EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks

Lucas Deckers[1, 2], Neetha Das[1, 2], Amir Hossein Ansari[1], Alexander Bertrand[1], and Tom Francart[2]

[1]KU Leuven, Dept. Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics. Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[2]KU Leuven, Dept. Neurosciences, ExpORL. Herestraat 49 bus 721, B-3000 Leuven, Belgium

Abstract

When multiple people talk simultaneously, the healthy human auditory system is able to attend to one particular speaker of interest. Recently, it has been demonstrated that it is possible to infer to which speaker someone is attending by relating the neural activity, recorded by electroencephalography (EEG), with the speech signals. This is relevant for an effective noise suppression in hearing devices, in order to detect the target speaker in a multi-speaker scenario. Most auditory attention detection algorithms use a linear EEG decoder to reconstruct the attended stimulus envelope, which is then compared to the original stimuli envelopes to determine the attended speaker. Classifying attention within a short time interval remains the main challenge. We present two different convolutional neural network (CNN)-based approaches to solve this problem. One aims to select the attended speaker from a given set of individual speaker envelopes, and the other extracts the locus of auditory attention (left or right), without knowledge of the speech envelopes. Our results show that it is possible to decode attention within 1-2 seconds, with a median accuracy around 80%, without access to the speech envelopes. This is promising for neuro-steered noise suppression in hearing aids, which requires fast and accurate attention detection. Furthermore, the possibility of detecting the locus of auditory attention without access to the speech envelopes is promising for the scenarios in which per-speaker envelopes are unavailable. It will also enable establishing a fast and objective attention measure in future studies.

Index Terms

Convolutional neural networks (CNN), auditory attention detection (AAD), electroencephalography (EEG), neuro-steered auditory prosthesis, brain-computer interface (BCI)

I. Introduction

In a competing multi-speaker scenario the human auditory system is able to focus on just one speaker, ignoring all other speakers and noise. This situation is called the ‘cocktail party problem’ (Cherry, 1953). Especially the elderly and people suffering from hearing loss have difficulties attending to one person in a noisy environment. In current hearing aids, this problem is mitigated by automatic noise suppression systems. However, when multiple speakers are present, these have to rely on heuristics such as level or direction to determine the target speaker, i.e,. the speaker the user wants to attend to. The emerging field of auditory attention detection (AAD), tackles the challenge of decoding auditory attention from neural activity. This research finds applications in the development of neuro-steered hearing prostheses that can automatically detect the person or direction to whom the user is attending and then amplify that specific speech stream while suppressing other speech streams and surrounding noise, aiming to increase speech intelligibility.

Recent research has shown that the neural activity, recorded using electroencephalography (EEG) or magnetoencephalography (MEG) in a competing two-speaker scenario, consistently tracks the dynamic variation of an incoming speech envelope during auditory processing, where the attended speech envelope is typically more pronounced (Ding and Simon, 2012; O'sullivan et al., 2014). This neural tracking of the stimulus can then be used to determine auditory attention. A common approach is stimulus reconstruction: the post-stimulus brain activity is used to decode and reconstruct the attended stimulus envelope (O'sullivan et al., 2014; Pasley et al., 2012). The reconstructed envelope is then correlated with the original stimulus envelopes, and the one yielding the highest correlation is then considered the attended speaker.

Other methods for attention decoding include the forward modelling approach: predicting EEG from the auditory stimulus (Akram et al., 2016; Alickovic et al., 2016), canonical correlation analysis (CCA)-based methods (de Cheveigné et al., 2018), and Bayesian state-space modeling (Miran et al., 2018). The current state-of-the-art models are capable of classifying auditory attention in a two-speaker scenario with high accuracy (80-90% correct) over a data window with a length of approximately 10 s. However, to quickly detect a switch in attention, detection in much shorter windows, down to a few seconds, is required.

In addition to decoding methods, some practical issues have also been investigated. In all these AAD approaches, access to clean speech streams is necessary. Therefore some integrated demixing and noise suppression algorithms have been developed to grant access to clean speech streams (Aroudi et al., 2018; Das et al., 2017; O'Sullivan et al., 2017; Van Eyndhoven et al., 2017). Researchers have optimized the number and location of concealable miniature EEG electrodes for wearability purposes, minimizing the subsequent loss in performance (Fiedler et al., 2016; Mirkovic et al., 2015; Narayanan Mundanad and Bertrand, 2018).

All studies mentioned above are based on linear decoders. However, since the human auditory system is inherently non-linear (Faure and Korn, 2001), non-linear models could be beneficial for reliable and quick AAD. Convolutional neural networks (CNN) are widely used and very successful in the field of image classification as they have become the preferred approach for almost all recognition and detection tasks (LeCun et al., 2015). Recent research has also shown promising results for CNN-based EEG classification. In seizure detection (Acharya et al., 2018a; Ansari et al., 2018a), depression detection (Liu et al., 2017) and sleep stage classification (Acharya et al., 2018b; Ansari et al., 2018b), CNN have shown promising classification capabilities for EEG data. A CNN for EEG-based speech stimulus reconstruction was presented recently (de Taillez et al., 2017), showing that deep learning is a feasible alternative to linear decoding methods.

Apart from decoding which speech envelope corresponds to the attended speaker, it may also be possible to decode the spatial locus of attention, i.e., not decoding which speaker is attended, but which location in space the person attends to. The benefit of this approach for neuro-steered auditory prostheses is that no access to the clean speech stimuli is needed. This has been investigated based on differences in the EEG entropy features (Lu et al., 2018), but the performance was insufficient for practical use (below 70% for 60-s data frames). However, recent research (Bednar and Lalor, 2018; Patel et al., 2018; Wolbers et al., 2011) has shown that the direction of auditory attention is neurally encoded, indicating that it could be possible to decode the attended sound position or trajectory from EEG. Especially the alpha power activity could be tracked to determine the locus of auditory attention (Frey et al., 2014; Haegens et al., 2011; Wöstmann et al., 2016).

