


Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech

Tobias de Taillez,*  Birger Kollmeier and Bernd T. Meyer

Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität, Oldenburg 26129, Germany

Keywords: auditory, auditory processing, hearing, neural networks, signaling pathways

Abstract

Previous research has shown that it is possible to predict which speaker is attended in a multispeaker scene by analyzing a listener's electroencephalography (EEG) activity. In this study, existing linear models that learn the mapping from neural activity to an attended speech envelope are replaced by a non-linear neural network (NN). The proposed architecture takes into account the temporal context of the estimated envelope and is evaluated using EEG data obtained from 20 normal-hearing listeners who focused on one speaker in a two-speaker setting. The network is optimized with respect to the frequency range and the temporal segmentation of the EEG input, as well as the cost function used to estimate the model parameters. To identify the salient cues involved in auditory attention, a relevance algorithm is applied that highlights the electrode signals most important for attention decoding. In contrast to linear approaches, the NN profits from a wider EEG frequency range (1–32 Hz) and achieves a performance seven times higher than the linear baseline. Relevant EEG activations following the speech stimulus after 170 ms at physiologically plausible locations were found. This was not observed when the model was trained on the unattended speaker. Our findings therefore indicate that non-linear NNs can provide insight into physiological processes by analyzing EEG activity.

Introduction

Normal-hearing listeners are able to focus on one speaker in a multi-speaker scenario and to effectively suppress concurring speakers (Mesgarani & Chang, 2012). This ability can be affected in hearing-impaired listeners (Shinn-Cunningham & Best, 2008), which can partially be compensated by spatial filtering in multichannel hearing aids, for example, by beamforming for enhancing speech from one specific speaker (Haykin & Liu, 2010). A passive system that automatically identifies and enhances the attended speaker is highly desirable in this application scenario. Attentional effects on auditory stimuli are detectable by analyzing rhythm-correlated brain activity like auditory steady-state response (Ding & Simon, 2012) or the p300 responses (Polich, 1986; Spencer & Polich, 1999). However, these techniques fail for continuous speech stimuli (Ding & Simon, 2012; Horton *et al.*, 2014). Another approach is to track the lateralization of alpha power bands for decoding spatial attention. The effect is strongest when shifting attention, while it is less pronounced when a sound source is continuously attended and is

therefore also not suitable for continuous attention decoding (Kerlin *et al.*, 2010). The current state-of-the-art technique for detecting auditory attention for continuous speech streams was proposed by Aiken & Picton (2008): this approach is based on the hypothesis that speech features of the attended speech stream such as the envelope are represented in neural brain activity. A source can be determined by measurement of this activity and subsequent correlation with individual speech streams of the speech sources. This was shown for electroencephalography (EEG) (Aiken & Picton, 2008; Di Liberto *et al.*, 2015; Mirkovic *et al.*, 2015; Biesmans *et al.*, 2017), magnetoencephalography (Akram *et al.*, 2016) and electrocorticography; Mesgarani & Chang, 2012).

Previous studies reached accuracies of 88% with a 60 s analysis window using EEG data (Mirkovic *et al.*, 2015) for decoding attention in a two-speaker paradigm. EEG offers high temporal resolution, is not invasive and is available as mobile system (Debener *et al.*, 2012).

The previously mentioned studies (Aiken & Picton, 2008; Mesgarani & Chang, 2012; O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015, 2016; Biesmans *et al.*, 2017) successfully applied a linear superposition of measured brain activity for stimulus reconstruction or attention decoding. In these studies, the authors point out two potential shortcomings of this approach. First, the auditory processing can be assumed to be non-linear due to dynamic compression, loss of fine structure information by integration in the auditory nerve or non-linear neuronal processes. Second, the mapping of this process is probably not invertible by a linear approach. The fairly low correlation values between the reconstructed and the attended

Correspondence: Tobias de Taillez, as above.
E-mail: tobias.de.taillez@uni-oldenburg.de

Received 8 June 2017, revised 23 November 2017, accepted 27 November 2017

Edited by John Foxe

Reviewed by Edmund Lalor, Trinity College Dublin, Ireland; and Hanjun Liu, Sun Yat-sen University, China

The associated peer review process communications can be found in the online version of this article.

envelope ($r = 0.054$; O'sullivan *et al.*, 2014) also support this assumption.

Therefore, in this study, we investigate non-linear machine learning methods with the aim of a better decoding of listeners' attention. Non-linear models for speech modeling have been used in automatic speech recognition (ASR) for many decades, but network topologies such as deep neural networks (NNs; Hinton *et al.*, 2012) or long short-term memory networks (Graves *et al.*, 2013) recently had a large impact on improving ASR systems. We propose to apply an artificial NN with a novel net architecture to replace the linear regression used in previous studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015) as an attention decoder. The network non-linearity reduces the error in the inverse mapping between EEG and audio stimulus to some extent. This approach also offers a tunable amount of parameters for a more sensitive and accurate reconstruction of the stimulus' envelope. A training paradigm is suggested in which consecutive *output* values of the net are considered for adjustment of model parameters, which contrasts with the standard procedure of using temporal context of *input* values. With this network, we investigate whether the non-linear model can profit from a wider EEG frequency range than previous models (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015), from a temporal segmentation of data, and whether the accuracy for efficient decoding can be increased.

