19th International Symposium on Hearing
19-24 June 2022, Lyon, France
DOI: 10.5281/zenodo.12623218

Robust decoding of the speech envelope from EEG recordings through deep neural networks

Michael Thornton[1], Danilo Mandic[1], and Tobias Reichenbach[2,*]

[1]Imperial College London, UK
[2]Friedrich-Alexander-University Erlangen-Nürnberg, Germany
[*]Corresponding author: tobias.j.reichenbach@fau.de

Abstract

During speech perception, a listener's brain activity tracks amplitude modulations in the speech signal, which are encoded in the speech envelope. The neural tracking of the speech envelope is modulated by cognitive factors such as attention to one of several competing speakers. Decoding the speech envelope from noninvasive electroencephalography (EEG) may be useful in future auditory prosthesis that could restore speech comprehension in noisy environments. Such applications require, however, a robust decoding of the speech envelope that functions in different acoustic conditions and that generalizes between different participants. Here we show that deep neural networks (DNNs) can lead to an enhanced decoding that has around 38% higher performance than the standard method, linear regression. The advantage of the DNNs persists across different acoustic scenarios and also when listener-independent decoders are used. We also show how an improved envelope decoding performance translates into a higher auditory attention decoding accuracy for the DNNs, in comparison to the method of linear modelling. Our work therefore demonstrates that DNNs show promise for data efficient, real-world auditory attention decoding.

1 Introduction

Conventional hearing aids are known to provide only a limited benefit to their users, especially when operating in noisy conditions (Lesica, 2018). The ability to determine the focus of a user's attention could enable the development of smart hearing aids with improved outcomes for those who suffer with hearing loss. Recent studies have demonstrated that auditory attention in multi-speaker scenarios can be decoded noninvasively from electrophysiological recordings such as the electroencephalogram (EEG) (Bleichner, Mirkovic, & Debener, 2016; Fiedler et al., 2017; Looney et al., 2010; Miran et al., 2018; O'Sullivan et al., 2014). One common paradigm for auditory attention decoding (AAD) is the method of backward linear modelling, whereby a set of coefficients are estimated in order to linearly reconstruct a speech feature from EEG recordings. Typically, the feature of choice is the speech envelope. The envelope of the attended speech stream is more strongly represented in a listener's EEG, and can be more accurately reconstructed from EEG recordings than the envelopes of the unattended speech streams. Therefore, in the backward modelling approach, a reconstruction score (typically Pearson's correlation coefficient between the reconstructed and the actual speech envelope) for each speech stream serves as a marker of selective attention.

Since the processing in the auditory system is inherently nonlinear, it is natural to ask whether nonlinear methods for auditory attention decoding can offer superior performance over linear methods. Nonlinear methods for backward modelling and AAD based on artificial neural networks have been introduced recently (Ciccarelli et al., 2019; de Taillez, Kollmeier, & Meyer, 2020). Artificial neural networks are heavily-parameterised, nonlinear models which are capable of representing a broad class of functions. In fact, they are universal function approximators (Mandic & Chambers, 2001). Deep neural networks (DNNs) are artificial neural networks which contain many layers of processing units (neurons). The correlation-based AAD technique described above can be also be used in conjunction with a DNN, by exchanging the linear backward model with a nonlinear DNN.

Deep neural networks are known to suffer from issues surrounding generalisability. This has been highlighted by some some recent investigations which did not achieve a competitive AAD performance across multiple datasets, when using DNNs which have elsewhere been reported to be effective (Ciccarelli et al., 2019; Geirnaert et al., 2021). In this work, we compared the performance of two nonlinear DNNs as well as one linear model for predicting the speech envelope from EEG recordings. We applied the DNNs to two distinct EEG datasets. Following a recent study, we examined a fully-connected feed-forward neural network (FCNN) (de Taillez, Kollmeier, & Meyer, 2020). We also considered a more lightweight convolutional neural network (CNN) based on the EEGNet architecture, which has been proposed for a range of brain-computer-interface applications (Lawhern et al., 2018).

2 Materials and Methods

2.1 Datasets and preprocessing

Two datasets from our research group were used in this work. The first dataset (termed Dataset 1 hereafter) was collected by Weissbart *et al.* (Weissbart, Kandylaki, & Reichenbach, 2020). A total of 13 native English-speaking participants were asked to attend to a single speaker narrating an audiobook in English, in noiseless and anechoic listening conditions. Each participant listened to 15 audiobook chapters in one recording session. The duration of each chapter was approximately 2.5 minutes, and each participant took breaks between chapters. Attendance to the audiobook was ensured by asking comprehension questions during the breaks.