The aim of this paper is twofold. The first goal is to further explore the possibilities of non-linear models for AAD based on CNNs. Our CNN builds on the work of (de Taillez et al., 2017), but rather than reconstructing the envelope we directly classify which of the two speakers was attended to, using the speech envelopes as inputs for the CNN. The second goal is to explore decoding the locus of spatial attention, i.e., the neural network is trained to decode the direction of attention (left or right), in which case it is only provided with the EEG data, not with the acoustic stimuli.

II. Materials and methods

A. Experiment setup

The data set used for this work was gathered previously (Das et al., 2016). Briefly, EEG data were collected from 16 normal-hearing subjects while they listened to two competing talkers, and were instructed to attend to one particular speaker. Every subject signed an informed consent form approved by the KU Leuven ethical committee.

The EEG data were recorded using a 64-channel BioSemi ActiveTwo system, at a sampling rate of 8196 Hz, in an electromagnetically shielded and soundproof room. The auditory stimuli were low-pass filtered with a cut-off frequency of 4 kHz and presented at 60 dBA through a pair of insert earphones (Etymotic ER3A). The stimuli were presented using the APEX 3 program (Francart et al., 2008). The auditory stimuli consisted of natural running speech, and were either presented dichotically (one speaker per ear), or after head-related transfer function (HRTF) filtering, to simulate speech from 90 degrees to the left and 90 degrees to the right of the subject. The order of presentation of both conditions was randomized over the different subjects. The stimuli were set to equal root-mean-square intensities and were perceived as equally loud.

The experiment was split into several trials of approximately 6 minutes duration in which the subject was instructed to attend to one particular speaker. Throughout the experiments, the attended ear of the subject was switched between trials to obtain an equal amount of data for both left and right attended ear per subject, which is important to avoid lateralization bias (Das et al., 2016). In total 8x6 min of unique trials and 12x2 min of repetitions (part of the same stimuli) resulted in an EEG data set of approximately 72 min for every subject.

The data set was split into a training set (70%, approximately 50 min), a validation set (15%, approximately 11 min), and a test set (15%, approximately 11 min). For every trial (6 min) the first part was chosen as the training set, followed by the test set and the validation set, which are thus non-overlapping. We included the 12x2 min of repeated stimuli to increase the amount of data for training. It is noted that this does not create dependencies across training, validation and test set since all repetitions of a specific stimulus interval were included in the same set. Afterwards, all sets were split into overlapping frames of 1, 2, 5, or 10 s long (separate experiments), called detection windows (50% overlap). For 10-s detection windows, the full training set consisted of 616 detection windows per subject, and thus 9856 detection windows overall.

B. Pre-processing

The EEG was filtered with an equiripple bandpass filter, and its group delay was compensated for. For use with linear models, the EEG was filtered between 1 and 9 Hz, which has been found to be an optimal frequency range for linear attention decoding (Ding and Simon, 2012; Pasley et al., 2012). For the CNN models a broader bandwidth between 1 and 32 Hz was used (see below). In both cases, the maximal bandpass attenuation was 0.5 dB while the stopband attenuation was 20 dB (high pass filtering) and 15 dB (low pass filtering). After the bandpass filtering the EEG data were downsampled to 20 Hz (for linear models) and 70 Hz (for CNNs).

The envelope of the speech stimuli was calculated according to the ‘powerlaw subbands’ method proposed by (Biesmans et al., 2017). A gammatone filter bank was used to split the speech into subbands. In each band the envelope was determined and power law compression with exponent 0.6 was carried out. The subband envelopes were then added to generate a broadband envelope, which was filtered with the same filter as used for the EEG recordings and then downsampled to 20 Hz for the linear decoder approach and to 70 Hz for the CNN approach. Eventually, all EEG data were normalized for each trial (approximately 6 min) by subtracting the mean and dividing by the standard deviation across the 6 minute trial. The stimulus envelopes were normalized as well.

C. Linear model: stimulus reconstruction

A linear stimulus reconstruction model (Das et al., 2016; O’sullivan et al., 2014) was used as a reference. In this model a spatio-temporal filter was trained and applied on the EEG data and its time shifted versions up to 250 ms delay, based on least-squares regression, in order to reconstruct the envelope of the attended stimulus. The reconstructed envelope was then correlated (Pearson correlation coefficient) with each of the two speaker envelopes over a data window with a pre-defined length (different lengths were tested), and the classification was made by selecting the speaker that yielded the highest correlation. Leave-one-out cross validation was used to select training and testing data.

D. Convolutional neural networks

CNNs consist of a structured sequence of operations, called layers, that are used to get desired outputs based on supervised learning. Every type of layer has a specific purpose. A first type is a convolutional layer, which consists of several parallel filters, called feature maps, that extract local data features. Feature maps are applied as a convolution, i.e., they slide over the data. After every convolutional layer usually a non-linearity such as a rectified linear unit (ReLu) is applied. The second type of operation is pooling. This layer is used for data dimensionality reduction, in which semantically similar features are combined into one. The last type of layer is the loss layer, in which the objective function and consequent error is calculated. The error, or loss function, for example the mean square error (MSE) between the network outputs and the desired outputs (labels), is minimized during the training process. A special type of loss is a softmax classifier, a combination of a normalized exponential function (softmax) and a logistic loss function that is used to minimize the cross-entropy between the network outputs and the desired outputs.

