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

Low-Latency Auditory Spatial Attention Detection Based on Spectro-Spatial Features from EEG

Siqi Cai[1,†], Pengcheng Sun[1,†] Tanja Schultz[2], and Haizhou Li[1,3]

Abstract— **Detecting auditory attention based on brain signals enables many everyday applications, and serves as part of the solution to the cocktail party effect in speech processing. Several studies leverage the correlation between brain signals and auditory stimuli to detect the auditory attention of listeners. Recently, studies show that the alpha band (8-13 Hz) EEG signals enable the localization of auditory stimuli. We believe that it is possible to detect auditory spatial attention without the need of auditory stimuli as references. In this work, we firstly propose a spectro-spatial feature extraction technique to detect auditory spatial attention (left/right) based on the topographic specificity of alpha power. Experiments show that the proposed neural approach achieves 81.7% and 94.6% accuracy for 1-second and 10-second decision windows, respectively. Our comparative results show that this neural approach outperforms other competitive models by a large margin in all test cases.**

I. INTRODUCTION

Humans have the ability to pay attention to a particular sound source or voice, even in multi-talker scenarios [1], that is also called cocktail party. Previous studies have revealed the role of specific neural processes involved and provided neural evidence for auditory attention modulation [2], [3], [4]. With the latest advancements in neuroscience, we are inspired to develop computational models that detect auditory attention as part of brain activities.

Recent findings show that auditory attention in cocktail party scenarios can be decoded from the recordings of brain activity, such as electrocorticography (ECoG) [3], magnetoencephalography (MEG) [5] and electroencephalography (EEG) [4], [6], [7], [8], [9], [10]. Among them, EEG provides a non-invasive and low-cost means of investigating cortical activity with a high temporal resolution, which makes it particularly suitable for brain-computer interface (BCI) applications [11]. Therefore, we are interested in the decoding of auditory attention from EEG signals in this paper.

Most of the studies on auditory attention detection seek to detect the envelope of the speech produced by the attended speaker from brain signals, that is referred to as the speech envelope reconstruction technique. Such technique requires the auditory stimulus, i.e. the clean speech signal recorded in a noise-free environment, to be available [4], [6], [7], [12]. Unfortunately, in real-world applications, such as hearing prostheses or speaker localization, it is unrealistic to obtain such clean speech signals. Inspired by the findings that alpha power is highly associated with spatial attention [13], [14], [15], [16], we hypothesize that we can detect the auditory spatial attention based on brain activities alone, without the need of clean speech envelopes.

Meanwhile, it was shown that the linear EEG decoder requires a very long decision window, with a duration of 10 seconds or more, for a reliable decision on auditory spatial attention [12]. A response delay of 10 seconds is out of the question for applications such as hearing aids. Thus, non-linear decoders with shorter decision windows are of high interest. The latest deep learning techniques provide new ways to understand the complex and highly non-linear nature of auditory processes in the human brain. Non-linear decoders [7], [8], [17] have shown superior performance to linear decoders in several low-latency settings. In this paper, we further the study of a non-linear decoder for low latency auditory spatial attention detection (ASAD).

The contributions of this paper come in three parts: (1) the design, implementation, extraction, as well as combination of spectral plus spatial features from the EEG alpha band to form spectro-spatial feature (SSF), (2) the application of convolutional neural network (CNN) based classification of auditory spatial attention, and (3) the combination of these two components to form a SSF-CNN system for ASAD, as illustrated in Fig. 1. The final SSF-CNN system is experimentally evaluated. We show that SSF-CNN outperforms other competitive models in both accuracy and latency.

II. AUDITORY SPATIAL ATTENTION DETECTION

A. Spectro-Spatial Feature (SSF)

The topological distribution of oscillatory cortical activity in the alpha frequency band is closely related to the location of spatial focus of attention [15], that prompts us to study a novel spectro-spatial feature extraction technique. We study the feature extraction in two stages as shown in Fig. 1.

First, a fast Fourier transform (FFT) is performed on the continuous time series of each electrode to obtain the power spectrum of the EEG signal. The average of squared absolute value in the frequency band is then taken as the individual measurement value of each electrode.