The second dataset (termed Dataset 2 hereafter) was collected by Etard and Reichenbach (2019). A total of 18 native English-speaking participants attended to speakers narrating audiobook chapters in several listening conditions: clean speech, speech in noisy conditions, and speech in competing-speaker scenarios. Additionally, 12 of the participants listened to a speaker narrate an audiobook in a foreign language, Dutch. For the noisy speech, background babble noise was synthesised and combined with the speech at three different signal-to-noise ratios (SNRs) of -1.4 dB, -2 dB, and -3.2 dB. There were two competing-speakers scenarios: either the male speaker was ignored, and the female speaker attended, or vice-versa. For each listening condition, EEG was recorded in four trials of approximately 2.5 minutes in duration.

Preprocessing was performed using default routines available in MNE-Python version 0.24.1 (Gramfort et al., 2013). To obtain the speech envelopes, we computed the absolute value of the Hilbert transform of each speech stream. The speech envelopes were low-pass filtered below 50 Hz (linear phase type 1 FIR anti-aliasing filter, Hamming window, 12.5 Hz transition bandwidth, -6 dB attenuation at 56.25 Hz, -53 dB stopband attenuation) and resampled to 125 Hz. To preprocess the EEG recordings, all channels were low-pass filtered below one of several upper passband

edges (linear phase type 1 FIR anti-aliasing filters, Hamming windows, −53 dB stopband attenuation). The considered upper passband edges were: 8 Hz, 12 Hz, 16 Hz, and 32 Hz. The EEG recordings were subsequently resampled to 125 Hz and high-pass filtered above one of two lower passband edges in order to remove slow drifts (linear phase type 1 FIR filters, Hamming windows, −53 dB stopband attenuation): 0.5 Hz, or 2 Hz. Finally, for every trial, each EEG channel was standardised to have zero mean and unit variance.

2.2 Linear models and deep neural networks

A linear backward model can be specified in the time domain by a matrix of parameters $\theta_{i,j}$. These are convolved with the EEG recordings to produce an estimate of the speech envelope:

$$\hat{y}_t = \sum_{i=1}^{C} \sum_{j=0}^{L-1} x_{t-j,i} \theta_{i,j}. \tag{1}$$

In this expression, y_t denotes the speech envelope sampled at time t, $x_{t,i}$ designates the EEG sampled at time t from electrode i, C represents the number of EEG channels being considered, and L is the filter length which describes how many temporal EEG samples are employed to estimate the speech envelope. The parameters were fitted through ridge regression.

The fundamental unit of any neural network is the "neuron". A neuron receives a pre-determined number of inputs, which it linearly combines according to its set of parameters (or weights). A nonlinear activation function is then applied to the resulting scalar quantity.

A feed-forward neural network consists of layers of neurons, with neurons in a particular layer receiving as inputs the outputs of neurons in preceding layers (neurons within a single layer do not connect with one another). If each neuron in a particular layer is connected with each neuron in the preceding layer, the neural network is described as 'fully connected' (FC). The fully-connected feed-forward neural network (FCNN) used in this work is depicted in Figure 1. If each neuron in a particular layer is instead only connected to a neighbourhood of input neurons, and all neurons in the same layer share the same parameters (weight sharing), then the neural network is described as a convolutional neural network (CNN). The CNN that was used in this work is shown in Figure 1.

The FCNN used in this work was inspired by the architecture of de Taillez *et al.* (de Taillez, Kollmeier, & Meyer, 2020). A spatiotemporal segment of EEG recordings is passed through several fully-connected feed-forward layers, with each layer containing fewer neurons than the preceding layer. The activation function is the hyperbolic tangent. The number of inputs is equal to $C \times T$, with C as the number of EEG channels used, and T as the number of temporal samples in the segment. The scalar output represents a point estimate of the speech envelope at the onset of the segment. Following de Taillez *et al.*, the number of neurons in each hidden layer decreases linearly from $C \times T$ to 1. The number of hidden layers is a tunable hyperparameter.