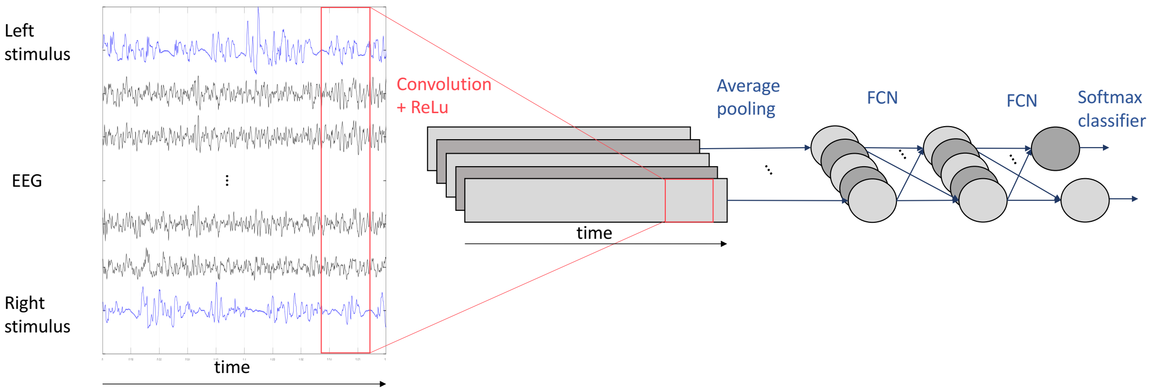


Fig. 1: CNN speaker classification model (CNN:S+D) structure. The stimulus envelopes (blue) and EEG data (black) are combined in a single matrix. The convolution is shown in red. The network outputs are 2 scalars that determine the attended stimulus.

E. Proposed CNN model

We developed two different CNN-based models. In the *CNN speaker classification model (CNN: S+D)*, the CNN is provided with the envelopes of the two speakers and trained such that it can select the envelope corresponding to the attended speaker. In the *CNN direction classification model (CNN:D)*, the network is not provided with the speaker envelopes, but trained to decode the spatial direction of attention. The 'S+D' refers to 'stimulus + direction', as the model is provided with information on both speech stimuli (S) as well as the (left-right) direction (D) of each of them (see further on).

In the CNN speaker classification model (CNN: S+D), both the stimulus envelopes and the EEG data are provided at the input. The left stimulus envelope is added to the data matrix on top of the EEG data, and the right stimulus envelope to the bottom, so the resulting matrix has 66 rows (64 EEG channels + 2 stimulus envelopes). The scalar labels indicate which speaker is the attended one. The location of the stimuli in the matrix with respect to the EEG channels does not influence the CNN training/testing since the first convolution is applied on all channels, including the stimulus envelopes. The input data structure is shown in Figure 1. The stimulus envelopes are shown in blue and the EEG data in black. The network should output a 0 at the top of the two outputs when it believes that the attention is directed towards the left (top signal in the input matrix) or a 1 at the top of the two outputs when it believes that the attention is directed towards the right speaker (bottom signal in the input matrix). The bottom output label is the binary complement if the top output. Since effects of attention are most distinguishable in the M/EEG signal starting 100 post-stimulus (Ding and Simon, 2011), a lag is introduced between the envelopes and EEG data. The stimulus envelope is shifted in time 7 samples with respect to the EEG data, corresponding to a 100 ms lag at the 70 Hz sampling rate.

The CNN:S+D structure is shown in Figure 1. The first step in the model is a convolutional layer, indicated in red. A [66 x 9] spatio-temporal filter is shifted over the input matrix, containing the envelopes and the lagged EEG data. Since the first dimension of the convolution filter is equal to the number of channels at the input, the resulting data has dimension [1 x time]. The 9 samples of the time dimension of the feature maps correspond to a the filter lag spanning 100-230 ms. Five (convolutional) filters are used in parallel, each one generating an output (shown in grey). A rectifying linear unit (ReLu) activation function is used after the convolution step. In the next step, average pooling, i.e., averaging the values of the resulting (parallel) data over the time dimension is done. Following the pooling, there is a two-layer fully connected network (FCN), containing 5 neurons in the first layer and 2 neurons in the second layer, with a sigmoidal activation function. The two FCN outputs are fed to a softmax classifier, consisting of a non-linear softmax function followed by a loss layer with a log likelihood loss function. This layer is used to calculate the error, which is minimized during training, based on the known binary labels that indicate whether attention is directed towards left (top row of the input matrix) or the right (bottom row of the input matrix). When using binary labels, the softmax is essentially reduced to a logistic loss. In future research, when expanding to multiple speakers, using the softmax will be necessary. The full CNN:S+D model contains approximately 3000 parameters.

The CNN:D model is similar to the CNN:S+D model. However, the stimulus envelopes are excluded at the input. The convolutional filter size is adjusted to [64 x time] since the input only contains 64 EEG channels. In this case only the speaker location can be used to determine the attended speaker.

It is noted that the CNN:S+D model can also use information about the speaker location during training since the left speaker is consistently put on top and the right speaker at the bottom of the input matrix. As a result, the CNN:S+D model can decode the direction of attention similar to the CNN:D model. To evaluate the system in a condition in which the direction of attention cannot be used, we added a third training condition, (CNN:S). The model structure is identical to the CNN speaker classification model (CNN:S+D), but now during training the location of the left and right stimulus envelopes with respect to the EEG data in the matrix are alternated. This means that the same EEG data is used twice but with different locations of the attended stimulus envelope in the data matrix. In this way the network cannot learn to use the direction of attention.

