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

Auditory Attention Detection with EEG Channel Attention

Enze Su[1,†], Siqi Cai[1,2,†], Peiwen Li[1], Longhan Xie[1,*], and Haizhou Li[2,3]

Abstract— **Auditory attention detection (AAD) seeks to detect the attended speech from EEG signals in a multi-talker scenario, i.e. cocktail party. As the EEG channels reflect the activities of different brain areas, a task-oriented channel selection technique improves the performance of brain-computer interface applications. In this study, we propose a soft channel attention mechanism, instead of hard channel selection, that derives an EEG channel mask by optimizing the auditory attention detection task. The neural AAD system consists of a neural channel attention mechanism and a convolutional neural network (CNN) classifier. We evaluate the proposed framework on a publicly available database. We achieve 88.3% and 77.2% for 2-second and 0.1-second decision windows with 64-channel EEG; and 86.1% and 83.9% for 2-second decision windows with 32-channel and 16-channel EEG, respectively. The proposed framework outperforms other competitive models by a large margin across all test cases.**

I. Introduction

Humans have the ability to distinguish between speakers and to pay selective attention to one speaker in a multi-talker scenario, i.e., cocktail party [1]. However, hearing aids users often experience difficulty of following a target speaker in the presence of noise and other competing speech sources [2]. Are we able to equip the hearing aids with the human ability of selective attention? Recently, neuro-steered hearing prostheses are studied to produce a better experience for people with hearing loss, in which auditory attention is decoded from recordings of brain activity and used to enhance the speech separation for the attended speaker.

Among the brain signals for auditory attention detection (ADD), such as electrocorticographic (ECoG) [3], magnetoencephalography (MEG) [4] and electroencephalogram (EEG) [5], EEG is a more realistic option for brain-computer interface (BCI) applications, because it's cheaper, non-invasive and easier to use. The techniques for EEG-based auditory attention detection can be grouped into linear and non-linear decoders [6].

Linear decoders have been well studied [5], where EEG responses are used to approximate the envelope of the speech attended by listeners, that is then compared with the original speech stimulus to reveal the attended or unattended speaker in a cocktail party scenario. Specifically, the reconstructed speech envelope from the cortical responses to a mixture of speakers is dominated by the salient spectral and temporal features of the attended speaker [3]. However, the correlation between the reconstructed and the attended envelope is fairly low [7]. A possible explanation is that the human auditory system is inherently non-linear [8] and the linear approach is probably not the best way to model the complex and dynamic nature of the brain [9]. Furthermore, the speech envelope reconstruction algorithm is not systematically optimized, e.g., jointly trained with the classifier, for auditory attention detection.

Recently, non-linear decoders have been studied to understand the complex and highly non-linear nature of auditory processes in the human brain, that show superior performance to linear decoders [6], [7], [10], [11], [12]. In this paper, we follow the CNN-based non-linear approach [10], [11], [12] with a particular focus on low-latency settings. We note that the placement positions of electrodes reflect the activities of the related brain areas. Furthermore, some EEG channels are more informative than others in terms of informing the decision-making process in the brain [13], [14]. At the same time, the distribution of effective channels may vary from subject to subject.

We propose a channel attention mechanism that predicts a channel mask on the fly. The channel mask corresponds to a spatial map of the EEG electrodes, that gives a differentiated weight to each of the EEG channels. An element in the mask is a continuous value, as opposed to on-off channel selection, that modulates the contribution of each EEG channel for optimal auditory attention performance. Such a channel mask may vary with the attended speaker, the speech content, the acoustic environment, the listening subjects, and so on. The question is how to devise a mechanism that dynamically predicts the mask according to each speech-EEG pair.

As far as we know, this is the first study on a channel attention mechanism for EEG-based auditory attention detection. The rest of the paper is organized as follows. Section II presents the channel attention mechanism and the CNN classifier. Experimental setup and the results are summarized in Section III. Finally, Section IV concludes the study.

