International Journal of Psychological Studies; Vol. 11, No. 4; 2019
ISSN 1918-7211 E-ISSN 1918-722X
Published by Canadian Center of Science and Education

Modeling Nonlinear Transfer Functions from Speech Envelopes to Encephalography with Neural Networks

Tobias de Taillez[1], Florian Denk[1], Bojana Mirkovic[2], Birger Kollmeier[1] & Bernd T. Meyer[1]

[1] Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

[2] Department of Psychology, Neuropsychology Lab, and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

Correspondence: Bernd T. Meyer, Medizinische Physik, Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany. E-mail: bernd.meyer@uol.de

Received: May 16, 2019 Accepted: July 2, 2019 Online Published: August 13, 2019

doi:10.5539/ijps.v11n4p1 URL: https://doi.org/10.5539/ijps.v11n4p1

Abstract

Different linear models have been proposed to establish a link between an auditory stimulus and the neurophysiological response obtained through electroencephalography (EEG). We investigate if non-linear mappings can be modeled with deep neural networks trained on continuous speech envelopes and EEG data obtained in an auditory attention two-speaker scenario. An artificial neural network was trained to predict the EEG response related to the attended and unattended speech envelopes. After training, the properties of the DNN-based model are analyzed by measuring the transfer function between input envelopes and predicted EEG signals by using click-like stimuli and frequency sweeps as input patterns. Using sweep responses allows to separate the linear and nonlinear response components also with respect to attention. The responses from the model trained on normal speech resemble event-related potentials despite the fact that the DNN was not trained to reproduce such patterns. These responses are modulated by attention, since we obtain significantly lower amplitudes at latencies of 110 ms, 170 ms and 300 ms after stimulus presentation for unattended processing in contrast to the attended. The comparison of linear and nonlinear components indicates that the largest contribution arises from linear processing (75%), while the remaining 25% are attributed to nonlinear processes in the model. Further, a spectral analysis showed a stronger 5 Hz component in modeled EEG for attended in contrast to unattended predictions. The results indicate that the artificial neural network produces responses consistent with recent findings and presents a new approach for quantifying the model properties.

Keywords: auditory attention, EEG, machine learning, neural networks, nonlinear models

1. Introduction

Auditory attention is an intuitive human skill and essential for participating in everyday conversations. In difficult acoustic situations, normal-hearing listeners excel in directing their attention to a source they want to listen to and can ignore competing sound sources with moderate effort [Shinn-Cuningham and Best 2008; Bronkhorst 2000]. In this context, establishing a transfer function between neurophysiological measurements and physical stimuli can help to better understand these processes.

To quantify the relation of acoustic stimuli and the resulting physiological responses, event-related potentials (ERP) captured with electroencephalography (EEG) and evoked with auditory clicks have become a standard tool in clinical examinations and research [Davis 1939; Picton, Woods, Baribeau-Braun and Haeley 1977] and were even investigated for aspects of auditory attention [Getzmann, Jasny, and Falkenstein 2017]. However, stimuli evoking ERP responses do not represent the vast majority of acoustic events that are encountered in everyday soundscapes - including spoken language - and it is therefore unclear if results obtained with simple clicks are transferable to complex acoustic scenes.

This issue was addressed by Lalor et al. [Lalor, Power, Reilly and Foxe 2009], who introduced an encoding (or forward) model to measure auditory event related potentials from continuous but artificial stimuli in an auditory spread spectrum analysis in normal-hearing listeners. In the follow-up study [Lalor and Foxe 2010] this approach was applied to continuous speech stimuli with the result of estimating a transfer function from speech envelope

to EEG with high temporal precision. Ding and Simon [2012] advanced this technique by using a magnetoencephalographic paradigm and found a frequency dependent transfer function that models neurophysiological responses to attended or unattended stimuli. In accordance with the above studies, our study does not attempt to reconstruct the best possible EEG prediction from the speech envelopes. This cannot be achieved at all due to the poor signal-to-noise ratio and the low information density of the envelopes. Instead, the long-term goal is a purely scientific one by trying to gain insights into the functioning of the auditory system and speech processing despite the poor SNR. As a first step towards this goal, a new method is proposed and grounded against recent study results. One limitation of these studies is that they are solely based on linear methods to model responses to speech sounds. Many psychoacoustic percepts do instead establish a nonlinear relation between the phyical stimulus and the observed psychophysical quantity such as loudness [Derleth, Dau and Kollmeier 2001] or frequency perception [Poirier and Wilson 2006], and therefore, a nonlinearity between speech envelope and physiological measurements of the auditory pathway can be assumed at least to some degree.

Mapping such black box mechanisms became recently approachable by the emergence of deep machine learning. These powerful methods have substantially improved the state-of-the-art in many pattern recognition tasks, which is often attributed to the capability of learning complex non-linear input-output relations by topologies like deep neural networks (DNN). Although the primary focus of DNN is the engineering, task-oriented application, there is a trend to analyze the how and why DNN deliver such high performance [Bach 2015; Sturm, Bach, Samek & Müller 2016; Zeiler & Fergus 2013]. Motivated by this progress, the central approach of the current study is to apply deep learning to model the encoding process that maps speech to brain activity. The combination of DNN-based analysis in neurophysiological research is still relatively scarce at the moment and the presented study takes some early steps to fill this gap.

To this end, a DNN is trained to model a transfer function from speech envelopes to electrocorticographic measurements by exposing it to the envelopes of the speech used in listening experiments, i.e., one attended and one unattended speech stream produced by two speakers in a spatial scene. As empirical data we employed benchmark data provided by Mirkovic et al. [Mirkovic, Bleichner and De Vos 2016] who used it by modeling the reverse transfer function of EEG to speech envelope. During experiments, subjects listened to two speakers embedded in a spatial acoustic scene over headphones while EEG was acquired with a high-density cap (84 electrodes). The envelopes of both speech streams were extracted from the respective time signals and used as input to the network.

In the current work the network is analyzed regarding effects of auditory attention and the modeled nonlinearity. The first is considered to validate the findings with recent studies and the latter to gain first insights into the additional nonlinear model capabilities of neural networks in neurophysiological context. First, a spectral analysis of the modeled EEG response for a continuous input stimulus is conducted to identify characteristic neurophysiological frequencies related to speech-envelope processing with respect to auditory attention effects. Second, the model properties in time domain are investigated by using a traditional short stimulus (in our case, a delta pulse) as input to the nonlinear model. This technique is used for auditory EEG since the first recorded EEG in the 1940s [Davis 1939; Picton, Woods, Baribeau-Braun and Haeley 1977]. An important model property in this context is the importance and contribution of nonlinear and linear processing learned by the DNN. We here propose to quantify these contributions by using exponential frequency sweeps as input to the DNN and analyzing the corresponding output. The technique was adopted from state-of-the-art methods to characterize the nonlinear behavior [Novak, Simon, Kadlec and Lotton 2009] of electro-acoustic systems like loudspeakers or amplifiers [Farina 2000]. The method facilitates estimation of the ratio of nonlinear to linear output components including the (non)-linear parts of the speech envelope to EEG transfer function.