All CNN models were implemented in MatConvNet, a CNN toolbox for MATLAB (Vedaldi and Lenc, 2015).

F. CNN training/testing

During training the cross-entropy between the network outputs and the corresponding labels, indicating the desired outputs, was minimized (log likelihood loss). The CNN was trained by updating the weights using back-propagation, based on stochastic gradient descent (Bottou, 2010). First, the CNN was trained using all available training data from all subjects. The network was initialized with random weights between -0.1 and 0.1. The initial learning rate was -0.1, which was decreased stepwise during training to assure convergence. An epoch is defined as one iteration over all training data. The learning rate was halved after respectively 10, 25, and 40 training epochs. The batch size was set at 50. The resulting generic, subject-independent CNN was used as an initialization for subject-dependent CNN retraining. During retraining the initial learning rate and the batch size were divided by 10. Both in generic training and per-subject retraining, L2 weight regularization was used with regularization parameter of 10^{-6}, as

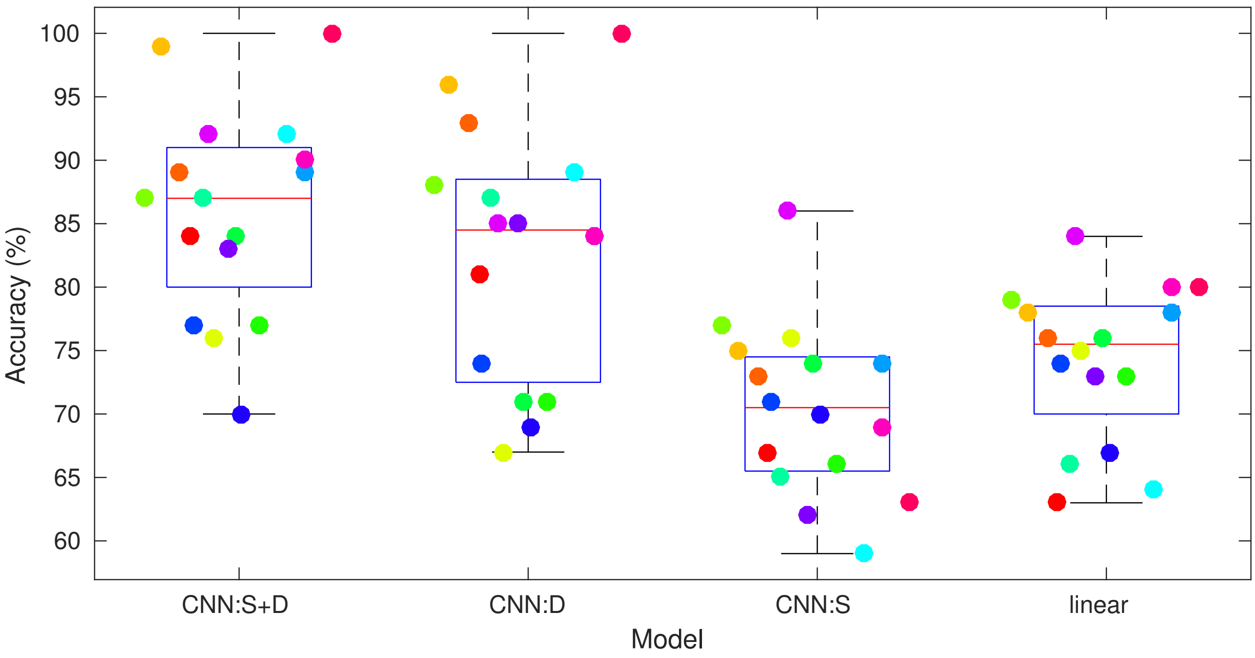


Fig. 2: Auditory attention detection performance for 10 s detection windows for each model: CNN:S+D, which uses stimulus envelope and direction information, CNN:D, which uses only direction information, and CNN:S which uses only stimulus envelope information, in comparison with the benchmark linear decoder. Per-subject results are shown using colored dots.

well as early stopping avoid overfitting. Training was stopped when no loss reduction was found for 10 consecutive training epochs. The training process never lasted more than 60 epochs.

During testing, the logistic loss layer of the softmax classifier was removed since it outputs the classification error. The two network output values were then compared. If the top output is larger, the left stimulus was considered attended and vice versa.

III. Results

A. Decoding performance

Four different detection window sizes were tested: 10, 5, 2 and 1 second windows. This defines the amount of data that is used to make a single decision. In the AAD literature, detection window sizes range from approximately 60 to 6 s. In this work the focus lies on shorter detection windows. This is done in order to avoid a ceiling effect in performance as well for practical reasons, because in neuro-steered hearing aid applications the detection time should ideally be short enough to follow attention switches of the user. The decoding accuracy is defined as the percentage of correctly classified detection windows. Figure 2 shows the subject-specific decoding accuracy for the various CNN models, compared with the linear model (Das et al., 2016), and applied to the same data set, for 10-s detection windows.

For 10 s detection windows, a Wilcoxon signed rank test yielded significant differences in detection accuracy between the the linear decoder model and the CNN:S+D model ($p < 0.001$), with an increase in median from 75.5% to 87%. The CNN:S+D also proved to be significantly better than CNN:D (increase of 2.5%, $p = 0.009$) as well as CNN:S (increase of 16.5%, $p < 0.001$).