II. Auditory Attention Detection with Channel Attention

We study a CNN classifier with channel mask (CM) for AAD, which is referred to as CNN-CM hereafter, as illustrated in Fig. 1. The CNN-CM neural architecture consists of a channel attention mechanism and a CNN classifier.

[1]Enze Su, Siqi Cai, Peiwen Li and Longhan Xie are with Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, Guangdong Province, China. Longhan Xie is the corresponding author. enzesu@hotmail.com, elesiqi@nus.edu.sg, lintean@qq.com, and melhxie@scut.edu.cn

[2]Haizhou Li and Siqi Cai are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. haizhou.li@nus.edu.sg

[3]Haizhou Li is also with Machine Listening Lab, University of Bremen, Germany.

† Equal contribution

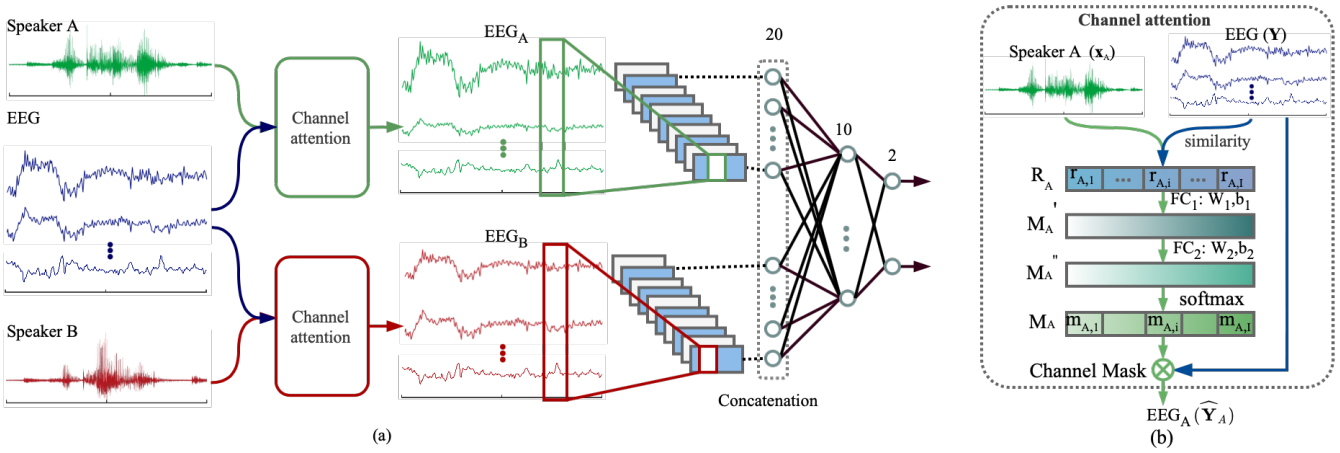

Fig. 1. The proposed CNN-CM neural architecture for auditory attention detection, which is trained as a whole with two output nodes for two attended speakers. (a) Overall architecture; (b) Channel attention mechanism that modulates the input multi-channel EEG signals with respect to speaker A.

During training, the CNN-CM network takes multi-channel EEG signals, speech envelopes A and B as input, and the attention labels as the supervisory signals. The channel attention mechanism is trained to generate the modulated EEG signals via an attention mask, while the CNN classifier is trained to make a detection decision. Both the channel attention mechanism and the CNN classifier are jointly trained for optimal attention decisions.

A. EEG Channel Attention

Humans pay selective attention in many everyday situations, such as auditory attention in cocktail party scenarios. The channel attention mechanism is motivated by such human ability, that seeks to adaptively select important features in machine translation [15], image classification [16], [17] and caption generation [18]. In this study, we would like to dynamically assign weights to channels to reflect the contributions of individual EEG channels for AAD. The channel attention mainly has two properties, 1) it explicitly models the correlation between EEG responses and speech stimuli, and 2) adaptively adjusts the weights of EEG channels.