This analysis is motivated by a method proposed earlier for separating the linear and nonlinear response characteristics of reverberant rooms [Novak, Simon, Kadlec and Lotton 2009]. With this method, we estimate the ratio of non-linear and linear processing by the DNN.

This paper is organized as follows. Section 2 introduces the experimental design including the listening experiments, the structure and training procedure of the proposed DNN model, as well as the approach for spectral and temporal analysis of model properties. Section 3 presents the results including model responses to click-like stimuli and exponential sweeps. Section 4 discusses the findings in the context of recent literature and presents a brief outlook on future developments before the paper is summarized in Section 5.

2. Materials and Methods

2.1 Auditory Attention Experiments

The data set described in the following was introduced in Mirkovic et al. [Mirkovic, Debener, Jaeger and De Vos 2015] for the task of auditory attention decoding with linear methods. Twenty young self-reported normal-hearing participants (German native speakers, mean age: 24.8, 8 male, 1 left-handed) listened over in-ear headphones (E-A-RTONE 3A) to two simultaneously presented stories produced by two speakers (both male) for 50 minutes in 5 consecutive 10-minute blocks. The speakers were virtually placed at 45 degree azimuth in a distance of one meter by convolving the time signals with head-related transfer functions measured in an anechoic room from a previous published corpus by Kayser et al. [2009]. The German stories read by the two speakers were 'Zwerg Nase' by Wilhelm Hauff presented to the left ear and a concatenation of 'Der Jäger und der Zwergenprinz' by Ulrich Jahn and 'Die Prinzessin im Felsenriff' by an unknown author. The locations of the speakers were kept constant throughout the experiment. Ten randomly selected listeners were instructed to attend the left speaker and ignore the right speaker; the remaining ten listeners attended the right speaker. High-density EEG was recorded with 84 channels positioned on the 96-channel EEG cap (Easycap, Germany). Six electrodes around each ear were not used since behind-the-ear EEG was measured simultaneously, which is however not used for the present study (for details refer to [Mirkovic, Bleichner and De Vos 2016]). Data was obtained using 16-bit BrainAmp amplifiers (Brain Products, Gilching, Germany) with an analog band-pass filter from 0.0153 to 250 Hz at a sampling rate of 500 Hz. Participants were instructed to visually focus on a crosshair on a screen to reduce eye movement.

Every 10 minutes, the experiment was paused, and the participants answered a content-related questionnaire to ensure auditory attention. One participant was excluded from further analysis due to incorrect responses; a second participant was excluded due to technical problems during recording. Both excluded participants attended the left speaker. To balance the data set, two randomly selected listeners who attended the right speaker were also excluded from analysis. All 20 participants signed an informed consent form and were equally paid. The study was approved by a local ethics committee and this data has first been published in Mirkovic et al. [Mirkovic, Bleichner and De Vos 2016] with another focus of the evaluation.

2.2 Preprocessing

To remove measurement artifacts from the EEG data, we performed an automatic channel rejection with the EEGLab toolbox [Delorme and Makeig 2004] based on the channel kurtosis and a subsequent interpolation of excluded channels taking into account the electrode position on the skull. The best practice procedure from the EEGLAB Toolbox was used for these purposes. This processing reduces the information contained in individual data sets but simultaneously enhances comparability between participants and therefore eliminates systematic subject-specific noise in the data set.

EEG sets were re-referenced to a common average reference and low-pass filtered at 125 Hz. Following a high-pass filter at 1 Hz to eliminate electrode drift, EEG signals were down-sampled to 250 Hz. Speech envelopes were calculated as the magnitude of the analytic time signal from the raw audio signals, low-pass filtered at 125 Hz and down-sampled to 250 Hz.

The data set of each participant was subdivided into blocks with a duration of one minute. Each block of EEG data was z-scored resulting in zero mean and unit variance over time for each channel, respectively. The chosen network architecture (see next section) compresses the final output with a standard function (hyperbolic tangent) which limits the output, i.e., the predicted EEG response, to a value range between -1 and +1. To ensure most EEG data points (99.9%) fall into this range, the EEG data further was normalized from unit variance to a standard deviation of 0.3. Speech envelopes were z-scored with respect to the complete duration of 50 minutes.

2.3 Network Layout

A deep feed-forward network as illustrated in Figure 1 was trained to predict EEG data from concurrent envelopes of attended and unattended speech. The training procedure used the combined data sets of all 16 participants to obtain a model that generalizes across these subjects, later tested by processing data from arbitrary listeners with the multi-listener model. Our network uses two envelopes values as input (one from each speech stream) to predict the EEG data related to auditory attention for the subsequent time segment with a duration of 600 ms. Latencies after that point are not considered because they presumably contain more processing related to semantic context; something not derivable from envelope values with a time-invariant DNN.

The network designed for these experiments consists of seven layers, where each layer is fully connected to the subsequent layer and each neuron has an adjustable threshold value. In this standard approach the activation of a

given neuron x^ with index i in layer j is calculated as

$$\hat{x}(i/j) = act\left(\sum_{n=1}^{N} x(n/j-1)\cdot w_{i,n} + b\right) \quad (1)$$

where N denotes the numbers of neurons in the layer j-1, b represents the neuron's learned bias, x(n | j-1) is the neuron with index n in the j-1 layer, $w_{i;n}$ is the adjustable weight between both neurons and act represents the activation function hyperbolic tangent.

The input layer with two neurons (one for each speech envelope stream value) is followed by layers with increasing number of neurons in each layer of 4, 8, 12, 16 and 20 neurons, all of which use a standard tanh activation function. Inspired by the layout of speech related neural networks [Martinez, Mallidi and Meyer 2017] this network size was chosen as a compromise between numbers of layers of the network and a reasonable number of adjustable parameters given the amount of training data specified above. Tests with slightly varying network depth and width did not show relevant differences in the outcomes as long as the parameter count was kept in the same range and therefore, we settled for the layout with a pyramidical scheme to make the dimension increase preferably smooth. The final layer maps the output of the 20 hidden neurons to a matrix of 84 (channels) times 150 (time samples) resulting in 12,600 output neurons (cf. Figure 1). Since the predicted envelope can assume arbitrary values, this corresponds to a regression problem for which the final activation function is also tanh. This is in contrast to classification problems, for which a softmax function is typically chosen that facilitates activations according to a 'winner-takes-all' rule. The total number of weights in the network is 265; 308. EEG predictions are obtained by processing subsequent input (envelope) samples and since we obtain a prediction for 150 EEG time samples (= 600 ms) for each single input envelope value (as illustrated in Figure 1, the resulting output matrices are highly overlapping (for up to 99.33 %); the final EEG prediction is obtained by averaging these overlapping output matrices. During training, weight updates are performed by measuring the error between the measured and the predicted EEG data for the averaged output for one-minute blocks of data (as described in the next section). The Nadam optimizer [Dozat 2016] was used with a logcosh cost function for adapting the DNN weights. This cost function approximates a mean squared error criterion for small deviations between prediction and target while resembling a linear loss function for large deviations. Therefore, typical high-power EEG artifacts arising from eye blinks or sudden muscle movements have a reduced influence on the DNN training process.

