batch normalization layer; FC – fully connected layer; DO – dropout layer

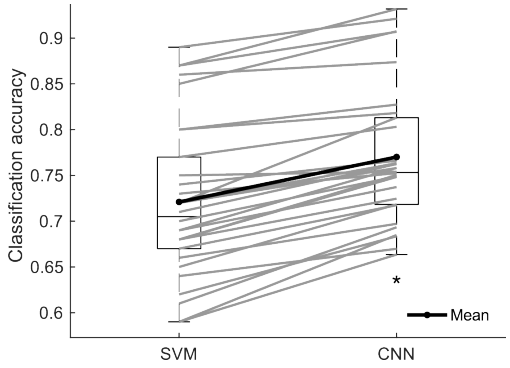

Fig. 3. SVM and CNN classification results. Each gray line represents data from one subject. CNN yielded a significantly higher average classification accuracy than SVM. *p<0.001

The training schemes for the CNN differed slightly from that used for the SVM to avoid overfitting due to overtraining. 10-fold cross validation was performed with 20 samples from each condition (40 samples in total) being held out as the validation set in each fold. The CNN was then trained for 40 epochs, and the model with the lowest validation loss across the 40 epochs was used as the final model to classify the testing set. The rationale is that if a model is overtrained in late epochs, the overfitting would lead to an increase in validation loss. By choosing the model with the lowest validation loss, we are technically stopping the training process before the model becomes too complicated to generalize properly, and thus avoiding overfitting. 20 iterations of random initialization were performed for this 10-fold cross validation. The testing accuracy was averaged across these iterations to estimate the classification performance of the model. We used a cross-entropy loss function, Adam optimizer with a learning rate of 0.0001, a lambda weight decay of 0.01, and a batch size of 50.

III. Results and Discussion

A. Classification accuracy

The average classification accuracy of the SVM approach was 72.10% (Tab. I), a slight drop from the result (∼75%) in [8], where ICA was adopted for artifact removal. Because ASR requires much less time to process than ICA and can be used in a real-time manner, it seems reasonable to replace ICA with ASR in a BCI system design for real-life applications.

The CNN method proposed in this study significantly improved the classification performance compared to the SVM approach (paired t-test, p<0.001, Fig. 3). It achieved a 77.01% decoding accuracy; moreover, each individual subject showed a performance gain over SVM. Given that attention was decoded from only 1.2 seconds of data, the proposed BCI system is highly efficient. The average equivalent ITR [14] is 11.11 bits/min, and the best ITR among all participants is 32.03 bits/min, on par with some visual BCI paradigms. In the future, we will attempt other advanced machine learning methods, such as convolutional long short-term memory (ConvLSTM) and adaptive learning [15] to seek for even better classification performance.

TABLE I
Classification accuracy & information transfer rate (ITR)

Classifier	Average accuracy	Average ITR (bits/min)	Best ITR (bits/min)
SVM	72.10%	7.30	25.00
CNN	77.01%	11.11	32.03

B. Performance gain with CNN

The gain in classification accuracy of CNN over SVM varied across participants. Fig. 4 shows that this improvement is strongly and negatively correlated with the SVM classification results (ρ=-0.541, p=0.002). The CNN method seems to have benefited subjects with a lower decoding score more than those with a higher one, and thus reduced the variability in decoding accuracy across subjects. One possible reason is that the SVM accuracy is low in some subjects not only because there is less distinguishing information in their EEG signals, but because such information is not reliably extracted from EEG using the CWT method. The CNN approach does not rely on hand-crafted features, but rather learns through training what features to use. It may help preserve information that is present, but does not get represented in generic, hand engineered features. The CNN approach thus may be especially beneficial to participants with a low SVM score.

C. Correlation with behavioral performance

The participants exhibit a wide range of behavioral performance in this study — some nearly achieved a perfect score in the attention task, while some others answered correctly in less than 70% of the trials. Interestingly, we observed a strong positive correlation between the participants' behavioral performance and their attention decoding accuracy using CNN (ρ=0.583, p<0.001, Fig. 5). This suggests that the variance in individual CNN classification results shown in Fig. 3 can be partially explained by how well a participant performed in the attention task. If the main reason for giving an incorrect response is that a subject's attentional focus drifted, the proposed BCI system has the potential to achieve

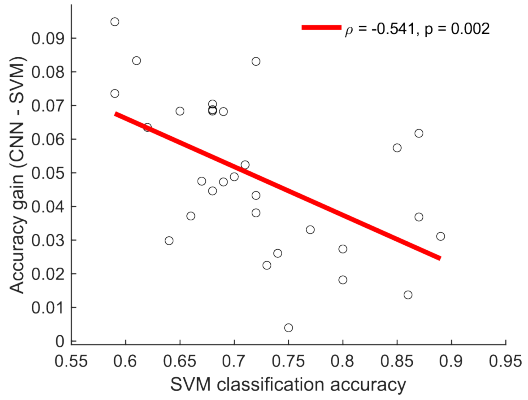

Fig. 4. The gain in classification accuracy from using CNN over SVM is negatively correlated with the SVM classification results. Each circle represents data point of one subject.