A better understanding of how a machine learning algorithm inverts the brain's audio processing could also contribute to understanding physiological processes. While statistical models such as NNs are often considered to be black boxes, several methods exist to analyze the salient cues for classification learned by the net. An algorithm to analyze the relevance of the input features for producing the output (also referred to as heat mapping) was recently proposed (Bach *et al.*, 2015; Sturm *et al.*, 2016). In this study, the algorithm is used to identify where and when neural activity occurs that is relevant for decoding of auditory attention. This includes the location and time of electrode activity that later results in correct speaker decoding, which enables an analysis of important cues in auditory processing that arise from physiological processes, such as the stimulus-response delay time, the skull region associated with high relevance and an interaction of these two.

In summary, three main goals are addressed in this study: first, we investigate whether current machine learning methods can contribute to improve the decoding of listeners' attention. Second, different input representations and network architectures are analyzed to optimize model parameters. Third, the spatial and temporal cues for reconstructing the attended speech envelope are explored by analyzing model parameters and the input–output relation of EEG data.

Materials and methods

Acoustic stimuli

The data set gathered by Mirkovic *et al.* (2016) was made available for our study and contains synchronized data streams of audio and EEG signals. Excerpts from German audiobooks were used as speech material ('A drama in the air' by Jules Verne and 'Two brothers' by Grimm brothers) read by two different professional male speakers. The speakers were virtually placed at $+45^\circ$ and -45° azimuth by calculating the convolution of the stimuli with head-related transfer functions that were recorded in an anechoic chamber (Kayser *et al.*, 2009). Both stories were played back continuously over tube in-ear headphones (E-A-RTONE 3A) with an external electromagnetically shielded sound generation box to minimize the interference of the headphones with EEG electrodes. The audio

presentation was paused every 10 minutes, and participants were asked to answer content-related questionnaires subsequently. This procedure was repeated five times, which resulted in 50 min of data per subject.

Participants

Twenty healthy, normal-hearing listeners participated in the experiments (mean age 25, eight male, one left-handed). They were randomly assigned to attend exclusively to the left or the right audiobook while keeping the number of participants per group even (10 for both sides). All participants were paid for their participation equally.

EEG recording procedure and preprocessing

EEG data were collected with a BrainAmp EEG amplifier system (BrainProducts, Gilching, Germany). A cap with a 96 Ag/AgCl electrode layout, equidistantly placed with a central frontopolar site as ground and a nose-tip reference, was used (Easycap, Herrsching, Germany) with 500-Hz sampling frequency. These electrodes do not follow the naming convention and/or locations of the 10/20-system (Jasper, 1958). Therefore, references to electrode positions are given with a corresponding 10/20-position. From the available 96 electrode locations, 12 were not used due to the presence of an additional ear-centered EEG devices called 'cEEGrid' (six electrodes on each side). To obtain a homogeneous data set, the cEEGrid data were not used in this study so that 84 channels were available. Divergent from the initial approach (Aiken & Picton, 2008), the frequency band used was widened due to findings from Di Liberto *et al.* (2015) who found that for phoneme level decoding, information up to 45 Hz can be useful. The preprocessing was performed off-line using customized Matlab/Python scripts and consisted of band-pass filtering between 1 and 32 Hz (2–8 Hz for resembling the linear regression approach used in earlier studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015) for comparison) as well as a subsequently down-sampling to 64 Hz. Data were re-referenced to a common average reference. No further artifact correction was conducted. The clean speech material was transformed to their respective absolute envelope by a Hilbert transformation, low-pass filtered with 8 or 32 Hz (see above) and down sampled to 64 Hz to match the EEG data. EEG data channels and audio envelopes were normalized to ensure zero mean and unit variance. For each participant, the five 10-min blocks of EEG recording and corresponding audio streams were concatenated so that a 50-min long data block was created. During post-processing, the data blocks of four participants were excluded: two data sets from the group that attended Story 1 (one due to a low score in the questionnaire and one due to technical problems) and two randomly selected sets from the group attending Story 2 (to reestablish a quantitatively evenly distributed data set). This resulted in the final data set with data from 16 subjects.

Neural network structure

The linear regression used in previous studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015) is nearly equivalent to a NN using a linear activation function and no hidden layers. For comparability, this linear mapping was therefore defined as the starting point for experiments with NNs and is also tested in the experiments to compare the previous linear approach with the proposed NN structure. Subsequently, more features of deep learning were added ranging