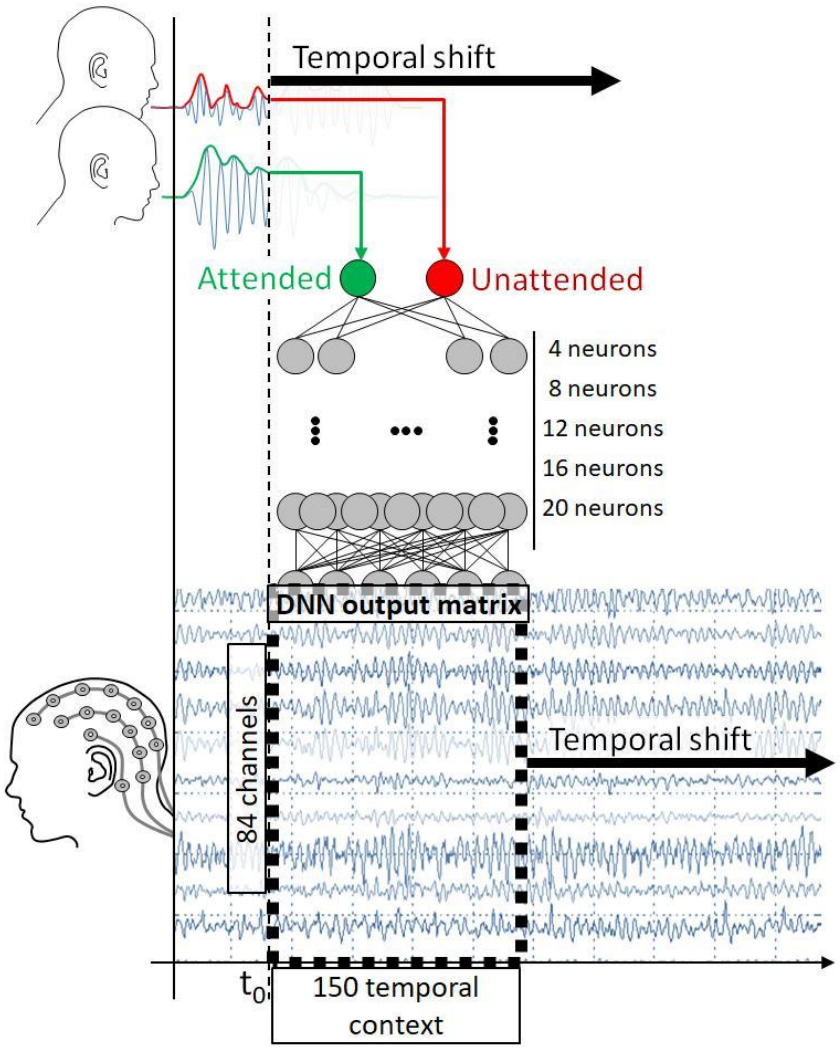

Figure 1. Illustration of the proposed DNN-based encoding model. Envelope samples of both speakers (attended and unattended) are fed into the network that predicts EEG responses arranged in a 2D matrix with the size of 84 (channels) 150 (temporal context corresponding to 600 ms). Processing of subsequent envelope samples produces overlapping EEG predictions that are averaged to obtain the final model output.

Training and evaluation were performed with a leave-one-out cross validation scheme with 25 evaluation cycles. From 50 minutes of data (EEG and speech envelopes) per participant, consecutive two-minute segments with even start indices were used for evaluation (total of 2 16 = 32 minutes) while the remaining data was used for training (48 16 = 768 minutes). The first half of the two-minute segment was used as a development set (or dev set in short) and the second half served as test set for further evaluation, resulting in two 16-minute sets. The training process itself was stopped as soon as the loss on the dev set did not decrease for 10 epochs. Since the training procedure uses randomized weight initialization, the whole leave one out cross validation was repeated 10 times. Since multiple repetitions of DNN training in the same cross-validation cycle cannot access new information but cover the influence of random DNN weight initialization, the resulting simulated EEG responses from each training repetition (including re-initialization) were averaged before further analysis.

To test the validity of our approach, a second identical model was trained by pairing the EEG data with input envelopes from a random time segment which should not be informative for attention decoding (but could reflect long-term properties of the driving stimulus including the spectral shape of the envelope). This model is named ’control model’ in the following.

2.4 Analysis of the Deep Encoding Model

The trained DNN is assumed to model the auditory speech processing to some extent and therefore our focus is to extract the learned properties of the DNN. This is approached by analyzing the spectral properties of the

modeled EEG also by exciting the network with clicks and sweeps to measure its time domain response to input stimuli. While click excitation closely resembles classical EEG experiments with auditory evoked potentials, using sweeps in this context has not been explored before. A DNN as a nonlinear system is not characterized completely by its impulse response, whereas using a signal with an exponential increasing frequency component (a sweep) enables a separation of the linear and nonlinear components of the response [Novak, Simon, Kadlec and Lotton 2009].

2.4.1 Spectral Properties of Model Responses

The spectral properties of the modeled EEG are analyzed by processing model response to the speech envelopes from the evaluation set. Data for the attended and unattended envelopes are processed separately, which allows contrasting the resulting spectra. The long-term spectra are obtained by averaging the absolute amplitude spectra from a moving Fast Fourier transform (FFT; window length 200 samples or 0.8 s) for each EEG channel separately. The exact FFT length is not a sensitive parameter for the analysis, however, too short windows (below 0.5 seconds) should be avoided to provide an adequate frequency resolution. The resulting spectra are averaged over EEG channels while the statistical significance of differences between attended and unattended long-term spectra is tested with a two-sided student t-test for each spectral bin separately over cross-validation cycles with an additional correction for multiple testing (200 tested spectral bins).

2.4.2 Click Response Analysis

To obtain the model response to clicks, a two-channel signal was used as input to the neurons for the attended and unattended stream, respectively, where all signal values were set to -1 except for one value set to 1 for the input channel of interest. The next 150 time samples (or 600 ms) produced after the 'click' were used as measurement. Differences between responses to attended and unattended excitation were tested with a two-sided t-test based on the underlying distributions of cross-validation cycles at each time point for $p < 10^{-6}$ with an additional correction for multiple testing (150 tested time points).