Second, we propose to convert these measurements of different decision windows into a sequence of 2-D images, so as to take full advantage of the spatial features of EEG signals. In this stage, EEG electrodes are projected from the 3-D space onto a 2-D plane according to the coordinate

[1]Siqi Cai, Pengcheng Sun, and Haizhou Li are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Siqi Cai is the corresponding author. elesiqi@nus.edu.sg, pengcheng.sun@u.nus.edu, and haizhou.li@nus.edu.sg

[2]Tanja Schultz is with Cognitive Systems Lab, University of Bremen, Germany. tanja.schultz@uni-bremen.de

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

† Equal contribution

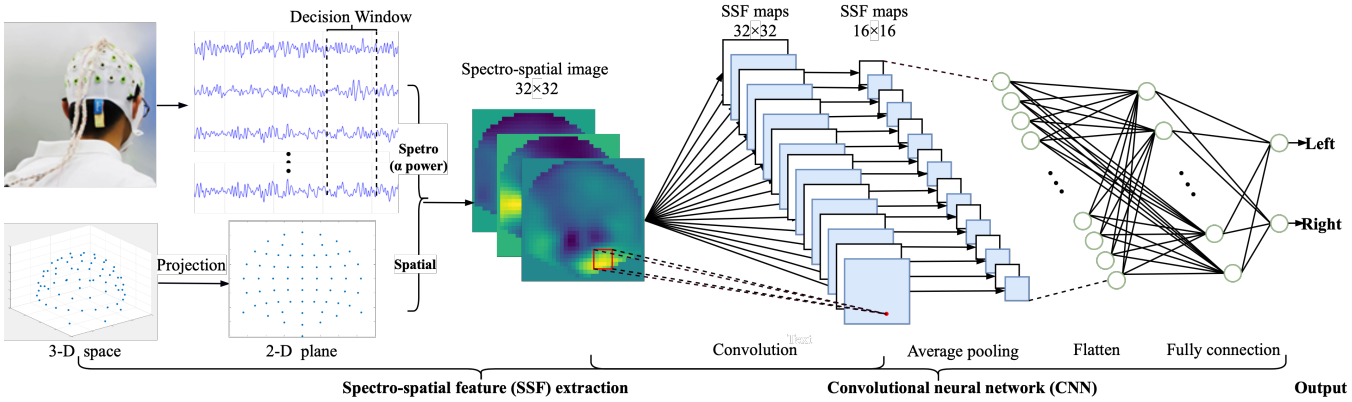

Fig. 1. The proposed convolutional neural network (CNN) with spectro-spatial feature (SSF) for auditory spatial attention detection, that is referred to as SSF-CNN model. The SSF-CNN network is trained to output two values, i.e., 0 and 1, to indicate the spatial location of the attended speaker.

information using Azimuth Equidistant Projection [18]. In practice, we project all points onto a plane tangent to the earth, and divide all the latitude and longitude lines into equal parts, which ensures that all points are accurately spaced and oriented from the center. Considering a human head approximately as a sphere, we select the top point of the head as the tangent point, so as to obtain the projection of the electrode on the 2-D image. The Clough-Tocher interpolant [19], which is based on cubic polynomial, is used to estimate the value of each grid over a 32×32 mesh, which represents the spatial distribution of EEG signals.

In this way, a topographical activity map of EEG signals can be generated, that represents the alpha frequency band within a time window. As the map takes both spectral and spatial information of EEG data into account, it is referred to as the SSF map. We then use the sequence of SSF maps derived from consecutive time windows to reflect the temporal evolution of brain activities, which serve as the input to the subsequent convolutional neural network.

Overall, the proposed SSF extraction facilitates the learning of the topographic specificity of alpha power from EEG signals [15]. We have good reason to expect that SSF is more expressive than the original EEG signals in attention detection. Moreover, it eliminates the need for handcrafting any features.