For 5 s detection windows, the overall performance was significantly lower (Repeated-measures ANOVA with factors decoder type and window length: $df = 3, F = 47.06, p < 0.001$) than for 10 s windows. The same significant differences between models were found as for the 10 s windows. In addition, for 5 s windows there was also a significant difference between the performance of the CNN:D model and that of the linear decoder ($p < 0.001$). The median decoding accuracy improved from 67.5% for the linear model, to 81.5% for the CNN:S+D model. The median decoding accuracy was also 3% ($p = 0.0469$) higher for the CNN:S+D model in comparison with the CNN:D model and 16.5% ($p < 0.001$) in comparison with the CNN:S model. For shorter detection windows (2 s and 1 s), the CNN:S model weights could not be optimized since no convergence was found.

To quantify the effect of the subject-specific retraining phase, we compared a subject-independent (generic) CNN:S+D model with the results above in which subject-specific retraining was done. These are the results from the generic decoder, which was trained on the full training data set, containing data from all subjects. The average decoding accuracy of the generic CNN:S+D model was 75% for 5 s detection windows and 81% for 10 s detection windows. This indicates that the subject-specific retraining led to a median increase of 6.5% for 5 s detection windows and 6% for 10 s detection windows.

B. *Effect of window length*

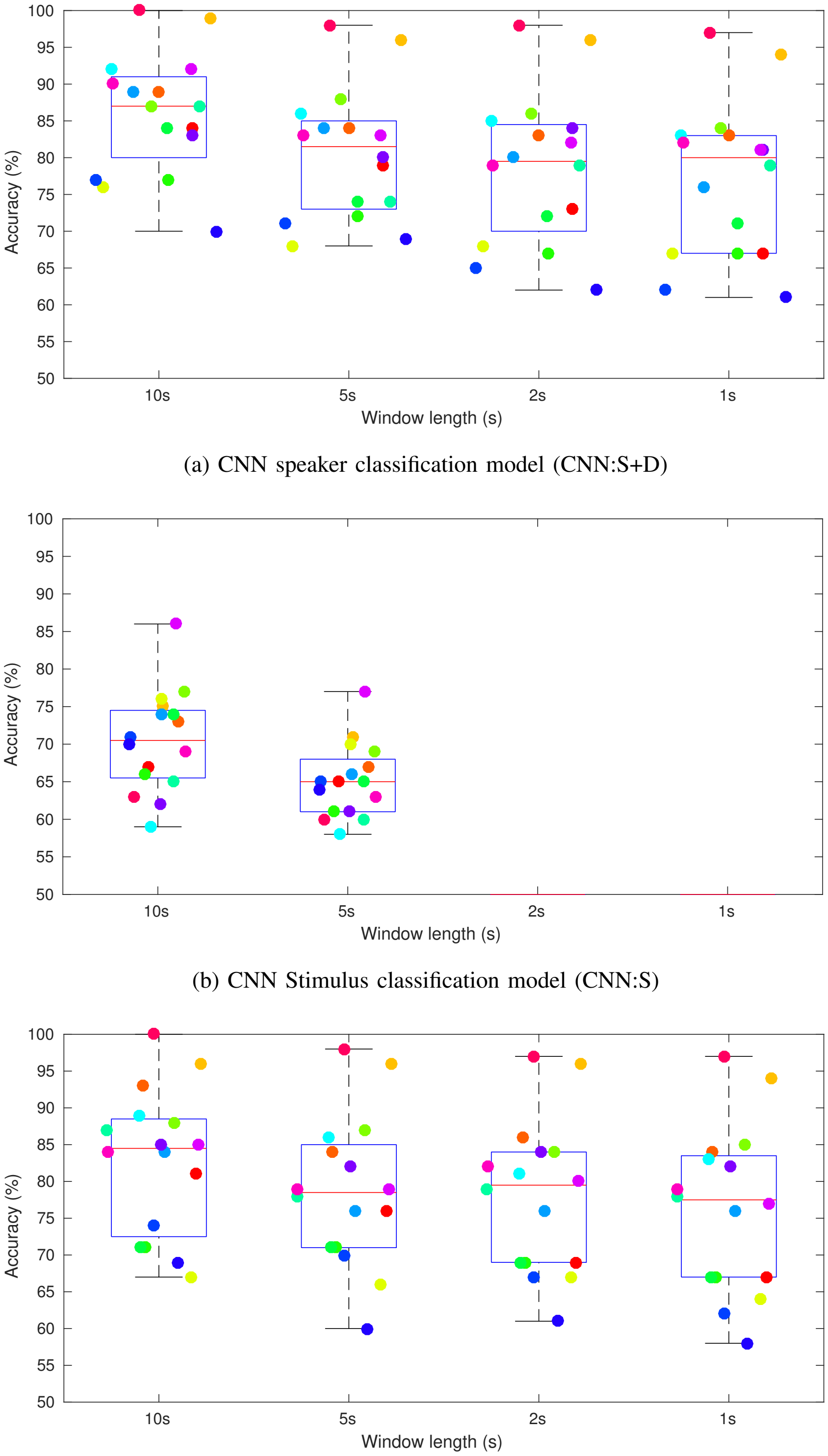
The CNN speaker classification (CNN:S+D) model network structure, shown in Figure 1, was tested for various detection window sizes, ranging from 10 s to 1 s. The results are shown in Figure 3.

For shorter detection windows, the CNN inputs carry less information and are therefore expected to yield a decreased performance. A repeated-measures ANOVA with factors model and window length showed a significant effect of window length ($df = 3, F = 35, p < 0.001$), and pairwise comparisons showed a significant difference between 10 s windows and 5 s windows ($p < 0.001$), 5 s windows and 2 s windows ($p = 0.042$) and between 2 s windows and 1 s windows ($p = 0.012$) However, for 8 out of 16 subjects, decoding accuracy was above 80%, even for 1-s detection windows.