from typical non-linear ‘tanh’ activation functions over additional hidden layers to methods for avoiding overfitting to the data such as dropout (Srivastava *et al.*, 2014). For attention decoding, we introduce a novel training scheme that performs a sample-wise prediction of the envelope but is able to include temporal context of the *output* (i.e., many consecutive samples) during training. While the inclusion of temporal context for the *input* is standard procedure, that is, in ASR (Hinton *et al.*, 2012), it has not been considered for output values in the context of physiological data. The resulting network was implemented in Keras (Chollet, 2015)/TensorFlow (Abadi *et al.*, 2016). The NN-based processing is illustrated in Fig. 1: for each point in time t_0 , a NN is trained to predict the current envelope value \hat{e}_0 , taking into account a temporal input context of $M = 27$ frames (corresponding to 420 ms; compare (Mesgarani & Chang, 2012)). During training, we use a training prediction window of length L , which will produce L adjacent envelope predictions using the same NN weights. The resulting time series $\hat{e}(n)$ is compared to the original (attended) envelope $e_a(n)$. The match between $e_a(n)$ and $\hat{e}(n)$ is measured with the cost function [mean-squared error (MSE) or correlation, respectively], which in turn is used to update the weights of the NN. This sample-wise approach allows to analyze long time series with a moderate amount of parameters. A smaller number of parameters allow successful training on comparatively small amounts of data (such as the EEG data in this study, which amounts to 50 min per participant). The NN consisted of an adjustable number of hidden layers with neurons of the non-linear activation and one output neuron with a linear activation. This modular setup provided the opportunity to investigate a range of neural net parameter settings. In speech processing, an intentional omission of training samples was found to be beneficial for generalization of classifiers. This technique is referred to as dropout. A dropout layer was included after each layer (except the output layer) with dropout strength of $\alpha = 0.25$, so that with a chance of 25%, the activation of a certain neuron will be set to zero. Further, two cost functions were used to measure the distance between the estimated and the actual attended envelope: first, the MSE was used as a cost function C_{MSE} . This is a widely used approach in deep learning (Bengio *et al.*, 2007; Bengio, 2012; Sturm *et al.*, 2016). With this cost function and the simple neural net, a linear regression is resembled by

$$C_{\text{MSE}}(\hat{e}_a, e_a) = \overline{[e_a - \hat{e}_a]^2},$$

with the attended e_a and predicted \hat{e}_a envelope. Second, a correlation-based cost function was employed that aims to maximize the correlation of prediction and stimulus:

$$C_{\text{corr}}(\hat{e}_a, e_a) = 1 - \frac{\text{cov}(\hat{e}_a, e_a)}{\text{std}(\hat{e}_a) \cdot \text{std}(e_a) + \varepsilon}.$$

A cost of zero implies a perfect correlation between attended envelope and predicted envelope, while cost values of 1 correspond to no correlation. Values between 1 and 2 indicate negative correlation. ε is a random number in the machine precision range (e.g., 10^{-30}) to avoid division by zero in rare cases.

Evaluation cycle

In the following, we describe the training and testing procedure of the neural net. First, 20% consecutive data samples were chosen from a randomly selected participant and starting point. The remaining 80% of the data were used for training as described in

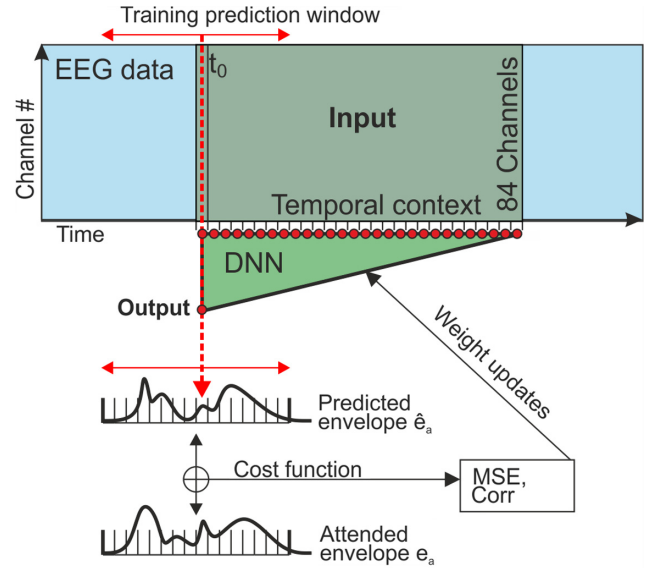


FIG. 1. Illustration of the neural network (NN) training process. From the electroencephalography (EEG) training set (blue), a time series of predicted envelope samples is obtained by shifting the NN (green) over the training prediction window (red arrows). For each t_0 , a temporal context (typically 27 frames) is included. The cost function [mean-squared error (MSE) or correlation] is calculated and used to update the weights of the NN. This process is repeated until convergence.

Section Neural network structure. The 20% were split into two 10% blocks, one for cross-validation (CV set) and one for evaluation (test set). The training algorithm uses the *Nadam* optimization criterion, which effectively prevents to get stuck in local minima in the error distribution by adaptively changing the learning rate. Training was performed according to standard deep learning procedure, that is, ‘early stopping’ was used as a method to prevent overfitting (Caruana *et al.*, 2001). Training was continued until no loss reduction was achieved on the CV set for five epochs. For evaluation, the NN weights are fixed and each envelope value of the test set is predicted with the same temporal input context as during training (27 samples). To determine how well this prediction matches the attended stimulus envelope, time windows of $\hat{e}(n)$ were correlated with $e_a(n)$ and the unattended envelope $e_u(n)$, respectively. The higher correlation value was counted as decision and rated against the ground-truth. The time windows had an overlap of at least 80% to ensure systematic testing. The resulting accuracy vector was averaged over time. Various analysis time window lengths were evaluated by this process. In a last step, the average accuracy P of a cycle is transformed into the information transfer rate (i.e., the effective bitrate) (Wolpaw *et al.*, 1998).

$$R = \log_2 N + P \cdot \log_2 P + (1 - P) \cdot \log_2 \frac{1-P}{N-1}.$$

N denotes the number of classes (in this study: $N = 2$ for attended and unattended streams). For comparability, the bitrate is scaled with the time window length to yield the number of correct decisions per minute (bit/min). To cover all participants and data points in test and training, the procedure is repeated 350 times; one iteration is referred to as *evaluation cycle*. Due to the random weight initialization during training, the repetition and the later averaging result in convergence and stable results. The random weights for each cycle also enforce intra-subject training.