B. Attention Detection with Convolutional Neural Network

Convolutional neural network (CNN) is a kind of feed-forward neural network, which makes use of convolution and pooling techniques for representation learning and classification decision. The fact that CNN is effective in image recognition [20] leads us to believe that it would also learn and classify well the sequence of topographical EEG maps.

As shown in Fig. 1, the CNN architecture starts with a convolution layer, which uses a kernel size of 3×3 and a stride of 1 with padding. The convolution layer has a rectifying linear unit (ReLU) activation function and is followed by an average pooling layer with a 2×2 kernel and 2 pixels stride, and three fully-connected (fc) layers with 512, 32, and 2 neurons, respectively. The batch normalization is applied to every convolution layer to reduce the effect of the distribution of internal neurons. To avoid overfitting, a dropout layer [21] is applied after the pooling layer and the first fc layer, respectively. Finally, a softmax output layer is added for binary decision. Cross-entropy loss is selected as the cost function, using the root mean square propagation algorithm (RMSProp) [22]. Both the learning rate and decay are set to 1×10^{-3}.

III. Experiments and Results

A. Experimental Setup

We conducted the auditory attention detection experiments on the dataset recorded at KU Leuven [23], denoted as KUL Dataset. Briefly, 64-channel EEG data was collected from eight male and eight female normal-hearing subjects. A subject was instructed to pay attention to one of two competing speakers. The EEG data was recorded with a BioSemi ActiveTwo device at a sampling rate of 8,192 Hz and an electrode positioning that follows the international 10-20 system. Four Dutch short stories, narrated by different male speakers, were used as the stimuli. The auditory stimuli were low-pass filtered with a cut-off frequency of 4 kHz and presented at 60 dB through a pair of in-ear earphones (Etymotic ER3).

The experiment for each subject was split into eight trials of 6 minutes duration. The auditory stimuli were either presented dichotically (one speaker per ear) or with two speakers coming from 90 degrees to the left and 90 degrees to the right of the subject, respectively. The latter stimuli was simulated based on a head-related transfer function (HRTF) filtering. Throughout the experiments, the order of presentation of both conditions was randomized over the different subjects. In total, 8 × 6 min = 48 min of EEG data was collected for each subject, accumulating to 12.8 hours of EEG data for all 16 subjects.

B. Data Preparation

The EEG data of each channel were re-referenced to the average response of the mastoid electrodes, then bandpass-filtered between 8 and 13 Hz, and subsequently down-sampled to 70 Hz. The frequency range was chosen based on non-linear auditory attention detection studies [7], [17].

Finally, EEG data channels were normalized to ensure zero mean and unit variance for each trial.

The data set was randomly split into a training (80%), a validation (10%), and a test set (10%) while preserving the distribution between left/right attention in the three partitions by subject. For each partition, the data segments were generated with a sliding window (referred to as *decision window*) with an overlap of 50%. Thus, for the 0.1-second decision window, the test set resulted in 5,760 decision windows per subject, totaling to 92,160 decision windows for 16 subjects.

C. Experiments with EEG of 64 Channels

The SSF-CNN model is an end-to-end network, which decides between left and right attention for each EEG data segment. As the test set is balanced between left-right attention, the chance-level is 50%. To avoid initialization bias, the accuracy is averaged over 10 runs with random initialization. We report the overall average detection accuracy and the average accuracy per subject for five decision window sizes ranging from 0.1 to 10 seconds. The results are presented in Fig. 2 and show an accuracy of 67.2% (SD: 4.57) for 0.1-second, 81.7% (SD: 5.37) for 1-second, 84.7% (SD: 6.13) for 2-second, 90.5% (SD: 5.71) for 5-second, and 94.6% (SD: 4.37) for the 10-second decision window. Overall, it is apparent that longer decision windows lead to higher detection accuracy. While there are exceptions where longer decision windows do not help, the accuracy trend over window size corroborates with findings in other studies [7], [8], [9], [10]. It is worth noting that SSF-CNN shows a more consistent accuracy trend over window size than the CNN [17] baseline, with a fewer number of exceptions.

