LSTMs for EEG-based Auditory Attention Decoding

1st René Pallenberg
Institute for Signal Processing
University of Lübeck
Lübeck, Germany

2nd Ann-Katrin Griedelbach
Institute for Signal Processing
University of Lübeck
Lübeck, Germany

3rd Alfred Mertins
Institute for Signal Processing
University of Lübeck
German Research Center for Artificial Intelligence (DFKI)
AI in Biomedical Signal Processing
Lübeck, Germany

***Abstract*—The cocktail party problem occurs when hearing aid users are in an environment with more than one speaker. For these people, it is difficult to follow a particular speaker. One possible solution would be to use electroencephalography signals recorded by the hearing aid to identify which speaker to listen to and amplify it. To improve this detection, we developed a long-short-term-memory-based neural network architecture that detects which of two speakers the person is currently paying attention to. The model was tested with two different preprocessing pipelines on two publicly available datasets. The network achieved 92% accuracy with an input window length of 0.25 seconds, the highest value for the dataset used.**

***Index Terms*—LSTM, Auditory attention, EEG, Attention, CNN**

I. Introduction

According to the World Health Organization's hearing report 2021, 20 % of people will experience some form of hearing loss in their lifetime, of which about 27% will require rehabilitation [1]. To help these people, better and better hearing aids are needed. One problem that hearing-impaired people often encounter is that they have difficulty following a particular speaker when several people are talking at the same time. Many hearing aids amplify the signal that is directly in the direction of the listener's gaze. However, this is not always the person who should be amplified. The basic idea to solve this problem is to use electroencephalography (EEG) signals to identify which of the multiple speakers the person wants to listen to and amplify that speaker.

Two datasets are available in which several subjects had to focus on one of two speakers [2], [3]. The subjects' EEG signals were recorded simultaneously. It was shown that by looking for correlations between the envelopes of the speaker's signals and the EEG signals, it is possible to determine which speaker the subject is attending to. These correlations were found in the frequency range 1-8 Hz [4]. By using canonical correlation analysis (CCA) to extract features for prediction of the attended speaker, accuracies of over 90% were achieved [5]. However, these approaches require window lengths of 30 seconds to achieve this high accuracy [6]. This makes their use in real-time applications such as decoding in hearing aids impossible. To achieve higher accuracy at short windows, methods that do not rely on correlation are currently being tested. For example, neural network-based methods that achieve high accuracies with time windows of one second or shorter have been developed in the last two years [6]–[11]. For overview purposes, we have taken the results of the various existing methods from the publications and summarized them in Table I.

Although these networks have already significantly improved accuracy, a real-world application requires even higher accuracy at shorter windows. In our work, we aim to improve the existing algorithms for the problem of speaker detection. For this, we use long short-term memory (LSTM) based networks. These have already been applied to the problem in different ways [9], [10]. The previously published networks followed an early fusion strategy of combining the EEG and audio data and then processing them together [7]–[9], [12].

However, the audio and EEG signals have different properties. These properties are not considered in the above-mentioned procedures. In this paper, we present the new 3LSTM architecture and its extension. This follows a late fusion strategy, where first the audio and EEG data are processed separately by LSTMs before the results are combined. For the evaluation, the current best networks have been reproduced [7], [8]. We compare the stability of the networks for two different preprocessing pipelines: a widely used one and a simplified variant [6].