Relevance of electrodes for attention decoding

The use of NNs for envelope reconstruction enables the application of recently developed algorithms that focus on a better understanding of processing principles in NNs. Bach *et al.* (2015) introduced an algorithm that estimates the contribution (or relevance) of NN input values to a correct classification result. This is achieved by propagating the relevance of a NN output to the input neurons. We use this approach to investigate the relevance of electrodes (and implicitly electrode positions) for attention decoding. This should separate input components with a small contribution (channels that do not encode the speech envelope and/or are dominated by noise) and important electrodes. The relevance algorithm allows an analysis of the complete input matrix that extends over time. Therefore, the relevance of electrode activity relative to an acoustic event (e.g., a word or syllable onset) can be studied. The portion of the output neuron that is explained by the specific input neuron is given by

$$r_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_j z_{ij}} r_j^{(l+1)} \text{ with } z_{ij} = x_i^{(l)} w_{ij}^{(l,l+1)},$$

where the relevance r_i of neuron i in layer l is calculated recursively from all upper layer relevances r_j . r_i is weighted with the activation z_{ij} between two neurons i and j . x_i represents the neuron activation, and w_{ij} , the weighting parameter. The relevance of the output neuron is defined as the match (correlation or MSE) between predicted and attended envelope. This match is calculated with a moving, overlapping time window of 16 samples (250 ms) and therefore provides results on this relatively short time scale. To identify cues that are relevant for a *correct* decoding, only time samples with a positive correlation were evaluated with the relevance algorithm. For each evaluated sample of the test set, an array of input relevance values was calculated, which are subsequently averaged over time.

Results

Decoding performance

In this section, we report the performance for decoding listeners' attention in dependence of four important parameters that are expected to affect the score. These are (a) the EEG bandwidth, (b) the length of the training prediction window, (c) the type of cost function and (d) the length of the analysis window (cf. Fig. 1). Also, the NN depth was varied with the best results were achieved with one hidden layer (data not shown). This presumably arises from the limited amount of training data, as deep NNs (i.e., with several hidden layers) are prone to overfitting when a larger number of weights are used for a rather small training set as it was performed here. The parameters (a–c) were optimized sequentially as a complete search of the four-dimensional search space was computationally not feasible. Our intuition is that a complete search is not required either, as for instance, the type of cost function should be optimal both for wideband as well as narrowband EEG data, that is, the factors presumably do not interact with each other. However, the full range of analysis window duration was analyzed for each configuration.

The results for identifying the attended speaker in terms of bits/minute are shown in Fig. 2. Note that the first tested condition in Panel (a) corresponds to the linear regression from previous studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015): a fully connected two-layer net without activation function is identical to a linear mapping. Using a narrowband EEG and envelope bandwidth, a training

prediction window of one sample, and the MSE cost function, the procedure used in Mirkovic *et al.* study (2016) is recreated, which serves as the linear baseline. It can be shown that a strong overall impact is achieved by increasing the bandwidth from 2–8 Hz to 1–32 Hz. When using broadband information, the transmitted bitrate is increased to 355% in comparison with the 2–8 Hz frequency range. These results were obtained with the C_{MSE} cost function and a training prediction window of one sample, which corresponds to the linear regression approach from previous studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015). Statistical significance of differences was tested with a one-sided Wilcoxon signed-rank test as the data are not normally distributed; the result of the significance test is denoted in the plot. Figure 2b shows the effect of the training prediction window, that is, the number of consecutive predicted samples of the net used to calculate correlation or MSE error during training. The bitrate increases with longer prediction windows and saturates for durations of 12 and 16 samples. Due to this saturation, and as the computational cost for training increases exponentially with longer windows, higher values for the window duration were not considered. For the training prediction window of 16 samples (250 ms), both cost functions were evaluated (Fig. 2c): maximizing the correlation between predicted and real envelope significantly increases the bitrate to 110% compared to the traditional MSE criterion.

On single frame level, a long analysis window provides the best accuracy with a classification rate of 0.976 for a 60 s window. However, the analysis of rather long segments introduces a significant delay. Further, a segmentation into smaller chunks and subsequent integration could provide an increased performance. Figure 2d shows how different durations of analysis windows affect the decoding performance. In the linear case, the analysis window duration is not a crucial factor. For non-linear neural nets, however, decoding performance is greatly increased when moving from 60 s to the optimum 2 s analysis duration. Choosing a window of two-seconds therefore constitutes the best trade-off between frame-wise accuracy ($P = 0.678$ on average) and subsequent integration of single decisions. Overall the evaluation over 350 cycles for each condition shows a maximum in decisions per minute for the correlation-based network including a bandwidth of 1–32 Hz and 16 samples filter length. In comparison with the linear regression from previous studies ($P = 0.88$ for $t = 60$ s corresponding to $0.47 \frac{\text{bit}}{\text{min}}$, compare (Mirkovic *et al.*, 2015)), this is a factor of 7.5 or a bit rate of $3.6 \frac{\text{bit}}{\text{min}}$ (compare Fig. 2d, black/diamond line). In terms of decisions per time window, this bit rate corresponds to a correct decision every 16.7 s. This is also reflected in the average correlation coefficient between prediction and attended envelope. For the linear regression, an average correlation coefficient of $\text{corr}_{\text{LR}} = 0.030$ was achieved whereas the highest performance condition yields $\text{corr}_{\text{Best}} = 0.131$.