Our choice of CNN was inspired by the EEGNet architecture of Lawhern *et al.* (Lawhern et al., 2018), which employs the exponential linear unit (ELU) as a nonlinear activation function, as well as batch normalisation and average pooling. Batch normalisation improves convergence during training by making the optimisation problem smoother (Ioffe & Szegedy, 2015; Santurkar et al., 2018). Average pooling is a form of downsampling. To regularise the CNN, we used L2 regularisation and a variant of dropout known as spatial dropout (Tompson et al., 2015). The scalar output of the CNN is formed by taking a linear combination of all of the activations in the final convolutional layer.

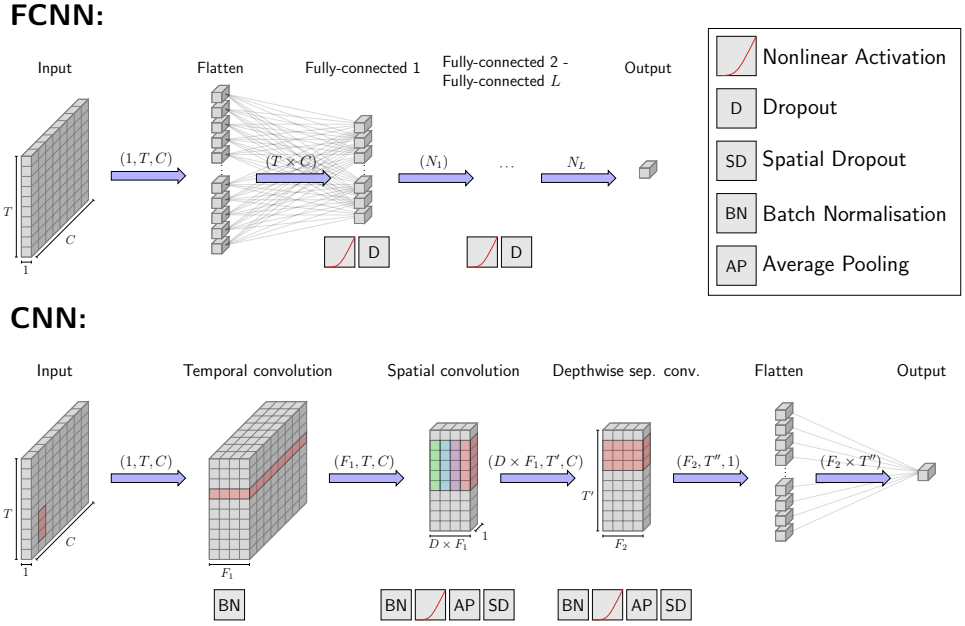


Figure 1 — The two neural network architectures used in this work. A spatiotemporal segment of T EEG samples and C channels is presented to both of the neural networks. Top: the FCNN architecture consists of several fully-connected hidden layers. The L^{th} hidden layer consists of N_{D_L} neurons. A nonlinear activation function is applied to the output of each hidden layer, followed by a dropout layer. Bottom: the CNN architecture employs convolutional layers, which make use of local connectivity and weight-sharing to reduce the number of parameters in the network. The 1^{st} convolutional layer consists of F_1 convolutional filters, each sharing common input but comprising distinct parameterisations. The 2^{nd} convolution flattens the channel dimension, and consists of $D \times F_1$ convolutional filters. The 3^{rd} convolutional layer implements the so-called depthwise separable convolution, which is similar to an ordinary convolutional layer consisting of F_2 low-rank convolutional filters. Following (Lawhern et al., 2018), we set $F_2 = D \times F_1$ to reduce the dimensionality of the hyperparameter search. Several operations may be applied to the activations of each convolutional layer, including a non-linear activation function, batch normalisation, and spatial dropout. Average pooling is a form of downsampling whereby the activations of a neighbourhood of neurons are replaced by the average activation. This form of downsampling is only applied along the temporal dimension.

2.3 Training procedure

The coefficients of the linear model were fitted through ridge regression, as discussed in Section 2.2. Following (de Taillez, Kollmeier, & Meyer, 2020), we optimised the DNN parameters by minimising the negative correlation coefficient between the reconstructed speech envelope and the target speech envelope. The NAdam optimiser was used (Dozat, 2016).

Dataset 1 consisted of 15 trials per participant, each of approximately 2.5 minutes in duration. We reserved 9 of these trials for model training, 3 for validation, and 3 for evaluation. Dataset 2 consisted for four trials per listening condition, per participant. Each trial had a duration of approximately 2.5 minutes. We used 8 trials for model training (4 clean-speech trials and 4 high-SNR speech-in-noise trials), and 4 trials for validation (from the low-SNR speech-in-noise condition). The remaining trials were used for evaluation.