1) Feature representation: The cosine similarity is chosen to measure the relationship between the speech stimuli and EEG responses [19] in this study, which does not involve any learning parameters. The cosine similarity between two time series, $\mathbf{x} = \{x_n\}_1^N$ and $\mathbf{y} = \{y_n\}_1^N$, is defined as,

$$\mathbf{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{n=1}^{N} x_n y_n}{\sqrt{\sum_{n=1}^{N} x_n^2}\sqrt{\sum_{n=1}^{N} y_n^2}} \tag{1}$$

Let $\mathbf{x}_A$ be the speech envelope, i.e., the auditory stimulus, from speaker A. The correlation between $\mathbf{x}_A$ and the i^{th} channel of EEG signals, $\mathbf{y}_i$, can be denoted as,

$$R_A = \mathbf{similarity}(\mathbf{x}_A, \mathbf{y}_i) \tag{2}$$

where $R_A = [r_{A,1}, \cdots, r_{A,i}, \cdots, r_{A,I}]$ and $r_{A,i}$ denotes the correlation between speaker A and I-channel EEG signals. Similarly we can have R_B for speaker B.

2) Predicting channel mask: In the neural attention mechanism, both R_A and R_B are taken as input by a gating mechanism to produce the channel mask $\{M_s : s \in \{A, B\}\}$. Two fully-connected (FC) layers are adopted to parameterise the gating mechanism to capture the nonlinear interaction among the channels [16]. The resulting channel mask for I-channel EEG signals $\mathbf{Y} = \{\mathbf{y}_i\}_1^I$ can be denoted as $M_s = [m_{s,1}, \cdots, m_{s,i}, \cdots, m_{s,I}]$,

$$M_s = softmax(W_2 \cdot (W_1 \cdot R_s + b_1) + b_2) \tag{3}$$

where a dimensionality-reduction layer with parameter W_1 and bias b_1 with reduction ratio r and *tanh* function as the activation function, and a dimensionality increasing layer with parameter W_2 and bias b_2, and followed by a sigmoid activation. Finally, the neural attention mechanism modulates the input EEG signals by applying the attention mask channel by channel $\hat{\mathbf{y}}_{s,i} = m_{s,i} \times \mathbf{y}_i$, or $\hat{\mathbf{Y}}_s = M_s \bigotimes \mathbf{Y}$ where $\bigotimes$ denotes a point-wise multiplication.

B. Auditory Attention Detection

Convolutional neural network (CNN) has been employed as a classifier for auditory attention detection with state-of-the-art performance [6], [7], [10], [11], [12]. Hence, we adopt the CNN as a backend classifier that takes two sets of modulated EEG signals as the input, one for speaker A and another for speaker B, and decides which speech stimulus is associated with the EEG responses in a binary decision.

As shown in Fig. 1, the neural architecture starts with a convolution layer, which uses a kernel size of 64×9 and a stride of 64×1. The convolution layer has a rectifying linear unit (ReLU) activation function and is followed by an average pooling layer with a 1×256 kernel, and two FC layers with 20 and 10 neurons, respectively. Finally, a softmax output layer is added for binary decision.

During training, we adopt the cross-entropy loss function as the cost function in the adaptive moment estimation algorithm (Adam) [20]. The learning rate is set to 1×10^{-3}. The channel attention mechanism and the CNN classifier are

jointly trained as a single system. At run-time, the system produces two values at the output nodes for decision-making.

III. Experiments and Results

A. Experimental Setup

In this study, experiments were carried out on a public auditory attention detection dataset [21], recorded at KU Leuven, which is referred to as KUL Dataset. Briefly, 64-channel EEG data were recorded from 8 male and 8 female normal-hearing subjects when they listened to two competing speakers and were instructed to attend to one speaker. EEG data were recorded at a sample rate of 8192 Hz using a BioSemi ActiveTwo system. Four Dutch short stories, narrated by different male speakers, were used as speech stimuli through a pair of insert earphones. The whole experiment was split into 8 trials and each trial lasts 6 minutes. The auditory stimuli were presented from 90° to the left and 90° to the right of the subject, respectively. Overall, the EEG data from 16 normal-hearing subjects was collected, and there were 48 minutes of data for each subject.