Relevance of electrodes and time lags

The configuration that yields the highest decoding performance (cf. Section Decoding performance) uses an EEG frequency range of 1–32 Hz, a 16 sample training prediction window, and correlation as a training cost function. The EEG data and the corresponding envelopes for this configuration are used to analyze the relevance of time lags between EEG and audio and of electrode channels (and therefore implicitly electrode positions). The relevance matrix was calculated by averaging over 350 evaluation cycles and sorted by energy per channel. The assumption behind this is that a high relevance only could appear at distinct points in time which would result in high root-mean-squared (RMS) values for an electrode in the

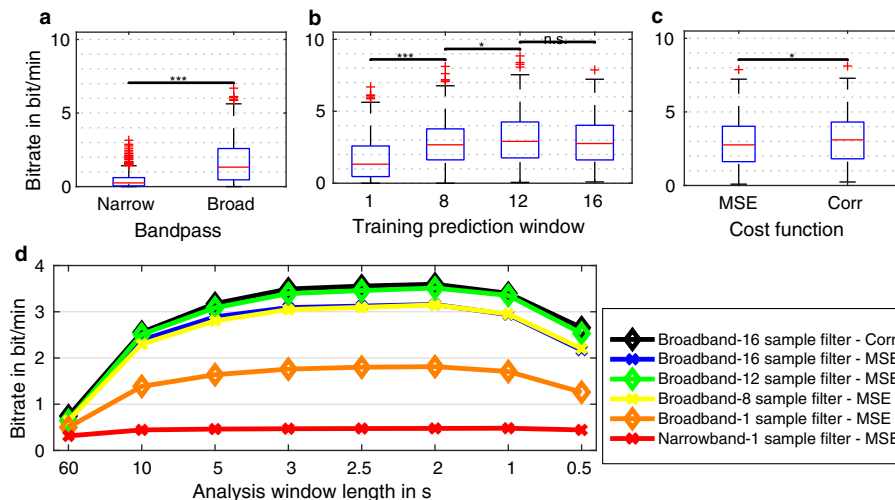


FIG. 2. Decoding performance in terms of bits/minute, obtained from 350 evaluation cycles for the respective condition. Significant differences $P < 0.001$ are indicated by three and $P < 0.05$ by one asterisk, respectively. (a) Performance for narrowband-filtered (2–8 Hz) and broadband-filtered (1–32 Hz) data, respectively. (b) Performance for different training prediction window lengths with *mean-squared error* (MSE) cost function and broadband data. (c) Difference in performance between MSE and correlation cost function, respectively. Length of training prediction window is 16 samples with broadband filtering. (d) Comparison in performance of all unique conditions from (a), (b) and (c) over evaluation window length. The red line corresponds to the linear regression also used in previous studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015).

relevance analysis. Therefore, the channels with high RMS presumably carry more information. Figure 3a shows the averaged relevance matrix with sorted channels. A high relevance can be observed at around 47 and 172 ms. Figure 3b shows the RMS per channel and a threshold for which the channel RMS starts to increase considerably. Channels above this threshold will be evaluated for Fig. 4. In Fig. 3c, topographic representations are displayed for these two distinct time points and in grand average over all time points. Clear relevance clusters can be distinguished above both ears for each time point. Additionally, an occipital lateral relevance cluster on the left hemisphere is observed for the early time point. The relevance of time points averaged over all channels is depicted in Fig. 3d. The electrodes of the 10/20-system close to the relevance clusters would be TP9/TP7/T7 as well as TP10/TP8/T8. If only the above-threshold channels are considered, the two peaks at 47 and 172 ms are clearly visible (compare Fig. 3e). These findings correspond to the case for which the reconstructor was trained on the attended stimulus.

In an additional experiment, we investigate the relation of EEG data with the envelope of the unattended speaker. This is achieved using the envelope of unattended speech as the training target. If the importance of EEG activity is related to the speaker attention task (and not a mere product of bottom-up processing), a clear difference between the mapping learned by the neural net should emerge. The results contrasting a training with unattended and with attended speech are shown in Fig. 4. The main peak for the attended case (120–300 ms) cannot be observed for unattended speech, which means that EEG data provide little information for reconstruction of the unattended speaker (which is in line with the good decoding results from the attended speaker shown in Fig. 2).

Discussion

Decoding performance

The aim of this study was to provide a method for decoding listeners' attention based on machine learning methods and to improve

results obtained in earlier studies (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015) that exploited linear solutions for the decoding task (Mirkovic *et al.*, 2016). Compared to these approaches, decoding performance was increased by a factor of seven using neural nets, which can be attributed to four main factors: (i) an increased bandwidth of EEG data compared to the narrowband representation applied in linear approaches, (ii) the use of temporal context during training through a training prediction window, (iii) optimization of the cost function that measures the difference between the current predicted and the target envelope and (iv) temporal segmentation of data into relatively short observation chunks. These factors are discussed in the following.