II. Methods

A. Channel Attention

The channel attention mechanism weights the input channels by multiplying the entries of each channel by a scalar. For computing these scalars, a neural-network-based attention block is created [7], [11], [13]. This block consists of one convolutional layer (CONV) and two fully connected layers (FC). The input signal is of size $X \in \mathbb{R}^{C \times T}$, where T is the number of time samples and C is the number of channels. First, one CONV layer with one filter is applied. The filter has input sample length $1 \times T$ and is followed by an exponential linear unit (ELU) activation function. Then channel-wise pooling is applied and output E_1 has size $C \times 1$. E_1 is further processed by two FC layers, the first layer has 8 outputs and the second C outputs and thus one scalar per channel. We tested two configurations that differ in the used activation and pooling functions. The formulas for the two

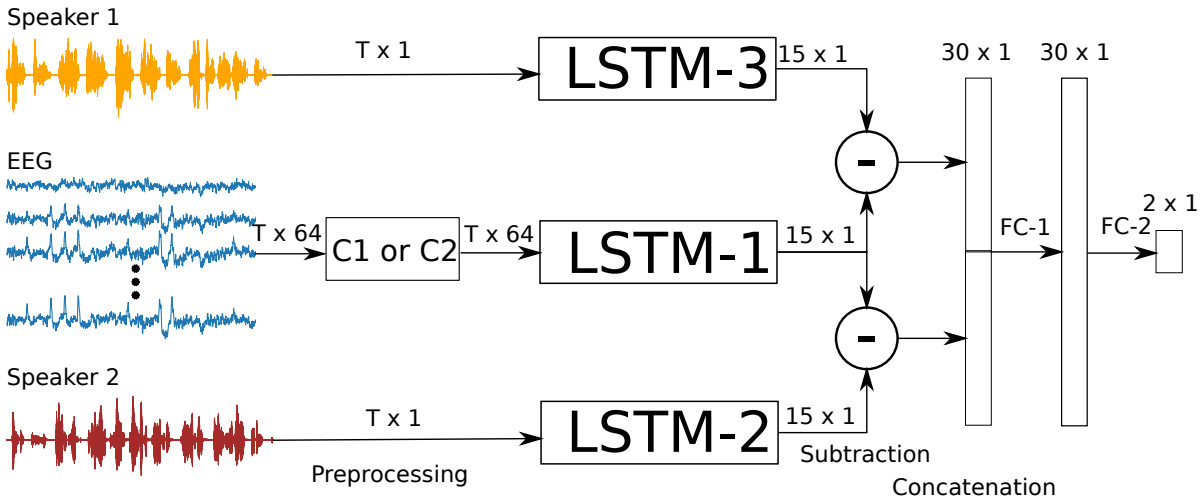

Fig. 1. Structure of the 3LSTM network consisting of LSTMs, FCs and optional attention blocks C1 and C2. T represents the number of samples. The input size after preprocessing is shown to the left of the LSTM symbols, while the output dimension is shown on the right.

different blocks $Y \in \mathbb{R}^{C\times T}$ are given as follows:
C1:

$$\begin{aligned} E_1 &= \mathrm{Max}(\mathrm{ELU}(\mathrm{CONV}_{1\times 5}(X))) \in \mathbb{R}^C \\ Y &= X \times \mathrm{ELU}(\mathrm{FC}_\mathrm{C}(\mathrm{ELU}(\mathrm{FC}_8(E_1)))), \end{aligned} \tag{1}$$

C2:

$$\begin{aligned} E_1 &= \mathrm{Ave}(\mathrm{ELU}(\mathrm{CONV}_{1\times 5}(X)) \in \mathbb{R}^C \\ Y &= X \times \mathrm{Tanh}(\mathrm{FC}_\mathrm{C}(\mathrm{Tanh}(\mathrm{FC}_8(E_1)))). \end{aligned} \tag{2}$$

B. Cosine Similarity-based Channel Attention

Adapted channel attention based on cosine similarity has also been tested [8]. The cosine similarity computes a score for the input vectors $x \in \mathbb{R}^T$ and $y \in \mathbb{R}^T$ as

$$l(x,y) = \frac{\sum_{n=1}^{N} x_n y_n}{\sqrt{\sum_{n=1}^{N} x_n^2}\sqrt{\sum_{n=1}^{N} y_n^2}}. \tag{3}$$

For one multichannel input $X \in \mathbb{R}^{C\times T}$, the scores $E \in \mathbb{R}^C$ are computed separately for each channel $E^i = l(x^i, y)$ with $x^i \in \mathbb{R}^T$. E is used as input for two FC layers. The results are multiplied by the input X.