2.4.3 Nonlinear and Linear Processing of the DNN-Based Model

Since the DNN is an inherently nonlinear system, its output y(t) is a nonlinear transformation of the input signal s(t). In linear systems, the output is given by the convolution of the input signal with the system's impulse response h(t). Therefore, a linear system is fully characterized by its impulse response. For nonlinear systems, the mathematical model has to be extended. One option is the generalized polynomial Hammerstein model [Novak, Simon, Kadlec and Lotton 2009] that consists of an ensemble of impulse responses that are convolved with ascending powers of the input signal and summed

$$y(t) = h_1(t) * s(t) + h_2(t) * s^2(t) + h_3(t) * s^3(t) + \dots \ , \qquad (2)$$

where the operator indicates a convolution. With this model, the DNN can be approximated by an ensemble of linear $h_1(t)$ and higher-order impulse responses $h_{1,2,\dots}(t)$, where the latter characterize the non-linear part of the system. We characterized these impulse responses of the current DNNs using the exponential sweep method [Novak, Simon, Kadlec and Lotton 2009] that is well-known from nonlinear system identification and outlined in Figure 2. A sine sweep with fixed amplitude (set to standard deviation of the original input signal with zero mean) and exponentially increasing frequency was fed into the input of the DNN. The output was then convolved with an inverted version of the sweep signal. The result is a temporal series of impulse responses, corresponding to the linear impulse response starting at t = 0 and nonlinear responses at negative times, where the temporal shifts depends on the order of nonlinearity and sweep parameters. The linear and harmonic impulse responses of each order were then separated by temporal windowing and analyzed independently. These impulse responses characterize the frequency-dependent nonlinearities of the DNN created for each distortion order. However, further processing is necessary to characterize the processes in the DNN that create these harmonic distortions (i.e., $h_{2,3,\dots}(t)$). Each harmonic distortion mechanism of even or odd order also creates distortion components at lower orders. For example, a portion of the DNN nonlinearity depending on the 7th power of the input signal also creates distortion components occurring in the measured impulse response of the 5th, 3rd and 1st (linear) orders. The linear transformation compensating for this effect as described in [Novak, Simon, Kadlec and Lotton 2009] that is based on addition theorems for forms of $\sin^N(x)$ was utilized to identify the underlying harmonic impulse responses of the DNN.

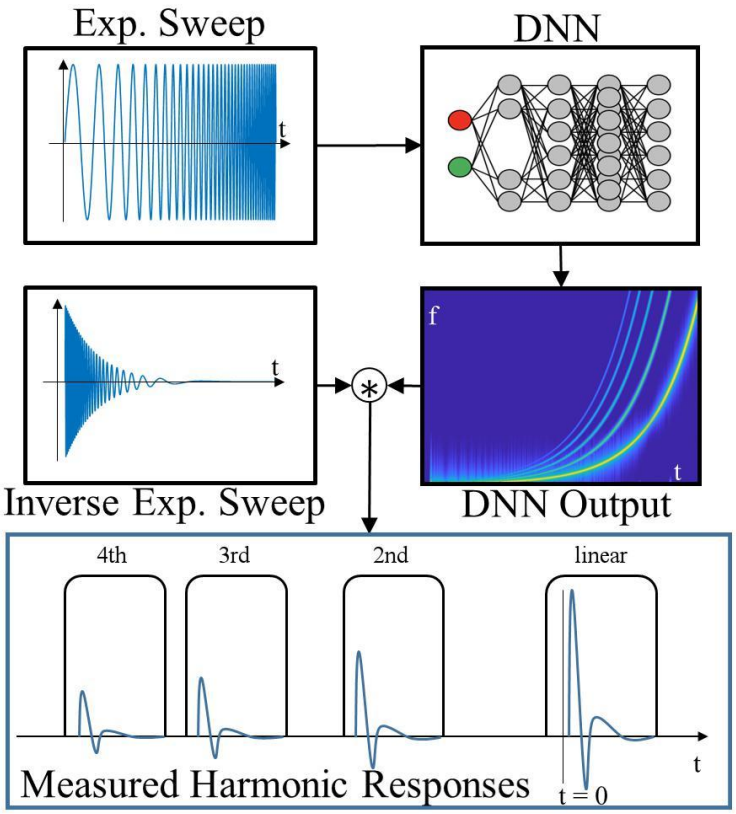

Figure 2. Flowchart of DNN nonlinearity analysis. An exponential sweep is fed into the DNN. The DNN output (shown here as a spectrogram plot) is convolved with an inverse of the sweep signal to obtain the linear and harmonic responses of the DNN, which are temporally separated depending on the distortion order. Temporal windowing (i.e., using only the temporal range indicated by the black curved windows) and a further transformation yields the individual linear and harmonic impulse responses of the DNN.

The exponential sweep signal covered a frequency range between 1 and 125 Hz (up to half of the sampling rate). Signal amplitude values were set between -1 and 1 in consistence with the speech envelope, although variable input amplitudes have diminishing effects on the outcome because the approach proposed by Novak et al. [Novak, Simon, Kadlec and Lotton 2009] used here yields the internal distortion models and not the generated distortion levels. The analysis was conducted for each trained DNN and subsequently averaged over cross-validation cycles.

3. Results

3.1 Spectral Analysis

Results of the spectral analysis of predicted EEG data are depicted in Figure 3, in which the shaded area corresponds to the standard deviation over cross-validation cycles. Large significant differences between attended and unattended spectra of the modelled EEG (channels averaged) are observed in the range from 4 to 7 Hz. An additional evaluation of the speech envelopes (data not shown) showed no peak at 5 Hz, so these spectral results probably do not just reflect the speech energy correlated processing. The spectral bin with the largest difference at 5 Hz is spatially analyzed by plotting the 5 Hz component of the power spectrum for each electrode, which is visualized as a topological plot in Figure 4.

The differences between the condition ‘attend’ and ‘unattended’ (shown in the right panel in Figure 4) are most pronounced bilaterally in temporal regions, where auditory cortex is located.