- (i) Exploiting broadband EEG information: The performance increase depicted in Fig. 2a shows that an extension from narrowband (2–8 Hz) to broadband (1–32 Hz) EEG data boosts performance of envelope-related decoding. This is in line with findings from Di Liberto *et al.* (2015) who found that for speech reconstruction on phoneme level, EEG frequency bands up to 45 Hz is useful. Although the authors state that temporal speech modulations predominantly occur in the range of 2–8 Hz, it seems that cues useful for discrimination in a two-speaker scenario extend to much higher modulations. A possible explanation is that the *difference* between two envelopes on time scales as short as 31 ms (which corresponds to 32 Hz) is represented in the EEG data, which could be exploited for onset detection (which in turn would be helpful for attributing the EEG data to the attended speaker). A frequency-specific analysis could help to disentangle the contributions of low- and high-frequency modulations for this decoding task, which will be subject of future research.
- (ii) Usage of time prediction windows: It could be also shown that the usage of training prediction windows is useful to improve performance. While in machine learning tasks such as ASR, usually a temporal context is provided for the input (as it is also performed in this study), the extension of providing temporal context during training for the *output* has not been used before for time series predicted by neural nets. The inclusion of

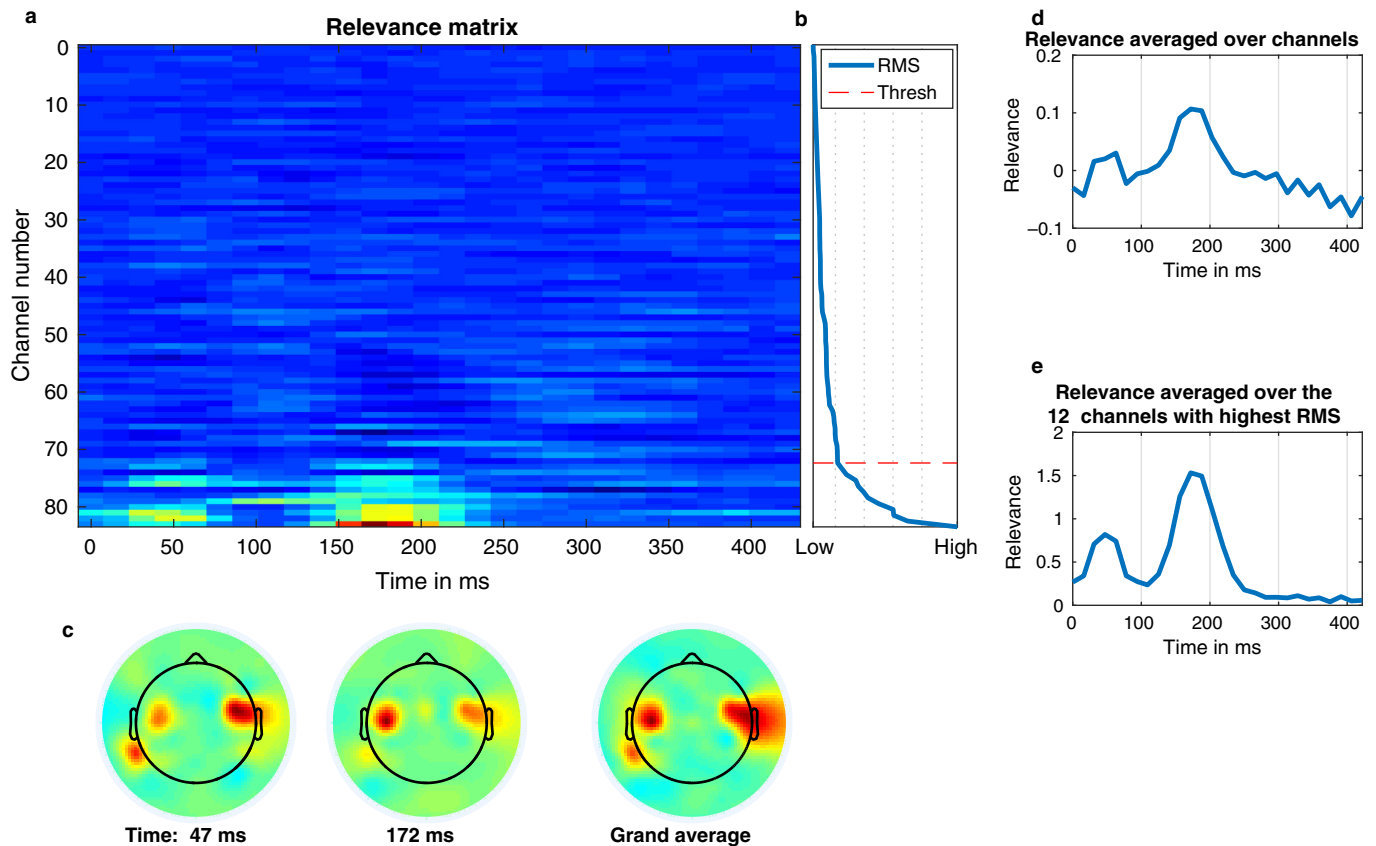


FIG. 3. (a) Relevance matrix averaged over 350 evaluation cycles trained for the attended speaker with correlation as loss function. Channels sorted after averaging for ascending root-mean-square (RMS). Blue indicates low and red high relevance. (b) RMS calculated for each channel. Threshold defines the point of the 'elbow' where channels begin to carry distinct relevance. (c) Topographic plots of electrode relevance distribution from the two columns of (a) at 47 and 172 ms (first two topographic plots on the left) and averaged over all time samples of (a) (plot on the right). Electrode locations of the 10/20-system in the vicinity of the relevance clusters: TP9/TP7/T7 and TP10/TP8/T8. (d) Relevance of (a) averaged over all channels. (e) Relevance of (a) averaged over just the 12 channels with highest RMS. Number of channels is defined by b).

temporal context of the output during training enabled the use of cost functions that require consecutive time samples such as correlation of time signals, which resulted in further improvements.

- (iii) The improvement achieved using a correlation-based cost function instead of MSE (compare Fig. 2b) might result from the invariance of correlation with respect to linear signal compression and to offsets between predicted and measured data. Such mismatches can occur as the predicted output cannot be normalized (which is the case when predicting short time segments with a neural net), but is compared to the normalized envelope. MSE on the other hand penalizes offsets and differences through linear scaling, while the selection criterion in the final selection step is also invariant to both factors.