C. *Model design choices*

In this section we describe a number of parameters that were explored to arrive at the model described in Section II.E. Below, for brevity, we only describe differences for 10-s detection windows, but the effect was similar for other window sizes.

As was shown in earlier research (de Taillez et al., 2017), expanding the frequency range to 32 Hz proved to enhance the AAD results. For this reason, the bandpass filter limits were adjusted to 1-32 Hz for the CNN:S+D model to test whether this finding is also valid for our data. This indeed had a significant effect, with median decoding accuracies improving by 10.5% ($p < 0.001$) when the frequency range increased from 1-9 Hz to 1-32 Hz



(a) CNN speaker classification model (CNN:S+D)

(b) CNN Stimulus classification model (CNN:S)

(c) CNN direction classification model (CNN:D)

Fig. 3: CNN models decoding accuracy in function of detection window sizes. The per-subject results are shown by the colored dots.

for the same CNN model. This was not true for the linear model. When the bandwidth was increased from 1-9 Hz to 1-32 Hz the median decoding accuracy dropped by 1%, although this difference was not significant.

In addition to the bandwidth, also the temporal context has a big influence on the CNN model performance. In the linear decoding literature, usually time windows from 0 to 250 ms post stimulus are used, and especially lags between 100 and 230 ms are found to be relevant for auditory attention decoding (Das et al., 2016; de Taillez et al., 2017; Ding and Simon, 2011; Power et al., 2012). We investigated temporal windows of 0-230 ms, 100-230 ms and 150-450 ms post stimulus, and found best performance for 100-230 ms. Adding this information inside the network by adjusting the time lag and convolutional filter size significantly improved the performance in comparison with the same CNN model with an allowed lag from 0 to 230 ms: for a 10 s detection window the median decoding accuracy increased from 82.5% to 87% ($p < 0.001$). Note that with increasing temporal context duration, the number of parameters in the model dramatically increases, which can explain why the shorter temporal integration window (but with the optimal time lags included) yielded better results than the model that included all time lags. The amount of overlap between frames (50% in the optimized model) allows to change the number of input samples during training, but this had little influence on the decoding performance. No significant differences were found when this overlap was increased to 90%.

Batch normalization (Ioffe and Szegedy, 2015) and dropout (Srivastava et al., 2014) are often used as regularization and generalization tools. In this CNN neither of these additions made a significant difference in the results. They were therefore omitted in the final network structure. Weight decay (L2 regularization), the batch size and the learning rate however had a big influence on the network performance, and were carefully optimized for this specific application. While initially 10 parallel filters were used in the convolutional layer, similar performance was found using just 5 (convolutional) filters, halving the number of parameters in the CNN. We therefore retained only 5 filters to decrease the computational complexity and minimize the number of free parameters.

In addition to directly classifying auditory attention in the CNN, we also explored a stimulus reconstruction model using neural networks, trained to reconstruct the attended stimulus rather than classify the attended speaker, similar to the model developed in earlier research (de Taillez et al., 2017). It however did not yield similarly encouraging results as the CNN:S+D model. The median decoding accuracy was only 71% for 5 s and 76% for 10 s detection windows for a subject-specific trained network.

IV. Discussion

We compared the auditory attention decoding results from a novel CNN-based model and a linear decoder and found significantly improved decoding performance when using the CNN-based model. In addition, we were able to decode the direction of attention (left-right), without access to the stimulus envelopes.

A. *Decoding accuracy*

All CNN models that were trained using the locus of attention resulted in a significant increase in decoding accuracy in comparison with the linear model. This shows that a non-linear decoding strategy can be beneficial for AAD. It is hard to directly compare the results of these models to the results presented in earlier CNN-based AAD work (de Taillez et al., 2017), due to the use of different data sets. However, similar to these previous findings, we found that extending the bandwidth to 32 Hz effectively enhanced AAD decoding performance for the CNN, contrary to the linear model, for which no significant differences were found. The CNN probably uses features that are not available to the linear model. For our dataset, a CNN-based stimulus reconstruction model, similar to the one proposed by (de Taillez et al., 2017), resulted in a poorer AAD accuracy than our proposed CNN:S+D and CNN:D models.

For detection windows of 1 or 2 seconds, the CNN:S+D median performance is still around 80%. We are not aware of other models that allow to perform such accurate AAD on such short windows. Compared to (de Cheveigné et al., 2018), for 1 and 2 s detection windows, our CNN:S+D method performs better by respectively by 14% and 10%. (Miran et al., 2018) proposed a real-time AAD algorithm with a high decision frequency, but an effective window length of 10 seconds (the effective window length is $W/(1-\lambda)$, with $W = 16$ samples, $\lambda = 0.975$ and a sampling rate of 64 Hz). (Akram et al., 2016) proposed an AAD system to quickly track changes in attention based on Bayesian modeling and achieved high performance. However, the Bayesian model requires both past and future data, so is not applicable in real-time applications.

The CNN:D model results were only slightly worse (only significant for 10 s detection windows) than the optimized CNN:S+D model, i.e, without access to the speech envelopes, the CNN is able to detect the attended speaker direction with nearly the same accuracy. This is an important result for neuro-steered auditory prostheses, as the *clean* speech signals are not available in realistic circumstances. The CNN:S model, in which the network had access to the clean speech envelopes, but not to the direction of attention, yielded worse performance than the CNN:D model. The CNN:S+D model performed better than the other two models. Taken together, this indicates that the CNN:S+D model effectively exploits both the direction of attention and relevant features from the stimulus envelopes and the EEG to determine the attended speaker.