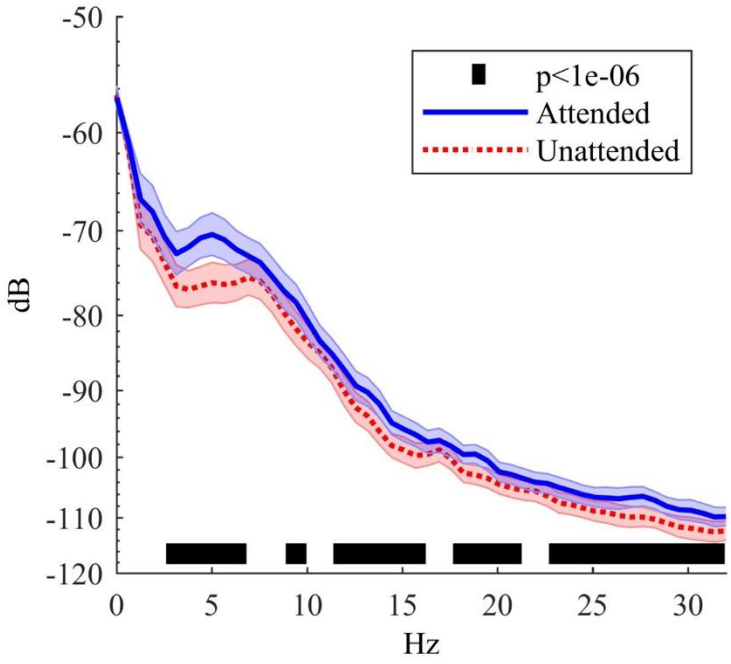

Figure 3. Long-term spectrum of the predicted EEG response obtained from 100 minutes of speech envelope inputs for both channels (attended and unattended) independently. The blue/solid and red/dotted curve correspond to the attended and unattended case, respectively. Black areas denote significant differences between attended and unattended; corrected for multiple testing.

3.2 Click Response of the Model

As described in Section 2.4.2, the model response to click-like stimuli is obtained by using two-channel signals for which a single sample in the attended or unattended channel is set to one. The resulting model responses were averaged over cross-validation cycles and plotted over EEG channels in Figure 5. In each subplot, the model responses for the conditions attended and unattended are plotted in blue and red, respectively. The largest response amplitudes are observed at the central electrodes and at the mastoid regions. For the Cz electrode, a more detailed view is presented in Figure 6, which additionally shows the responses from the control model in green and black (cf. last paragraph in Section 2.3). The model response for the attended conditions is significantly different from the unattended condition. Significant differences occurred from 130 to 220 ms and 250 to 370 ms as well as for several single points for higher latencies above 400 ms. Additionally, the two conditions were significantly different for the single time frame from 0 to 25 ms. For the control-model that was trained with unrelated input envelopes, the procedure was repeated and no significant differences were found between the envelope channels as expected.

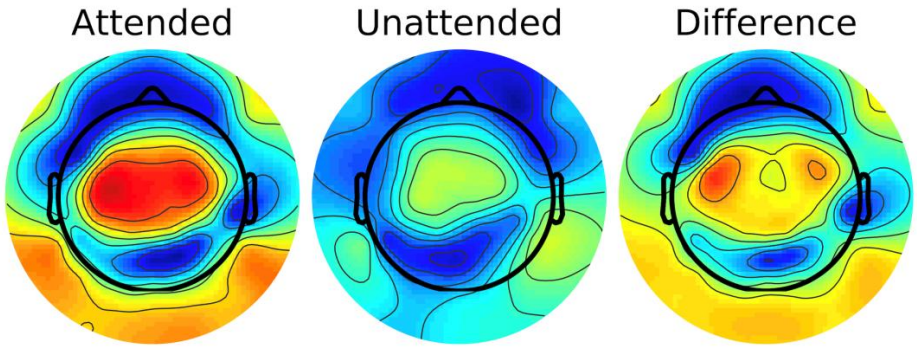

Figure 4. Topological plot of the spectral 5 Hz component for the attended (first) and unattended (middle) processing and the difference between these (right). The difference plot is z-scored individually while the first two plots share a color scale. Red indicates strong and blue low spectral activity in all three plots.

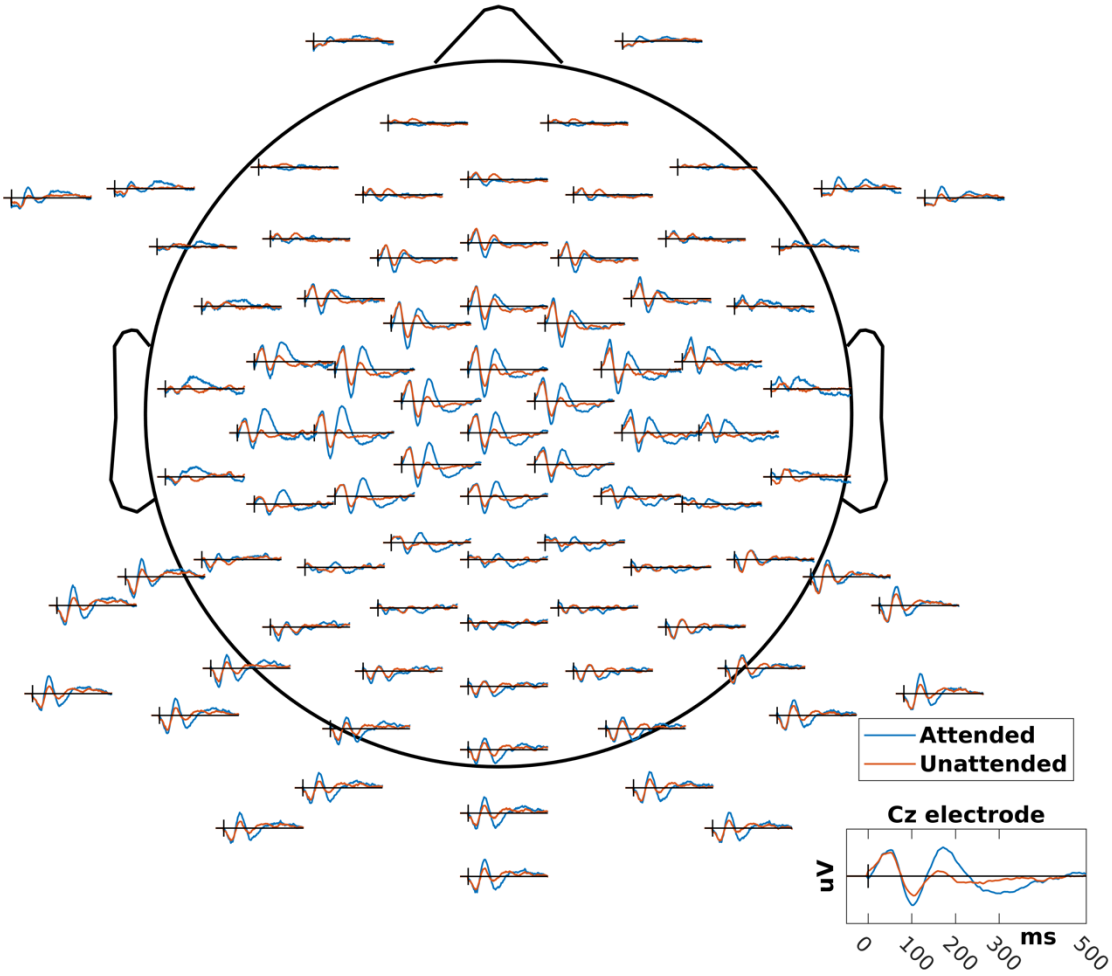

Figure 5. Topographic plot of EEG click response of the model for each electrode position. Data for the Cz electrode is shown separately in the bottom right.