C. 3LSTM

The designed network is inspired by an LSTM-based network which is also used for auditory attention detection [10]. An overview of our 3LSTM network can be seen in Fig. 1. The first block consists of three independent LSTMs for each of the input signals. These are the 64 EEG channels and the two audio stimuli. T is the number of time samples, which depends on the chosen window length and sampling frequency. LSTM-1 receives the EEG signals as input ($T \times 64$), LSTM-2 the audio signal from Speaker 1 ($T \times 1$) and LSTM-3 the audio signal from Speaker 2 ($T \times 1$). Each LSTM has 15 hidden states and produces a 15×1 output. In the next step, the outputs of the LSTMs are combined. To do this, the output of LSTM-1 is subtracted from the output of LSTM-2 and subtracted separately from the output of LSTM-3. The two resulting outputs are then concatenated. Finally, these results are passed into FC layers. The first, FC-1, has 30 units. The second one, FC-2, provides two output units that are applied to a softmax function. This network is extended to LSTM-C1 and LSTM-C2 by the previously described channel-attentions C1 and C2 to the EEG signal, see (1) and (2).

D. Convolutional Neural Networks

We compared our network to three different networks. The first one is a basic convolutional neural network (CNN) which was used in [6], [8]. It consists of one CONV layer followed by max-pooling and two FC layers. This network is extended to CNN-C1 and CNN-C2 by the previously described channel-attentions C1 and C2, see (1) and (2). This attention is applied to the EEG signal as the first step after preprocessing. The third network is the CNN-CM, which applies the described cosine similarity-based channel attention (CM) [8]. This attention is used two times, once for each speaker. The EEG signal is applied to both attentions separately, producing two 64-channel outputs that are concatenated and then processed by a CONV layer and two FC layers similar to CNN and CNN-C. An overview of the published results of these networks can be found in Table I.

III. Experiments and Data

A. Datasets

For the evaluation, we used two publicly available datasets.

DTU: The dataset was created by the Technical University of Denmark (DTU) [15]. It contains recordings of 18 normal hearing subjects. Two naturally spoken stories were played to the subjects. The stories were divided into 50-second segments. Each subject listened to 60 trials, in which two audio stimuli were presented simultaneously. The subjects had to concentrate on one of the stories. In parallel, the subject's EEG signals were recorded. There was a pause between each

TABLE I

RESULTS OF AUDITORY ATTENTION METHODS FOR INPUT WINDOW LENGTHS OF ONE AND TWO SECONDS TAKEN FROM THE LITERATURE AND THE RESULTS ACHIEVED WITH OUR METHODS FOR THE SAME WINDOW LENGTHS. THE FIRST VALUE IS THE MEAN ACCURACY, AND THE VALUE AFTER THE ± IS THE STANDARD DEVIATION. THESE RESULTS WERE GENERATED ON THE KUL AND THE DTU DATASETS, AS DESCRIBED IN SECTION III-A.

	KUL	
Model	***1s***	***2s***
CCA [14]	60	64
CNN [6]	78.9 ± 11.6	80.4 ± 11.7
CNN-C [7]	81.1 ± 11.9	82.1 ± 12.0
CNN-FC [7]	83.6 ± 10.3	86.9 ± 8.8
CNN-CM [8]	86.5 ± 8.0	88.3 ± 7.9
3LSTM	89.0 ± 5.7	86.9 ± 7.2
3LSTM-C2	**92.8 ± 6.2**	**93.0 ± 6.2**
	DTU	
CCA [14]	60	64
CNN [6]	69.2	71.2
CNN-CF [7]	79.3 ± 8.17	82.9
3LSTM	85.5 ± 9.1	80.1 ± 11.7
3LSTM-C2	**89.6 ± 7.7**	**88.4 ± 8.9**

of the segments. The direction of the stimuli was randomly reversed. Three different virtual auditory environments (VAEs) were simulated in which the two audio sources were positioned at a distance of 2.4 meters in ± 60 degree from the subjects' line of gaze. The audio signals were presented through plug-in earphones at 65 dBA, and the EEG was recorded using a 64-channel EEG system with a sampling rate of 512 Hz [3]. The data without any preprocessing is publicly available.