B. Data Processing

The EEG data of each channel were firstly re-referenced to the mean of the response of all channels. Then, all the EEG data were bandpass filtered between 1 and 50 Hz, and subsequently down-sampled to 128 Hz. The speech stimuli were first passed through a Gammatone filterbank ranging from 150 Hz to 8 kHz. All of the sub-bands were power-law compression with 0.6 [22]. Finally, the speech envelopes were transformed into their respective absolute envelopes by a Hilbert transformation, low-pass filtered with 50 Hz, and down-sampled from 512 Hz to 128 Hz to match the EEG data.

The data set was randomly split into a training set (60%) and a validation set (20%), and a test set (20%). For each partition, data segments were generated with a sliding window, denoted as the decision window, with an overlap of 50%. We maintain a balanced number of speaker A/B attention samples by subject, i.e, a random guess will give a 50% accuracy. All the repetitions were discarded to keep the training, validation, and test set mutually exclusive. We are particularly interested in low-latency attention detection, therefore, we only report the detection accuracy for four short decision windows: 0.1-, 0.5-, 1-, and 2-second. To avoid initialization bias, the experiments of each subject were carried out 10 times with random initialization to report a subject average accuracy.

C. Experiment Results

A comprehensive comparative study was carried out. We re-implemented two reference baselines, namely the stimulus reconstruction (linear) model [5], and the CNN (non-linear) model [10], on KUL Dataset with 64-channel EEG. The difference between our CNN-CM model and the CNN baseline lies in the additional channel attention mechanism.

As shown in Table I, the CNN-CM model significantly outperforms the linear decoder with a large margin for all

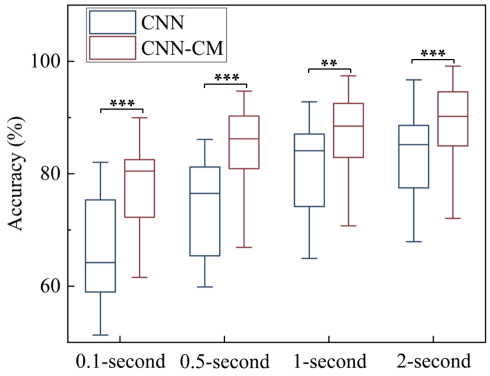

Fig. 2. Auditory attention detection accuracy (%) of CNN-CM and CNN models with 64-channel EEG for different decision windows. Statistically significant differences: $^{**}p < 0.01$, $^{***}p < 0.001$. It is observed that a larger decision window leads to a higher accuracy with a lower variance.

decision windows. These results corroborate with previous studies [7], [10], [11], [12]. It is noted that the detection accuracy increases as we increase the decision window size. Encouragingly, the CNN-CM model has seen a mean accuracy of 77.2% (SD: 8.24) for 0.1-second decision window, which represents an improvement of 15.9% over the linear baseline for 2-second decision window. Moreover, the average detection accuracy of the CNN-CM model exceeded 86% at a temporal resolution of around 1 second, comparable to the human's time lag when switching attention [23]. We are not aware of other decoders that achieve similar accuracy under such low latency settings.

As shown in Fig. 2, the CNN-CM model outperforms the non-linear CNN model in [10] with consistent improvements in AAD accuracy with 5.4% for 1-second and 2-second decision windows, 9.7% for 0.5-second decision window, and 10.7% for 0.1-second decision window, respectively. With the channel attention mechanism, our CNN-CM model significantly outperforms the CNN model (paired t-test: $p < 0.01$).

D. Effect of Electrode Reduction

We would like to examine if the proposed channel attention mechanism is still effective with a reduced number of EEG channels.

Specifically, we reduced the EEG signals from 64-channel to 32-channel and 16-channel following the electrode locations of the international 10/20 system, respectively. As summarized in Table I, though AAD performance decreases with low-density EEG systems [13], the CNN-CM model continues to outperform both linear and non-linear baselines. It is noteworthy that the CNN-CM model with 16-channel EEG achieves even better results than the state-of-the-art CNN model across all different decision windows.