3.3 DNN Nonlinearity

Results of the analysis of the contribution of linear and nonlinear processing are shown in Figure 7. The impulse responses for the linear and higher order non-linear cases were obtained by the scheme illustrated in Figure 2. Subplots a) - e) show individual time-domain impulse responses for the attended and unattended cases, respectively.

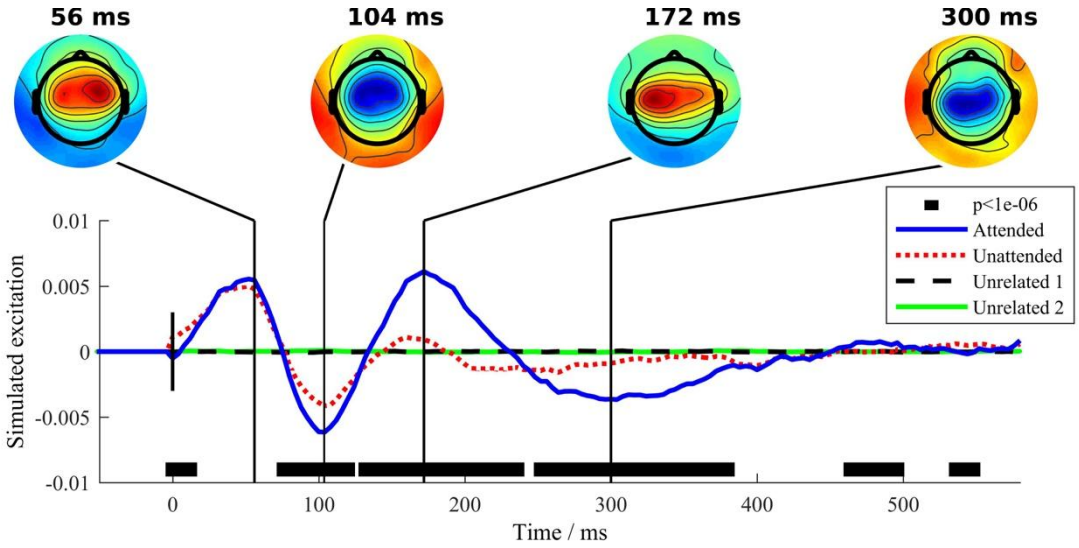

Figure 6. Click response of the DNN-based model obtained from the DNN output neuron for the Cz electrode. The blue/solid and red/dotted graphs show responses for input corresponding to the attended and unattended case, respectively. Green/solid and black/dashed EEG responses are evoked from the model trained on temporally misaligned speech envelopes. Black markers at the bottom indicate significant differences between the attended and unattended response.

Subplot f) shows the spectrogram calculated from the DNN impulse response after deconvolution for an attended case. The spectrum of each component (from a) to e)) is plotted in g) to contrast spectral differences.

The linear impulse response has a similar structure as the click response (as shown in Figure 6) but exhibits a lower level. The nonlinear impulse responses can be divided into two different groups: While for the even harmonics (2nd and 4th) no influence of attention is evident, the odd components (3rd and 5th) are quite strongly modulated at a latency of 200 ms but only the 3rd harmonic shows additional attentional effects at 100 ms and 300 ms.

The contribution of each harmonic component is quantified by measuring the root mean square (RMS) average over the 600 ms long time series in relation to the linear RMS value. The 2nd harmonic component is relatively small with 2.5 % RMS in relation to the linear component while the 3rd harmonic impulse response has the highest contribution with 12.9 %. Also, the second even harmonic (4th) is small in RMS with 1.25 % of the linear component in contrast to the 5th harmonic with 7.56 % RMS. The decreasing share of higher order harmonics is somewhat expected from the series expansion kind of the Taylor series.

The responses for the 3rd and 4th harmonic are inverted with respect to the amplitude-axis as compared to the linear impulse response, i.e., they exhibit a local maximum where the linear response exhibits a minimum. This surprising observation is discussed in the next section.

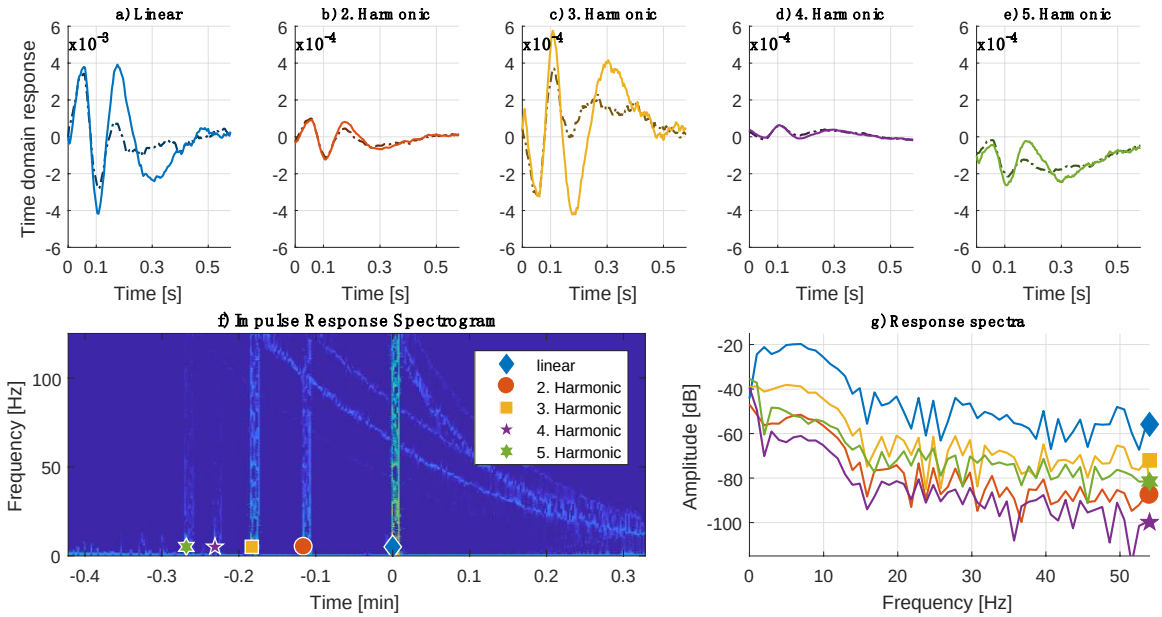

Figure 7. Time-domain representations of harmonic components are shown in Panels a)-e) with the solid line showing the DNN response to the attended sweep and the dotted line to the unattended sweep. Panel f) shows the inverse spectrogram that is acquired by convolving the DNN sweep response with the inverse filter followed by a short-time Fourier transformation. The linear (blue) and nonlinear components (red, yellow, purple and green) are labeled with corresponding markers. Individual spectra of each attended component are plotted in Panel g) with identical color coding and marker-style as in Panel (f).