KUL: The dataset was created by the Department of Neuroscience at KU Leuven [16]. It contains datasets from 16 normal-hearing subjects. As audio stimuli, four audiobooks, divided into two parts of six minutes each, were used. For each subject, two of the stories were randomly selected so that the subject hears only two different speakers at the same time. First, the subject had to listen to two parts of a story, switching the condition and the stimulated ear. Then the same stories were played again, but this time the listener had to listen to the other story. This was repeated with the two remaining stories. So in total, there were eight different presentations with a length of 48 minutes. The audio was presented via plug-in headphones at 60 dBA, with low-pass filtering at 4 kHz. EEG was recorded using a 64-channel EEG system with a sampling rate of 8192 Hz. The available data has been preprocessed. The EEG signals were filtered with a high-pass filter with a cutoff of 0.5 Hz and sampled down to 128 Hz.

B. Preprocessing

We tested two different preprocessing pipelines for the KUL dataset and one for the DTU dataset. In Pipeline 1, the EEG signals were bandpass filtered between 1 and 32 Hz, downsampled to 64 Hz, and standardized. The standardization was performed channel by channel as

$$f(x) = \frac{x - \bar{x}}{\sigma_x + \epsilon}. \tag{4}$$

Here x is one sample, $\bar{x}$ is the mean, and σ_x is the standard deviation of the corresponding channels samples. The envelopes of the audio signals were extracted by applying the built-in Hilbert transform from SciPy and extracting the absolute value [17]. The result was low-pass filtered with a cutoff frequency of 32 Hz, downsampled to 64 Hz, and standardized.

Pipeline 2 followed the preprocessing used in [7], [8]. The EEG channels were referenced to the mean, bandpass filtered between 1 and 50 Hz and sampled at 128 Hz. For the audio stimuli, the auditory-inspired method developed by Briesman et al. was used [12]. The audio signal was decomposed into subbands by applying a gammatone filter bank. The center frequencies were placed uniformly on the Erb scale. For the resulting sub-bands, the envelopes were calculated by taking the absolute value. The resulting envelopes were summed together to create the overall envelope. This signal was given a power of 0.6, bandpass filtered between 1 and 50 Hz and downsampled to 128 Hz.

C. Setup

The described models receive the preprocessed EEG data and the audio stimulus as input and are designed to predict the attended speaker. The models were trained and tested separately for each of the subjects. The data are divided into non-overlapping windows of constant length and split into test and training data. A 5-fold cross-validation was performed so that in each run 80% of the windows were used for training and the rest for testing. The training data was augmented by creating overlapping windows on the training set with an overlap of 50%. This procedure ensured that no data from the test was contained in the training. Window lengths of 0.25, 0.5, 1, and 2 seconds were tested. Networks were implemented in Tensorflow 2.4 using the Adam optimizer with a step size of 0.001. Each model was trained for 100 epochs.

D. Decoding Performance

The results of the described models on the KUL dataset using preprocessing Pipelines 1 and 2 can be found in Table II. Here, the accuracy and standard deviation were averaged for all 5 training/test splits and subjects. Results for the DTU dataset are shown in Table III.

IV. DISCUSSION

Compared to the previous methods tested on the datasets used, the new architecture of the 3LSTM achieved a major improvement. The previously published best value of accuracy for 1-second windows was 86% (CNN-CM) on the KUL and 79% (CNN-CF) on the DTU dataset, see Table I. The base version of the 3LSTM has already increased this to 89% for the KUL and 85% for the DTU dataset. By extending the 3LSTM with the channel attention C2 this could be further increased to 92% and 89% respectively.