- (iv) The length of the analysis window was analyzed for a range of combinations of the preceding three factors. When using a configuration similar to the linear approach applied earlier (O'sullivan *et al.*, 2014; Mirkovic *et al.*, 2015; red curve in Fig. 2d), a separation into smaller time segments only had minor effects, that is, temporal separation seems to be a negligible factor for linear decoding. For all other conditions that use either broadband data or longer training prediction windows, we observed a trade-off between the length of the analysis window and the number of binary decisions that can be made in a given (here: 1 min) time segment. Best decoding performance was obtained

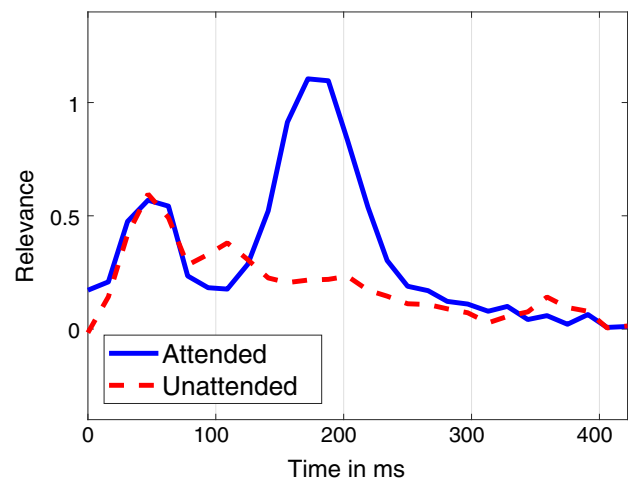


FIG. 4. Time lag relevance averaged over the 12 channels with highest RMS from the relevance matrix for the correlation loss function. Attended/unattended refers to training on the attended/unattended envelope. For both conditions, correlation was used as cost function together with 16 samples training prediction window length.

with 2-s windows, which is consistent over all test conditions. This duration therefore seems to be long enough to capture syllables reflected in the envelope, yet short enough to ignore

unrelated speech events, that is, temporally distant phonetic events do not influence each other. For decoding with a non-linear neural net, this factor increased overall decoding performance by a factor of seven.

Relevance of time lags and suppression of interfering speaker

The relevance analysis has shown two electrode clusters to be of special importance for envelope reconstruction. These clusters are located in an occipital/lateral position near to the auditory cortex. Both show high relevance at two distinct points in time (47 and 172 ms, respectively). The second relevance peak is in line with results for linear decoding from Mirkovic *et al.* (2015), and the relevant activations arise from regions that are physiologically plausible. This indicates that neural nets linked with relevance analysis can confirm earlier results from physiological studies (Power *et al.*, 2012). Moreover, the existence of an additional peak (47 ms, which was not reported in Mirkovic *et al.* study (2015)) hints at improved exploitation of the EEG input data for decoding, that is, it suggests that additional information can be derived with non-linear neural nets. When the network is trained on the unattended stimulus, the second relevance peak vanishes (Fig. 4), that is, the second maximum after 120 ms helps to discriminate between unattended and attended speakers, which could be caused by attentional top-down effects for suppressing interfering talkers. A first maximum at 47 ms has a high relevance for envelope reconstruction for the attended and unattended case and therefore does not contribute for attention decoding.

Conclusion

In this study, a NN was proposed to map listeners' EEG signals to the envelope of the attended speaker's signal in a spatial two-speaker scenario. The proposed network was trained using the temporal context of the output, which enabled the evaluation of time-dependent cost functions such as correlation. We found the duration of the analysis window to be an important parameter when using a non-linear neural net: When segmenting the data into short blocks and subsequent combination of single decisions, the decoding performance was strongly increased compared to 1-min segments. We also found that – in contrast to other reconstruction schemes – NNs profit from broadband EEG input (1–32 Hz). The resulting approach outperforms previous systems (Di Liberto *et al.*, 2015; Mirkovic *et al.*, 2015) by a factor of seven in terms of binary decisions per minute.