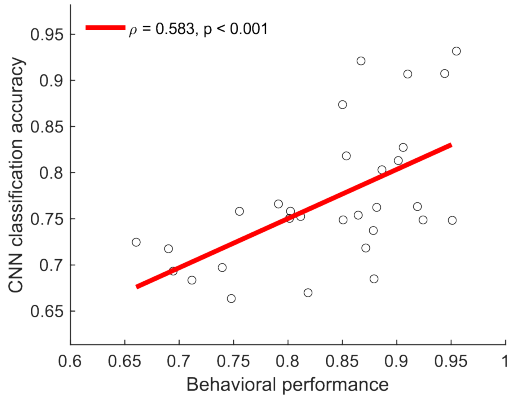

Fig. 5. The CNN classification accuracy is positively correlated with the subject's behavioral performance in the attention task.

even better accuracy and efficiency if the user is always fully engaged and motivated, which is more likely during real-life applications.

IV. CONCLUSIONS

This study proposed a method to decode auditory attention from single-trial EEG for the purpose of building a BCI system. We adopted a subspace-based artifact removal pipeline, which can process signals in a real-time manner. The CNN approach yielded high classification accuracy and efficiency, outperforming a SVM baseline as well as previous studies. The CNN decoding results are strongly correlated with the participants' behavioral performance in the attention task, suggesting a possible improvement in decoding, when used in real-life applications, where users are highly motivated.

ACKNOWLEDGMENT

The authors would like to thank Mr. Yangyang Xia (Department of Electrical and Computer Engineering, Carnegie Mellon University) for his advice on the proposed CNN architecture.

REFERENCES

[1] Reza Abiri, Soheil Borhani, Eric W. Sellers, Yang Jiang, and Xiaopeng Zhao, "A comprehensive review of EEG-based brain-computer interface paradigms," *Journal of Neural Engineering*, vol. 16, no. 1, pp. 011001, 2019.

[2] Zhimin Lin, Chi Zhang, Ying Zeng, Li Tong, and Bin Yan, "A novel P300 BCI speller based on the Triple RSVP paradigm," *Scientific Reports*, vol. 8, no. 1, pp. 3350, dec 2018.

[3] Do Won Kim, Jae Hyun Cho, Han Jeong Hwang, Jeong Hwan Lim, and Chang Hwan Im, "A vision-free brain-computer interface (BCI) paradigm based on auditory selective attention," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2011, pp. 3684–3687, IEEE.

[4] Netiwit Kaongoen and Sungho Jo, "A novel hybrid auditory BCI paradigm combining ASSR and P300," *Journal of Neuroscience Methods*, vol. 279, pp. 44–51, 2017.

[5] Minqiang Huang, Jing Jin, Yu Zhang, Dewen Hu, and Xingyu Wang, "Usage of drip drops as stimuli in an auditory P300 BCI paradigm," *Cognitive Neurodynamics*, vol. 12, no. 1, pp. 85–94, 2018.

[6] Winko W. An, Hakim Si-Mohammed, Nicholas Huang, Hannes Gamper, Adrian KC Lee, Christian Holz, David Johnston, Mihai Jalobeanu, Dimitra Emmanouilidou, Edward Cutrell, Andrew Wilson, and Ivan Tashev, "Decoding auditory and tactile attention for use in an EEG-based brain-computer interface," in *2020 8th International Winter Conference on Brain-Computer Interface (BCI)*. feb 2020, pp. 1–6, IEEE.

[7] Winko W An, Barbara Shinn-cunningham, Hannes Gamper, Dimitra Emmanouilidou, David Johnston, Mihai Jalobeanu, Edward Cutrell, Andrew Wilson, Kuan-jung Chiang, and Ivan Tashev, "DECODING MUSIC ATTENTION FROM " EEG HEADPHONES ": A USER-FRIENDLY AUDITORY BRAIN-COMPUTER INTERFACE," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[8] Winko W An, Alexander Pei, Abigail L Noyce, and Barbara Shinn-cunningham, "Decoding auditory attention from single-trial EEG for a high-efficiency brain-computer interface," in *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020, pp. 3456–3459.

[9] Xiang Zhang, Lina Yao, Xianzhi Wang, Jessica Monaghan, David Mcalpine, and Yu Zhang, "A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers," jun 2021.

[10] Bill Gardner and Keith Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone MIT Media Lab Perceptual Computing-Technical Report 280," Tech. Rep., Media Lab, MIT, 1994.

[11] Chi-Yuan Chang, Sheng-Hsiou Hsu, Luca Pion-Tonachini, and Tzyy-Ping Jung, "Evaluation of Artifact Subspace Reconstruction for Automatic Artifact Components Removal in Multi-channel EEG Recordings," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 1114–1121, jul 2020.

[12] Yuqi Deng, Inyong Choi, and Barbara Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, vol. 207, pp. 116360, feb 2020.

[13] Winko W. An, Kin Hung Ting, Ivan P.H. Au, Janet H. Zhang, Zoe Y.S. Chan, Irene S. Davis, Winnie K.Y. So, Rosa H.M. Chan, and Roy T.H. Cheung, "Neurophysiological Correlates of Gait Retraining with Real-Time Visual and Auditory Feedback," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1341–1349, jun 2019.

[14] Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[15] Kuan-jung Chiang, Dimitra Emmanouilidou, Hannes Gamper, David Johnston, Mihai Jalobeanu, Edward Cutrell, Andrew Wilson, Winko W An, and Ivan Tashev, "A Closed-loop Adaptive Brain-computer Interface Framework : Improving the Classifier with the Use of Error-related Potentials .," in *10th International IEEE EMBS Conference on Neural Engineering*, 2021.