It is a clear advantage to have a low number of EEG electrodes because we considerably simplify data acquisition and reduce the preparation time. Hence, our CNN-CM model is more suitable for neuro-steered hearing prostheses.

IV. Discussion and Conclusions

All experiments have confirmed the effectiveness of the proposed channel attention mechanism. We consider that two

TABLE I

AUDITORY ATTENTION DETECTION ACCURACY (%) AND ITS STANDARD DEVIATION (±) IN A COMPARATIVE STUDY. † DENOTES OUR RE-IMPLEMENTATION OF THE LINEAR MODEL IN [5]. ‡ DENOTES OUR RE-IMPLEMENTATION OF THE CNN MODEL IN [10]. $^{a,\ b,\ c}$ DENOTE THE SIGNIFICANT INCREASE OF AAD ACCURACY OVER THE CNN METHOD [10] WITH $p < 0.05$, $p < 0.01$, AND $p < 0.001$ RESPECTIVELY.

Model	EEG Channels	Decision window (second)			
		0.1	0.5	1	2
Linear [5]†	64	-	55.6	58.1	61.3
CNN [10]‡	64	66.5 ± 9.22	74.6 ± 8.89	81.1 ± 8.45	82.9 ± 8.17
CNN-CM	64	**77.2 ± 8.24**^{c}	**84.3 ± 8.56**^{c}	**86.5 ± 7.99**^{b}	**88.3 ± 7.89**^{c}
CNN-CM	32	**74.9 ± 7.94**^{c}	**81.6 ± 8.13**^{c}	**84.2 ± 8.18**^{a}	**86.1 ± 7.74**^{b}
CNN-CM	16	**72.8± 7.62**^{b}	**79.2 ± 8.16**^{b}	**82.0 ± 8.07**	**83.9 ± 7.92**

main factors have contributed to the significant improvement of the CNN-CM model over the baselines. One is the task-oriented feature representation with a channel attention mask. The attention mechanism dynamically focuses the attention on effective EEG channels, as evidenced by the reduced standard deviation of accuracy across the participating subjects over the CNN baseline in Table I. Another is the joint training between the channel attention mechanism and the CNN classifier, which allows for feature representation and classifier to be optimized for attention detection performance. In future work, we will study auditory attention detection on subjects with hearing loss based on the proposed channel attention mechanism.

ACKNOWLEDGMENT

This research work is supported by Programmatic Grant No. A18A2b0046 and A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

The work by Haizhou Li is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

The work by Longhan Xie is also funded by the National Natural Science Foundation of China (Grant No. 52075177).

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] K. Chung, "Challenges and recent developments in hearing aids: Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms," *Trends in Amplification*, vol. 8, no. 3, pp. 83–124, 2004.

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, p. 233, 2012.

[4] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[5] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[6] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[7] T. de Taillez, B. Kollmeier, and B. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech." *The European journal of neuroscience*, 2017.

[8] P. Faure and H. Korn, "Is there chaos in the brain? I. Concepts of nonlinear dynamics and methods of investigation," *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, vol. 324, no. 9, pp. 773–793, 2001.

[9] H. Korn and P. Faure, "Is there chaos in the brain? II. Experimental evidence and related models," *Comptes rendus biologies*, vol. 326, no. 9, pp. 787–840, 2003.

[10] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, 2018.

[11] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, 2020.

[12] S. Cai, E. Su, Y. Song, L. Xie, and H. Li, "Low latency auditory attention detection with common spatial pattern analysis of EEG signals," *Proc. Interspeech 2020*, pp. 2772–2776, 2020.

[13] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, 2015.

[14] T. Otaiby, F. Abd El-Samie, S. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, 08 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[19] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in Neuroscience*, vol. 13, p. 1267, 2019.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] N. Das, T. Francart, and A. Bertrand, "Auditory Attention Detection Dataset KULeuven," 2019.

[22] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2016.

[23] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach," *Frontiers in neuroscience*, vol. 12, p. 262, 2018.