As can be seen in Tables III and II, the late fusion strategy chosen in combination with the use of LSTMs in place of CNNs achieved higher accuracies across all experiments. Comparing the results of the 3LSTM with the published results

TABLE II
AVERAGED ACCURACY AND STANDARD DEVIATION OVER ALL SUBJECTS AND RUNS ON THE KUL DATASET FOR THE CORRESPONDING WINDOW LENGTHS IN SECONDS AND PREPROCESSING PIPELINE.

Model	Window length in seconds			
	0.25	***0.5***	***1***	***2***
Pipeline 1				
CNN	78.3 ± 7.2	78.5 ± 7.6	78.1 ± 8.1	73.9 ± 8.3
CNN-C1	75.2 ± 8	77.9 ± 7.9	79.4 ± 7.7	75.7 ± 9.2
CNN-C2	74.0 ± 8.1	78.2 ± 7.9	79.9 ± 7.7	78.8 ± 9.2
CNN-CM	76.0 ± 8.1	79.1 ± 8.0	76.7 ± 10.1	79.5 ± 8.8
3LSTM	**84.6 ± 5.7**	**88.0 ± 5.6**	**87.9 ± 5.5**	85.6 ± 7.1
3LSTM-C1	79.8 ± 8.9	86.2 ± 7.2	85.3 ± 7.5	80.9 ± 10.9
3LSTM-C2	83.1 ± 6.6	88.0 ± 5.9	87.9 ± 6.6	**86.0 ± 8.4**
Pipeline 2				
CNN	83.6 ± 6.9	82.9 ± 6.6	80.8 ± 9.0	78.3 ± 7.4
CNN-C1	87.7 ± 5.5	85.2 ± 9.2	82.2 ± 10.9	75.2 ± 13.4
CNN-C2	88.9 ± 6.4	90.8 ± 6.5	90.6 ± 7.1	89.6 ± 6.4
CNN-CM	83.4 ± 7.1	86.1 ± 6.6	85.1 ± 7.4	81.9 ± 8.8
3LSTM	90.4 ± 3.9	90.5 ± 4.3	89.1 ± 5.3	79.0 ± 11.0
3LSTM-C1	90.9 ± 7.1	91.9 ± 4.3	89.8 ± 5.7	83.1 ± 10.4
3LSTM-C2	**91.8 ± 3.7**	**92.8 ± 4.2**	**92.8 ± 6.2**	**93.0 ± 6.2**

TABLE III
AVERAGED ACCURACY AND STANDARD DEVIATION OVER ALL SUBJECTS AND RUNS ON THE DTU DATASET FOR THE CORRESPONDING WINDOW LENGTHS IN SECONDS AND PREPROCESSING PIPELINE.

Model	Window length in seconds			
	0.25	***0.5***	***1***	***2***
Pipeline 1				
CNN	88.1 ± 7.3	88 ± 7.6	87.5 ± 9.2	85.1 ± 10.6
CNN-C1	88.7 ± 7.8	88.4 ± 8.2	88.9 ± 8.6	88.5 ± 8.7
CNN-C2	88.1 ± 8.2	86.9 ± 9.0	87.9 ± 9.4	88.9 ± 8.8
CNN-CM	87.6 ± 7.9	86.1 ± 8.5	86.9 ± 9.1	87.4 ± 9.1
3LSTM	87.7 ± 7.2	85.9 ± 8.7	85.2 ± 9.9	80.3 ± 11.6
3LSTM-C1	**89.9 ± 6.8**	**89.7 ± 7.2**	**90.0 ± 9.0**	88.4 ± 9.1
3LSTM-C2	88.0 ± 7.3	88.8 ± 7.3	89.6 ± 7.7	**88.4 ± 8.9**
Pipeline 2				
CNN	78.7 ± 9.9	75.6 ± 9.4	73.8 ± 9.8	69.8 ± 9.9
CNN-C1	80.3 ± 11.5	76.9 ± 12.5	72.4 ± 13.3	68.8 ± 11.8
CNN-C2	85.8 ± 9.8	84.4 ± 9.9	81.7 ± 10.3	79.9 ± 11.0
CNN-CM	80.8 ± 10.8	80.0 ± 10.9	79.7 ± 11.1	76.8± 10.7
3LSTM	85.5 ± 8.5	82.5 ± 9.2	77.1 ± 10.9	67.5 ± 12.8
3LSTM-C1	86.2 ± 5.8	**85.6 ± 9.2**	81.1 ± 11.1	74.8 ± 12.9
3LSTM-C2	**86.6 ± 8.3**	85.4 ± 10.5	**83.8 ± 11.4**	**81.9 ± 11.9**