The spectral profiles of linear and harmonic components (Figure 7g) have a similar general appearance with a high response at frequencies below 10 Hz and a relatively low response at frequencies above 15 Hz. The local peaks at 50 Hz can be attributed to the frequency of AC current. To measure the total nonlinear contribution the broadband average power ratio R between linear and nonlinear impulse responses is calculated as

$$R = 10 \cdot log_{10} \left[\frac{P_{lin}}{\sum_{i=2}^{5} P_i} \right],$$

where P denotes the power of the indexed components, respectively. We obtain a ratio of R = 6.1 dB, which corresponds to a 24.5% contribution of nonlinearities to the output of the DNN.

4. Discussion

This study investigated if EEG responses to continuous speech can be modeled with a nonlinear DNN. In the following, the properties of the DNN for the auditory attention task are discussed, as well as limitations of the model and future directions for model extensions.

4.1 Analysis of the Model Response for Attended and Unattended Stimuli

The separate two-channel input to the DNN network proposed in this paper allowed a separate analysis of attended auditory processing versus unattended, respectively. Attention effects were analyzed by contrasting model responses to the attended and unattended case in the temporal and the spectral domain. The response of the model to a short audio burst on the attended and unattended channel separately showed significant differences in excitation. The largest effect was observed between 110 ms and 220 ms where attended processing deviates from the unattended processing. Since auditory attention (here the focus on a specific speaker) is the result of a conscious decision and therefore related to top-down processing, a mechanism should be active at a certain point along the auditory pathway (and consequently, with a certain delay between stimulus and response) that selects the task-related features (e.g., the speech representations from the selected speaker [Mesgarani and Chang 2012]). Our data shows a deviation between the attended and unattended streams for time delays as low as 100 ms, which suggests that attention could modulate relatively early stages along that pathway which is in line with a study from Akram, Simon, and Babadi [2017] who found the latency of 100 ms especially important for auditory attention. However, the strongest attention-based modulations are observed for delays between 130

ms and 370 ms, consistent with other studies. Lalor and colleagues [Lalor and Foxe 2010; Power, Foxe, Forsde, Reilly and Lalor 2012] performed an auditory evoked spread spectrum analysis for which a group average of attended processing was reported with three peaks at approximately 40 ms, 90 ms and 160 ms, respectively. Our corresponding peaks at 56 ms, 104 ms and 172 ms show small (16 ms, 14 ms and 12 ms) delay differences in comparison. These results may arise from differences in the calculation of the envelope used or from methodological delays between the two approaches. We also observed a significant difference between attended and unattended response from the short time windows between 0 ms and 25 ms physiologically not plausible since attention effects have not been reported for such low time delays. We attribute this to a property of the learning algorithm which should avoid a drift of excitation patterns over longer time periods, and therefore the integral of the click impulse response should be zero. Presumably, the asymmetric excitation pattern of the unattended processing – skewed slightly to negative values as shown in Figure 6 – requires the network to initialize the output with a small positive offset.

4.2 Contribution of Nonlinear Processing

The analysis using an exponential sweep as input to the DNN model allowed for a separation of linear and nonlinear contributions, with the result that the most pronounced contribution is associated with the linear component, while the nonlinear components combined contribute approximately 25% to the output RMS. Odd numbered harmonic responses contribute the larger part of the nonlinear DNN output. This observation is plausible as we applied a symmetric tanh activation function that should only create nonlinearities of odd harmonic order. Harmonic responses of even order are thus presumably a consequence of the threshold adaptation in each neuron of the DNN. While the average of both input stream is zero due to the z-scoring in the preprocessing step, the speech envelopes are inherently skewed, and the median of both input streams is below zero. The subsequent adjusting through the learned threshold values correct the input signal with an offset to match them better to the working area of the DNN. This also explains why no effect of attention is noted for the harmonic responses of even order. Also, the amplitude inversion of the output function in the third and fourth harmonic is interesting: It seems these terms compensate a high model output from the linear and quadratic term, and therefore play the role of effectively inhibiting the modeled EEG response. Without further detailed research it cannot be excluded that the selected activation function has an influence on the type of nonlinear effects shown here. Most likely it is even crucial to use different activation functions at different levels of the network to enable the full potential of nonlinear modelling. Pilot tests with other activation functions have shown that the nonlinearities that occur can have different properties. In order not to exceed the scope of this paper, no further analysis was carried out, although there are certainly still interesting aspects to be investigated in this area. Also the ratio between linear and nonlinear components is most likely tied to the model architecture and depth and the observed ratio only a first impression of the order of magnitude. Although the method presented does not provide irrefutable evidence for nonlinear aspects in auditory speech processing, the approach opens the possibility to investigate this matter both qualitatively and quantitatively. Future work should therefore include more detailed assessment of nonlinear mapping functions in EEG data and their relation to higher cognitive processes, since this approach might enable a better understanding of sensory processing. Generally, the analysis of nonlinearities in the DNN indicate that nonlinear components substantially contribute to the performance in encoding models, which could be an advantage over the well-established linear regression analysis [Lalor, Power, Reilly and Foxe 2009; Mirkovic, Bleichner and De Vos 2016; Picton, Woods, Baribeau-Braun and Haeley 1977].

4.3 Spectral Analysis of Modeled Signals

The separate spectral analysis of model responses has shown that generally a higher spectral power is observed for the attended case than for the unattended input neuron. Note the input signals had been normalized before, i.e., the higher activation is a property of the model and not a mere artifact of higher energy of the input stimulus. The highest differences were observed for the 5 Hz component, which was reported to be modulated in attention tasks before: Hickok and colleagues [Hickok and Poeppel 2007] showed a higher activation in the theta band range with a dominant spectral modulation power from 4 to 7 Hz is related to top-down selection of an attended speaker who produces speech with a typical syllable rate. In our experiments, we found the regions with the highest 5 Hz power to be located in the region of auditory cortices and therefore in the vicinity of Brodmann area 22 also related to theta band activity (4–7 Hz) and speech modulation processing [Giraud and Poeppel 2015]. Also it was found that theta band activity indeed encodes syllable encoding [Pefkou, Arnal, Fontolan and Giraud (2017)]. Therefore, we assume the specific higher activation around 5 Hz observed in our model arises from the top-down modulation of attention on a syllable level.