A relevance analysis of envelope reconstruction provided insight about temporal relations between stimulus and EEG activation (and implicitly location) and therefore indirectly about the temporal and spatial processing encoded by neural activity. Specifically, EEG activity around 170 ms was found to be relevant for reconstructing the envelope of the attended speaker. A back tracing of relevant activity to the electrode position showed relevant activity in physiologically plausible locations. This indicates that a relevance analysis of neural net models can provide insight into physiological processes involved in auditory attention.

Acknowledgements

This work was funded by the DFG (SFB/TRR 31 'The Active Auditory System', Research Unit FOR 1732 'Individualized Hearing Acoustics', Cluster of Excellence 1077/1 'Hearing4all'). The authors want to thank Bojana

Mirkovic and Stefan Debener emphatically for sharing their data and for fruitful discussions.

Conflict of interest

No potential conflict of interest was reported by the authors.

Author contributions

Tobias de Taillez and Bernd T. Meyer designed the study, developed the methodology and wrote the manuscript. Tobias de Taillez performed the experiments and did the data analysis. The study was supervised by Bernd T. Meyer and Birger Kollmeier. Birger Kollmeier contributed ideas both early in the study (general concept) as well as regarding experimental details such as the temporal segmentation of the EEG data.

Data accessibility

Due to copyright issues, the underlying data set of EEG and audio data cannot be made available for public. The code for the NN is accessible on the following github repository: https://github.com/tdeTaillez/neural_networks_auditory_attention_decoding

Abbreviations

ECoG, electrocorticography; EEG, electroencephalography; MEG, magnetoencephalography; MSE, mean-squared error; NN, neural network; RMS, root-mean-squared.

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A. *et al.* (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Aiken, S.J. & Picton, T.W. (2008) Human cortical responses to the speech envelope. *Ear Hearing*, **29**, 139–157.
- Akram, S., Presacco, A., Simon, J.Z., Shamma, S.A. & Babadi, B. (2016) Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, **124**, 906–917.
- 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**, e0130140.
- Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade*. Springer, Berlin, Heidelberg, pp. 437–478.
- Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*. The MIT Press, Cambridge, MA, pp. 153–160.
- Biesmans, W., Das, N., Francart, T. & Bertrand, A. (2017) Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE T. Neur. Sys. Reh.*, **25**, 402–412.
- Caruana, R., Lawrence, S. & Giles, C.L. (2001). Overfitting in neural nets: back-propagation, conjugate gradient, and early stopping. *Advances in Neural Information Processing Systems*. The MIT Press, Cambridge, MA, pp. 402–408.
- Chollet, F. (2015) Keras, *GitHub*, <https://github.com/fchollet/keras>
- Debener, S., Minow, F., Emkes, R., Gandras, K. & Vos, M. (2012) How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology*, **49**, 1617–1621.
- Di Liberto, G.M., O'Sullivan, J.A. & Lalor, E.C. (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.*, **25**, 2457–2465.
- Ding, N. & Simon, J.Z. (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.*, **107**, 78–89.
- Graves, A., Jaitly, N. & Mohamed, A.R. (2013). Hybrid speech recognition with deep bidirectional LSTM. *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, pp. 273–278.

- Haykin, S. & Liu, K.R. (2010) *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Hoboken, NJ.
- Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V. *et al.* (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Proc. Mag.*, **29**, 82–97.
- Horton, C., Srinivasan, R. & D'Zmura, M. (2014) Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'. *J. Neural Eng.*, **11**, 046015.
- Jasper, H.H. (1958) The ten twenty electrode system of the international federation. *Electroen. Clin. Neuro.*, **10**, 371–375.
- 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 J. Adv. Sig. Pr.*, **2009**, 6.
- Kerlin, J.R., Shahin, A.J. & Miller, L.M. (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J. Neurosci.*, **30**, 620–628.
- Mesgarani, N. & Chang, E.F. (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, **485**, 233–236.
- 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. *J. Neural Eng.*, **12**, 046007.
- Mirkovic, B., Bleichner, M.G., De Vos, M. & Debener, S. (2016) Target speaker detection with concealed EEG around the ear. *Front. Neurosci. Switz.*, **10**, 349.
- O'sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A. *et al.* (2014) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex*, **25**, 1697–1706.
- Polich, J. (1986) Attention, probability, and task demands as determinants of P300 latency from auditory stimuli. *Electroen. Clin. Neuro.*, **63**, 251–259.
- 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. *Eur. J. Neurosci.*, **35**, 1497–1503.
- Shinn-Cunningham, B.G. & Best, V. (2008) Selective attention in normal and impaired hearing. *Trends Amplif.*, **12**, 283–299.
- Spencer, K.M. & Polich, J. (1999) Poststimulus EEG spectral analysis and P300: attention, task, and probability. *Psychophysiology*, **36**, 220–232.
- Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.
- Sturm, I., Lapuschkin, S., Samek, W. & Müller, K.-R. (2016) Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Meth.*, **274**, 141–145.
- Wolpaw, J.R., Ramoser, H., McFarland, D.J. & Pfurtscheller, G. (1998) EEG-based communication: improved accuracy by response verification. *IEEE T. Rehabil. Eng.*, **6**, 326–333.