The CNN:S model performs significantly worse than the linear benchmark for 10 s windows ($p = 0.0210$) and for 5 s windows ($p = 0.0266$). There are two plausible causes for this discrepancy. Firstly, although the fully trained CNN:S model should be able to detect the attended envelope in the input using a filtered representation of the EEG, the CNN cannot implement the same evaluation metrics (correlation) that the linear model uses to differentiate between the speakers. Another possibility for the poor performance may be that the network was poorly initialized and the hyperparameters may not have been optimal, although the same training conditions were applied in the other CNN models.

The proposed CNN:D model yielded much higher performance than the entropy-based classification presented in literature (Lu et al., 2018), in which the average decoding performance proved to be insufficient for real life use (< 80%) for detection windows of 60 s. The CNN:D model with 5 s detection window yielded similar or better decoding performance.

B. *Future improvements*

It has been shown that other speech representations than the envelope, such as a spectrogram or phoneme representation, carry additional valuable information that can be beneficial for AAD (Brodbeck et al., 2018;

Broderick et al., 2018; Di Liberto et al., 2015). In this work, only broadband stimulus envelopes are used. The current CNN structure makes it easy to add additional information channels to the matrix.

Two simple CNN designs are proposed in this work. More complex CNN architectures may benefit more from generalization features such as dropout and batch normalization. In further research deeper network approaches could be explored as well.

For a practical neuro-steered hearing aid, it may be beneficial to make soft decisions. Instead of the translation of the continuous softmax outputs into binary decisions, the system could output a probability of speaker 1 or 2 being attended, and the corresponding noise suppression system could adapt accordingly. In this way the integrated system could benefit from temporal relations or the knowledge of the current state to predict future states. The CNN could for example be extended by a long short term memory (LSTM) network.

C. Applications

The main bottleneck for the implementation of neuro-steered noise suppression in hearing aids thus far has been the detection speed. If we assume that a listener needs no more than 1-2 s to switch attention between speakers, ideally, an auditory attention detection system should be able to make a decision within 1-2 s, with high accuracy. We estimate that with proper heuristics, an minimum accuracy of around 80-90% may be required. While these estimates still need to be validated, it would seem that our CNN-based system was able to overcome this major bottleneck for 7 out of 16 subjects (for minimally 80% accuracy; 2 out of 16 subjects for 90% accuracy). A remaining challenge with current linear and CNN AAD solutions is the inter-subject variability. Especially for short detection windows the results can vary up to 35% between subjects. The goal should be to create an algorithm that is both robust and able to quickly decode attention within the estimated limits for all subjects.

Another difficulty in neuro-steered hearing aids is that the clean speech envelopes are not available. This has so far been addressed using sophisticated noise suppression systems (Aroudi et al., 2018; O'Sullivan et al., 2017; Van Eyndhoven et al., 2017). If the speakers are spatially separated, our CNN direction classification model might elegantly solve this problem by steering a spatial filter towards the direction of attention, without requiring access to the envelopes of the speakers at all. Note that in a practical system, in particular with superdirectional beamformers, the system would need to be extended to more than two possible directions of attention.

For application in hearing aids, a number of other issues need to be investigated, such as the effect of hearing loss (Holmes et al., 2017), acoustic circumstances (Das et al., 2016), background noise and speaker locations (Das et al., 2018), mechanisms for switching attention (Akram et al., 2016) etc. The system would also need to be extended to handle a multi-speaker scenario.

For implementation in hearing aids, the computational complexity would need to be reduced. Especially if deeper, more complex networks are designed, CNN pruning will be necessary. By introducing feature map-wise, kernel-wise, and intra-kernel strided sparsity (Anwar et al., 2017), a CNN can be pruned. Then a hardware DNN implementation, or even computation on an external device such as a smartphone could be considered. Another practical obstacle are the numerous electrodes used for the EEG measurements. Similar to the work of (Fiedler et al., 2016; Mirkovic et al., 2015; Narayanan Mundanad and Bertrand, 2018), it should be investigated how many and which electrodes are minimally needed for adequate performance.

Fast and accurate detection of the locus of attention can be an important tool in future fundamental research. Thus far it was not possible to measure compliance of the subjects with the instruction to direct their attention to one ear. Not only will our CNN approach enable this, but it will also allow to track the locus of attention in almost real-time, which can be useful to study attention in dynamic situations, and its interplay with other elements such as eye gaze, speech intelligibility and cognition.

In conclusion, three novel EEG-based CNN approaches for auditory attention and direction detection have been developed and compared with the current linear AAD model. The results showed a clear decoding performance improvement in comparison with the existing AAD models. The first CNN model exploits both the direction information and relevant features from the stimulus envelopes and EEG data to determine the attended speaker. The second CNN model extracts the direction of the attended speaker only based on the EEG. The third model only uses speech envelopes without information on the direction of the corresponding speakers. Although there are still some practical problems, the proposed models that include the locus of attention approach the desired real-time detection performance. Furthermore, as it does not require the clean speech envelopes, the CNN:D model has potential applications in realistic noise suppression systems for hearing aids.

V. Acknowledgements

The work was funded by KU Leuven Special Research Fund C14/16/057 and C24/18/099, FWO project nrs. 1.5.123.16N and G0A4918N, the ERC (starting grant to Tom Francart, 637424) under the European Union's Horizon 2020 research and innovation programme.

References

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adeli, H. (2018a). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in biology and medicine*, 100:270–278.

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adeli, H., and Subha, D. P. (2018b). Automated EEG-based screening of depression using deep convolutional neural network. *Computer methods and programs in biomedicine*, 161:103–113.