To assess the performance of our proposed SSF-CNN system we used two benchmarks, a linear decoder baseline [4] and a non-linear CNN model [17]. For the former, we re-implemented the stimulus (speech envelope) reconstruction model [4], in which the EEG signals approximate the envelope of the attended speech. The reconstructed stimulus is then compared with the original by calculating the correlation between them. Strong correlation indicates the presence of attention. For the latter, we refer to the results reported by Vandecappelle et al. [17], who performed auditory attention detection experiment on the same KUL Dataset. The authors applied a non-linear CNN model with a $[64, T]$ matrix as input, that represents 64 EEG channels and T samples in a decision window. While our SSF-CNN model leverages our proposed spectro-spatial features as input to the CNN.

Table I shows the auditory spatial attention detection accuracy of the proposed SSF-CNN model in comparison to the two benchmark models. The SSF-CNN models consistently outperforms both the linear [4] and the CNN [17] model. The differences become more prominent with shorter window length. In particular, the linear decoder accuracy drops significantly when operating on short decision windows, i.e., 58.1% with 1-second decision window, while the SSF-CNN model (81.7%) maintains the performance level reasonably well. The variation of decoding accuracy with window length is consistent with the literature [7], [9], [17], [16].

TABLE I

ATTENTION DETECTION ACCURACY (%) ON KUL DATASET OF 64 AND 32 CHANNEL EEG (#EEG) FOR FIVE DECISION WINDOW SIZES.

Model	#EEG	Auditory stimulus	Decision window (second)				
			0.1	1	2	5	10
Linear [4]	64	with	-	58.1	61.3	67.5	75.8
CNN [17]	64	without	65.9	80.8	82.1	83.6	85.6
SSF-CNN	64	without	**67.2**	**81.7**	**84.7**	**90.5**	**94.6**
SSF-CNN	32	without	-	**76.1**	**80.1**	**86.2**	**89.4**

We carried out a significance test to confirm that the SSF-CNN model improves significantly (paired t-test, $p = 0.039$) over its CNN counterpart [17]. Since the major difference between the models lies in the EEG feature representation, we believe that the performance improvements are in fact a result of the topographic specificity of alpha power signals, which acts as a spatially selective filter of attention in cocktail party scenarios [13], [14], [15].

It is worth noting that 1-second decision window is close to the time lag required by humans to switch attention [12]. To test the low-latency limit, we further evaluated SSF-CNN and the CNN model with window length that is shorter by an order of magnitude, i.e., 100 millisecond (0.1 second). It is encouraging to see that the SSF-CNN model not only outperforms the CNN model at same window length, but also the linear model with 1-second, and 2-second test windows.

To the best of our knowledge, the SSF-CNN model achieves the best accuracy on KUL dataset with all decision windows ranging from 0.1 to 10 seconds. Since it eliminates the need for a reference auditory stimulus, the proposed SSF-CNN model represents a very appropriate solution for neuro-steered hearing prostheses and other everyday applications, and remains viable even for low-latency requirements.

D. Experiments with EEG of 32 Channels

Results reported so far, relied on 64-channel EEG data. Since a lower number of EEG electrodes has multiple advantages, we reduced the number of electrodes from 64 to 32 channels, following the international 10/20 system [24].

In Table I, we compare the detection accuracy of the SSF-CNN model between 32-channels and 64-channel signals. The detection accuracy for the 32-channel version is 76.1% (SD: 6.84), 80.1% (SD: 7.56), 86.2% (SD: 6.05), and 89.4% (SD: 8.09) for 1, 2, 5, and 10-second decision windows, respectively. While the accuracy of 32-channel data is generally lower than that of 64-channel data, the mean accuracy remains around 80% with a 2-second decision window. In addition, the 32-channel SSF-CNN model outperforms the linear model with 64-channel EEG over all decision windows lengths. From Table II, we also observe that the 32-channel SSF-CNN model compares favorably [17] for longer window sizes. In sum, the proposed SSF-CNN method detects the auditory spatial attention accurately even with a reduced set of EEG channels, which is an important feature for real-world application.