4.4 Model Input features

Since the DNN is trained on a simple representation of acoustic stimuli only, the EEG model output does not closely resemble the EEG signal actually measured, rather the envelope-related component of EEG data. On the one hand, this is a drawback of the model, since it is unable to accurately reconstruct EEG. Hence, it is probably not well-suited for solving decoding tasks (i.e., determining which speaker was attended in a multi-speaker scenario) by comparing the predicted and the measured EEG signal. Furthermore, our model requires the separate speech envelopes of both speakers, which are not available in real-world scenarios. On the other hand, the model produces physiologically plausible responses, which sheds light on the underlying mechanisms, for instance, electrode positions in auditory attention. This aspect could contribute for designing better brain-computer interfaces (BCIs) and attention decoding algorithms, which could be applied for creating BCI-controlled hearing devices that enhance the attended source, for instance by spatial filtering of multi-microphone signals in hearing aids [Van Eyndhoven Francart and Bertrand 2016]. Future research for deep encoding models should incorporate additional features to learn the mapping from acoustic stimuli to electrical brain activity, since it is unlikely that the speech envelope fully covers attention-related processes. For instance, the feature set proposed by Di Liberto, O'Sullivan, and Lalor [2015] that extends the feature range by including spectrograms, phonemes as well as phonetic features including manner and place of articulation could be incorporated to determine a more accurate mapping function for future models. Such sophisticated input improvements are also likely to lead to finer granular physiological outcomes with respect to auditory attention, but at present the simplistic approach is a good starting point to verify the model assumptions and compare the results with previous studies. This approach would probably require a larger data set than the one used in the current study to train a DNN. The corresponding databases could be recorded with mobile EEG systems to provide sufficient training material, which would also have the advantage of including real-world data (in contrast to well-controlled lab situations). For these datasets, long short-term memory networks [Hochreiter and Schmidhuber 1997] could be considered, which would be interesting due to their recurrent structure that might be beneficial for modeling top-down processes in auditory attention.

5. Summary

This paper investigated using nonlinear deep neural networks for modeling the EEG response of listeners for an auditory attention task. The aim was not to predict the EEG data as accurately as possible, but to gain insights into speech processing despite the poor SNR. The test data was continuous speech which represents a natural stimulus that the normal-hearing listeners could easily follow in the spatial two-speaker scenario analyzed here. A neural network was trained to predict the concurrent EEG from the envelopes of both speech streams. The resulting model was analyzed by using traditional click-like stimuli for the network that exhibits separate input neurons for the attended and the unattended speech envelope stream, respectively. The response to these stimuli exhibits extreme values at time lags between input and response at 50, 110, 170 and 300 ms, which were differently modulated by attention: Late model responses (170 and 300 ms) differed strongly for the attended and unattended case, while earlier responses differed to a lesser extent (110 ms) or not at all (50 ms). We investigated the contribution of linear and nonlinear components of the mapping from envelopes to neurophysiological data by using exponential sweeps as input to the network, which can be used for separating linear and nonlinear properties of the DNN. The analysis shows that the linear mapping is the most important, while nonlinear components still play a major role with a 25% contribution to the average output of the model.

Acknowledgements

This work was funded by DFG (Research Unit FOR 1732 "Individualized Hearing Acoustics", Cluster of Excellence 1077/2 "Hearing4all"). The Titan Xp GPU used for designing the neural network in this research was donated by the NVIDIA Corporation.

List of Abbreviations

DNN – Deep Neural Network

EEG – Electroencephalography

BCI – Brain-Computer-Interface

RMS – Root-Mean-Square

STFT – Short-Time Fourier Transformation

References

Akram, S., Simon, J. Z., & Babadi, B. (2016). Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1896-1905. https://doi.org/10.1109/TBME.2016.2628884

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, *86*(1), 117-128.

Davis, P. A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of neurophysiology*, *2*(6), 494-499. https://doi.org/10.1152/jn.1939.2.6.494

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Derleth, R. P., Dau, T., & Kollmeier, B. (2001). Modeling temporal and compressive properties of the normal and impaired auditory system. *Hearing Research*, *159*(1-2), 132-149. https://doi.org/10.1016/S0378-5955(01)00322-7

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457-2465. https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., & Simon, J. Z. (2011). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, *107*(1), 78-89. https://doi.org/10.1152/jn.00297.2011

Dozat, T. (2016). Incorporating nesterov momentum into adam, Proc. ICLR 2016.

Van Eyndhoven, S., Francart, T., & Bertrand, A. (2016). EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Transactions on Biomedical Engineering*, *64*(5), 1045-1056. https://doi.org/10.1109/TBME.2016.2587382

Farina, A. (2000, February). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*. Audio Engineering Society.

Getzmann, S., Jasny, J., & Falkenstein, M. (2017). Switching of auditory attention in "cocktail-party" listening: ERP evidence of cueing effects in younger and older adults. *Brain and cognition*, *111*, 1-12. https://doi.org/10.1016/j.bandc.2016.09.006

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, *15*(4), 511. https://doi.org/10.1038/nn.3063

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393. https://doi.org/10.1038/nrn2113

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., & Kollmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, *2009*, 6. https://doi.org/10.1155/2009/298605

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European journal of neuroscience*, *31*(1), 189-193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of neurophysiology*, *102*(1), 349-359. https://doi.org/10.1152/jn.90896.2008

Martinez, A. M. C., Mallidi, S. H., & Meyer, B. T. (2017). On the relevance of auditory-based Gabor features for deep learning in robust speech recognition. *Computer Speech & Language*, *45*, 21-38. https://doi.org/10.1016/j.csl.2017.02.006

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233. https://doi.org/10.1038/nature11020

Mirkovic, B., Debener, S., Jaeger, M., & 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. https://doi.org/10.1088/1741-2560/12/4/046007

Mirkovic, B., Bleichner, M. G., De Vos, M., & Debener, S. (2016). Target speaker detection with concealed EEG around the ear. *Frontiers in neuroscience*, *10*, 349. https://doi.org/10.3389/fnins.2016.00349

Novak, A., Simon, L., Kadlec, F., & Lotton, P. (2009). Nonlinear system identification using exponential swept-sine signal. *IEEE Transactions on Instrumentation and Measurement*, *59*(8), 2220-2229. https://doi.org/10.1109/TIM.2009.2031836

Pefkou, M., Arnal, L. H., Fontolan, L., & Giraud, A. L. (2017). θ-Band and β-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech. *Journal of Neuroscience*, *37*(33), 7930-7938. https://doi.org/10.1523/JNEUROSCI.2882-16.2017

Picton, T. W., Woods, D. L., Baribeau-Braun, J., & Healey, T. M. (1977). Evoked potential audiometry. *J Otolaryngol*, *6*(2), 90-119.

Poirier, F. J., & Wilson, H. R. (2006). A biologically plausible model of human radial frequency perception. *Vision research*, *46*(15), 2443-2455. https://doi.org/10.1016/j.visres.2006.01.026

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & 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. https://doi.org/10.1111/j.1460-9568.2012.08060.x

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in amplification*, *12*(4), 283-299. https://doi.org/10.1177/1084713808325306

Sturm, I., Lapuschkin, S., Samek, W., & Müller, K. R. (2016). Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods*, *274*, 141-145. https://doi.org/10.1016/j.jneumeth.2016.10.008

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. Proc. *European conference on computer vision*, pp. 818-833. https://doi.org/10.1007/978-3-319-10590-1_53