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., and Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, 124:906–917.

Alickovic, E., Lunner, T., and Gustafsson, F. (2016). A system identification approach to determining listening attention from EEG signals. In *24th European Signal Processing Conference (EUSIPCO), Aug 28-Sep 2, 2016. Budapest, Hungary*, pages 31–35. IEEE.

Ansari, A. H., Cherian, P. J., Caicedo, A., Naulaers, G., De Vos, M., and Van Huffel, S. (2018a). Neonatal seizure detection using deep convolutional neural networks. *International journal of Neural Systems*, page 1850011.

Ansari, A. H., De Wel, O., Lavanga, M., Caicedo, A., Dereymaeker, A., Jansen, K., Vervisch, J., De Vos, M., Naulaers, G., and Van Huffel, S. (2018b). Quiet sleep detection in preterm infants using deep convolutional neural networks. *Journal of Neural Engineering*, 15(6):066006.

Anwar, S., Hwang, K., and Sung, W. (2017). Structured pruning of deep convolutional neural networks. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 13(3):32.

Aroudi, A., Marquardt, D., and Daclo, S. (2018). EEG-based auditory attention decoding using steerable binaural superdirective beamformer. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada*, pages 851–855. IEEE.

Bednar, A. and Lalor, E. C. (2018). Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG. *NeuroImage*, 181:683–691.

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). 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*, 25(5):402–412.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Brodbeck, C., Hong, L. E., and Simon, J. Z. (2018). Transformation from auditory to linguistic representations across auditory cortex is rapid and attention dependent for continuous speech. *bioRxiv*, page 326785.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Das, N., Bertrand, A., and Francart, T. (2018). EEG-based auditory attention detection: boundary conditions for background noise and speaker positions. *Journal of Neural Engineering*, 15(6):066017.

Das, N., Biesmans, W., Bertrand, A., and Francart, T. (2016). The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *Journal of Neural Engineering*, 13(5):056014.

Das, N., Van Eyndhoven, S., Francart, T., and Bertrand, A. (2017). EEG-based attention-driven speech enhancement for noisy speech mixtures using N-fold multi-channel Wiener filters. In *Signal Processing Conference (EUSIPCO), 2017 25th European, Kos, Greece*, pages 1660–1664. IEEE.

de Cheveigné, A., Wong, D. D., Di Liberto, G. M., Hjortkjær, J., Slaney, M., and Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172:206–216.

de Taillez, T., Kollmeier, B., and Meyer, B. T. (2017). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*.

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465.

Ding, N. and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859.

Ding, N. and Simon, J. Z. (2012 doi:https://doi.org/10.1152/jn.00297.2011). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1):78–89.

Faure, P. and Korn, H. (2001). 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*, 324(9):773–793.

Fiedler, L., Obleser, J., Lunner, T., and Graversen, C. (2016). Ear-EEG allows extraction of neural responses in challenging listening scenarios—a future technology for hearing aids? In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA*, pages 5697–5700. IEEE.

Francart, T., Van Wieringen, A., and Wouters, J. (2008). APEX 3: a multi-purpose test platform for auditory psychophysical experiments. *Journal of Neuroscience Methods*, 172(2):283–293.

Frey, J. N., Mainy, N., Lachaux, J.-P., Müller, N., Bertrand, O., and Weisz, N. (2014). Selective modulation of auditory cortical alpha activity in an audiovisual spatial attention task. *Journal of Neuroscience*, 34(19):6634–6639, doi:https://doi.org/10.1523/JNEUROSCI.4813–13.2014.

Haegens, S., Nácher, V., Luna, R., Romo, R., and Jensen, O. (2011). α-oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences*, 108(48):19377–19382, doi:https://doi.org/10.1073/pnas.1117190108.

Holmes, E., Kitterick, P. T., and Summerfield, A. Q. (2017). Peripheral hearing loss reduces the ability of children to direct selective attention during multi-talker listening. *Hearing research*, 350:160–172.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

Liu, N., Lu, Z., Xu, B., and Liao, Q. (2017). Learning a convolutional neural network for sleep stage classification. In *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress, Shanghai, China*, pages 1–6. IEEE.

Lu, Y., Wang, M., Zhang, Q., and Han, Y. (2018). Identification of auditory object-specific attention from single-trial electroencephalogram signals via entropy measures and machine learning. *Entropy*, 20(5):386.

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., and Babadi, B. (2018). Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach. *Frontiers in Neuroscience*, 12.

Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering*, 12(4):046007.

Narayanan Mundanad, A. and Bertrand, A. (2018). The effect of miniaturization and galvanic separation of EEG sensor devices in an auditory attention detection task. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, Hawai*, pages 77–80. IEEE.

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, 25(7):1697–1706.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., and Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal* *of Neural Engineering*, 14(5):056001.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251.

Patel, P., Long, L. K., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2018). Joint representation of spatial and phonetic features in the human core auditory cortex. *Cell reports*, 24(8):2051–2062 doi:https://doi.org/10.1016/j.celrep.2018.07.076.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? a late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9):1497–1503.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Van Eyndhoven, S., Francart, T., and Bertrand, A. (2017). EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Transactions Biomedical Engineering*, 64(5):1045–1056.

Vedaldi, A. and Lenc, K. (2015). Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM.

Wolbers, T., Zahorik, P., and Giudice, N. A. (2011). Decoding the direction of auditory motion in blind humans. *Neuroimage*, 56(2):681–687.

Wöstmann, M., Herrmann, B., Maess, B., and Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences*, pages 3873–3878 doi:https://doi.org/10.1073/pnas.1523357113.