IV. CONCLUSIONS

In this paper we proposed a novel spectro-spatial feature representation for EEG that serves as input into a CNN model

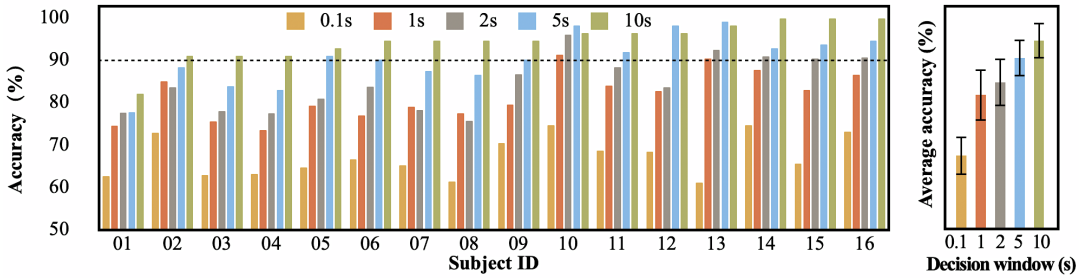

Fig. 2. Auditory spatial attention detection accuracy of the proposed SSF-CNN model with 64-channel EEG over different decision windows and for 16 subjects. The subjects are ranked according to the accuracy for the 10-second decision window. The horizontal dotted line shows a reference point of high accuracy at 90%.

to perform auditory spatial attention detection. The resulting SSF-CNN system consistently and significantly outperforms two benchmark models, a conventional linear model and a state-of-the-art CNN model, over various window lengths. Furthermore, the SSF-CNN achieves encouraging results even with extremely short decision window length and a reduced number of EEG channels. Most importantly, the proposed feature representation does not require any clean reference signal. The combination of these outcomes make the SSF-CNN a highly competitive candidate for real-life applications such as neuro-steered hearing aids.

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 and Tanja Schultz is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

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] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party"," *Journal of Neuroscience*, vol. 30, no. 2, pp. 620–628, 2010.

[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] 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.

[5] 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.

[6] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of neural engineering*, vol. 13, no. 5, p. 056014, 2016.

[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] 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.

[9] 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.

[10] M. Jaeger, B. Mirkovic, M. G. Bleichner, and S. Debener, "Decoding the attended speaker from EEG using adaptive evaluation intervals captures fluctuations in attentional listening," *Frontiers in Neuroscience*, vol. 14, 2020.

[11] H.-J. Hwang, S. Kim, S. Choi, and C.-H. Im, "EEG-based brain-computer interfaces: a thorough literature survey," *International Journal of Human-Computer Interaction*, vol. 29, no. 12, pp. 814–826, 2013.

[12] 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.

[13] M. Wöstmann, B. Herrmann, B. Maess, and J. Obleser, "Spatiotemporal dynamics of auditory attention synchronize with speech," *Proceedings of the National Academy of Sciences*, vol. 113, no. 14, pp. 3873–3878, 2016.

[14] A. Bednar and E. C. Lalor, "Where is the cocktail party? decoding locations of attended and unattended moving sound sources using EEG," *NeuroImage*, vol. 205, p. 116283, 2020.

[15] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *Neuroimage*, vol. 207, p. 116360, 2020.

[16] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Transactions on Biomedical Engineering*, 2020.

[17] 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.

[18] J. P. Snyder, *Map projections–A working manual*. US Government Printing Office, 1987, vol. 1395.

[19] I. Amidror, "Scattered data interpolation methods for electronic imaging systems: a survey," *Journal of electronic imaging*, vol. 11, no. 2, pp. 157–176, 2002.

[20] E. R. De Rezende, G. C. Ruppert, and T. Carvalho, "Detecting computer generated images with deep convolutional neural networks," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2017, pp. 71–78.

[21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[22] T. Kurbiel and S. Khaleghian, "Training of deep neural networks based on distance measures using RMSProp," *arXiv preprint arXiv:1708.01911*, 2017.

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

[24] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalography and clinical neurophysiology*, vol. 66, no. 4, pp. 376–382, 1987.