of the CNN-LSTM, which combines the EEG and audio data and processes them with a single LSTM, the advantage of early fusion becomes clear [9]. In the work, only a median accuracy of 75% for the KUL and 55% for the DTU could be achieved for 2-second windows [9].

A weakness of the 3LSTM is that it achieves significantly worse results for the window lengths of 1 and 2 seconds than for 0.5 seconds. Here the LSTM with the given 15 hidden units seems not to be able to keep the context information about the larger number of time steps. Since for the intended application of hearing aids the latencies should be as short as possible, we did not try to optimize the parameters for these window lengths. Moreover, by using the channel attention C2, it is already possible to improve the prediction for two 2 seconds to a level similar to that for 0.5 seconds, see Table II.

Overall 3LSTM-C2 is the most stable one, especially for a window length of 2 seconds. In combination with Pipeline 2 always the highest accuracy was achieved. For Pipeline 1, the model either achieved the highest accuracy or was close to it. Attention C2 achieves better results than Attention C1. This could be due to the pooling function, where C2 uses the average instead of the maximum. The average contains information about the overall distribution of the data, while the maximum is very susceptible to outliers. Also, tanh seems to be the more effective activation function since it limits the range of values to the interval from -1 to 1, which seems to be more suitable for weighting in contrast to the elu function. This is also evident in the comparison of CNN, CNN-C1, and CNN-C2. Again, CNN-C2 is the best of the variants across all setups. In particular, for short window lengths of 0.25s, the 3LSTM has led to significant improvements over previous methods. With an accuracy of 91.8%, the 3LSTM-C2 achieves better results for 0.25s than the previous best method CNN-CM for 0.5s on the KUL dataset [8]. It should be noted that the values achieved in this study cannot be directly compared to those from previous studies due to differences in the partitioning of training and test data. Cai et al. performed 10 random splits of the data, with 60% allocated for training, 20% for validation, and 20% for testing [7], [8]. In contrast, we used a 5-fold cross-validation approach, where 80% of the data was used for training and 20% for testing.

The applied preprocessing pipeline has a strong impact on the performance of CNN-based networks. This effect is especially noticeable on the KUL dataset, where the CNN-C2 achieves an accuracy of 74% for 0.25s with Pipeline 1. With Pipeline 2, however, the same network achieves an accuracy of 88% for 0.25s. This effect can also be observed with the 3LSTM, but the difference is weaker at 6%.

In previous works with CNNs, the difference in performance between subjects was very high, as indicated by the relatively large standard deviation of up to 12. As described by Vandecappelle et al., this effect only occurs when using neural networks instead of correlation-based methods [6]. One possible explanation for the large variability in accuracy between subjects could be that the number of training samples for the neural networks used in classification was insufficient. It can be observed that for LSTMs and Pipeline 2, the variance increases when the window length increases. This issue could be addressed by increasing the overlap of windows or by performing data augmentation.

For the KUL dataset, it can also be seen that the standard deviation for the 3LSTM-based methods is lower than for the CNN-based methods. This suggests that the model generalizes better across different subjects, although the differences between them can still be large. For example, the lowest accuracy achieved by the LSTM-C2 for one subject at 0.25s was 81%, while the highest was 98%. For the DTU dataset, these differences in the standard deviation between the CNNs and 3LSTMs occur only sporadically.

V. CONCLUSION

The goal of this work was to improve auditory attention detection. We developed an LSTM-based network architecture

to detect which of two speakers a person is listening to. In contrast to the methods used in the past, our architecture pursued a late fusion strategy instead of an early fusion one. The network was applied to two publicly available datasets (DTU and KUL) [2], [3]. The networks were compared with current state-of-the-art networks. Two different preprocessing methods were tested [12].

The LSTM-based networks achieved high accuracies, especially for time windows of 0.25 seconds and 0.5 seconds. With an accuracy of 92% for a window length of 0.5 seconds on the KUL dataset, the achieved accuracy is 8% higher than the previously published ones achieved by the CNN-CM on the same dataset [8]. Our work shows not only that the late fusion strategy using multiple LSTMs achieves higher accuracy than the current state-of-the-art, but also that it can be combined with much less complex preprocessing (Pipeline 1).

This is another important step to enable a real-world application. The next steps in development include using audio signals from speakers that were not yet separated. Furthermore, the network architectures are only working with two speakers while the datasets provide only experiments with two competing speakers. The presented networks have to be adapted to allow a different number of speakers. Future work should also test the networks on more realistic datasets recorded with portable EEG devices such as CEEGrids [18]

ACKNOWLEDGMENTS

This work was funded by the Bundesministerium für Wirtschaft und Energie (BMWi) through the KI-SIGS project.

REFERENCES

[1] S. Chadha, K. Kamenov, and A. Cieza, "The world report on hearing, 2021," *Bulletin of the World Health Organization*, vol. 99, no. 4, pp. 242–242A, Apr. 2021.

[2] 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, p. 056014, Sep. 2016.

[3] S. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, Apr. 2017.

[4] E. Alickovic, T. Lunner, and F. Gustafsson, "A system identification approach to determining listening attention from EEG signals," in *2016 24th European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: IEEE, Aug. 2016, pp. 31–35.

[5] A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.

[6] 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," *eLife*, vol. 10, p. e56481, Apr. 2021.

[7] S. Cai, E. Su, L. Xie, and H. Li, "EEG-Based Auditory Attention Detection via Frequency and Channel Neural Attention," *IEEE Transactions on Human-Machine Systems*, pp. 1–11, 2021.

[8] E. Su, S. Cai, P. Li, L. Xie, and H. Li, "Auditory Attention Detection with EEG Channel Attention," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Nov. 2021, pp. 5804–5807.

[9] I. Kuruvila, J. Muncke, E. Fischer, and U. Hoppe, "Extracting the Auditory Attention in a Dual-Speaker Scenario From EEG Using a Joint CNN-LSTM Model," *Frontiers in Physiology*, vol. 12, p. 700655, 2021.

[10] Y. Lu, M. Wang, L. Yao, H. Shen, W. Wu, Q. Zhang, L. Zhang, M. Chen, H. Liu, R. Peng, M. Liu, and S. Chen, "Auditory attention decoding from electroencephalography based on long short-term memory networks," *Biomedical Signal Processing and Control*, vol. 70, p. 102966, Sep. 2021.

[11] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention from EEG," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2022.

[12] 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, pp. 1–1, Jan. 2016.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Computer Vision – ECCV 2018*, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 3–19.

[14] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, Jul. 2021.

[15] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018.

[16] N. Das, T. Francart, and A. Bertrand, "Auditory Attention Detection Dataset KULeuven," Aug. 2020.

[17] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[18] B. Holtze, M. Rosenkranz, M. Jaeger, S. Debener, and B. Mirkovic, "Ear-EEG Measures of Auditory Attention to Continuous Speech," *Frontiers in Neuroscience*, vol. 16, 2022.