During training, batches of EEG data were presented to the DNNs, and a corresponding batch of predicted speech envelope values was produced. These were correlated against the actual speech envelope values, and the DNN parameters were updated via a NAdam gradient descent step in order to maximise the correlation coefficient. After iterating through all batches of data (one epoch), the correlation score was evaluated on the validation dataset. The DNN hyperparameters were tuned via a random search.

For each analysis, we trained 15 linear models with different regularisation parameters spaced evenly on a logarithmic scale (ranging from 10^{-7} to 10^{7} inclusive). The model that achieved the highest correlation score on the validation dataset was selected for testing.

2.4 Analysis procedure

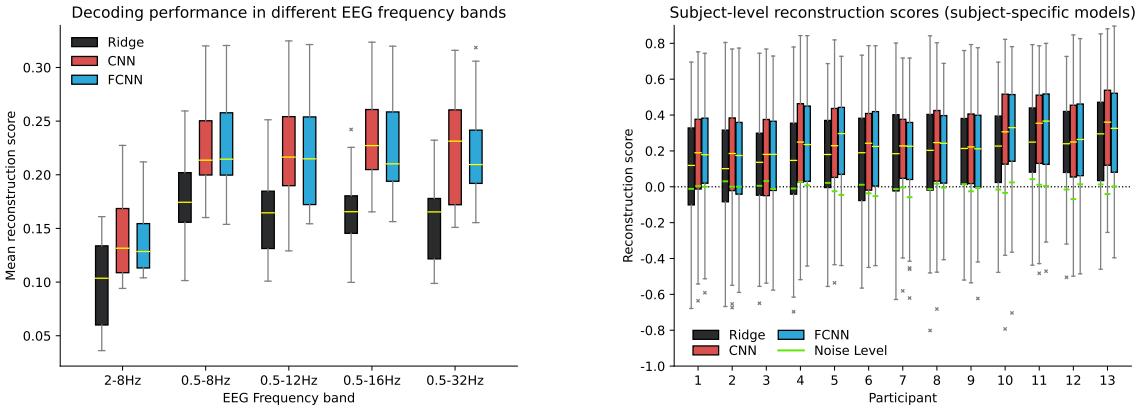
To evaluate the models, the EEG data were split into contiguous windows. The predicted speech envelope values in each widow were correlated against the actual speech envelope values. The mean and variance of the correlation score over all windows were calculated, since these quantities are of interest in AAD applications. To construct a null distribution, the predictions for each window were also correlated against the true speech envelope in unrelated windows.

In Section 3.4 we applied the linear model as well as both DNNs to EEG recorded in competing-speakers scenarios. The correlation-based method was used to decode auditory attention using each DNN or the linear model. The performance of each backward model was quantified through the attention decoding accuracy for a given window size W. The information transfer rate (bit rate) B was also calculated according to (McFarland, Sarnacki, & Wolpaw, 2003; Wolpaw et al., 1998):

$$WB = \log_2 N + P \log_2 P + (1 - P) \log_2 \frac{1 - P}{N - 1}, \tag{2}$$

where N is the number classes in the classification problem (two in this case), and P is the attention decoding accuracy. The bitrate in (2) is scaled by the window duration W to obtain an effective bitrate that takes into account the size of window.

The neural networks were implemented in PyTorch version 1.10.0 (Paszke et al., 2019). Statistical analyses were conducted using Scipy version 1.7.1 and Statsmodels version 0.11.1 (Seabold & Perktold, 2010; Virtanen et al., 2020).



(a) (b)

Figure 2 — Two distinct DNN architectures, as well as a linear model, were used to relate EEG recordings to the envelope of clean speech. The decoding performance for all three methods when different EEG frequency bands are used is shown in (a). Each boxplot comprises the mean envelope reconstruction score (correlation coefficient) for each participant. The subject-level results for the EEG frequency band 0.5-8 Hz are depicted in (b). Each boxplot represents the median and range of the reconstruction score when a 2 s correlation window is employed. The median reconstruction scores of the null distributions are shown in green.

3 Results

3.1 Subject-specific models

We first evaluated the performance of the DNNs and linear models at reconstructing the envelope of clean speech, using Dataset 1. We considered several different EEG frequency bands, and we found that using the 0.5-8 Hz band yielded the most effective linear decoders (Figure 2a). For this frequency band, the spread of reconstruction scores for all participants is reported in Figure 2b. We used this frequency band for all subsequent analyses.

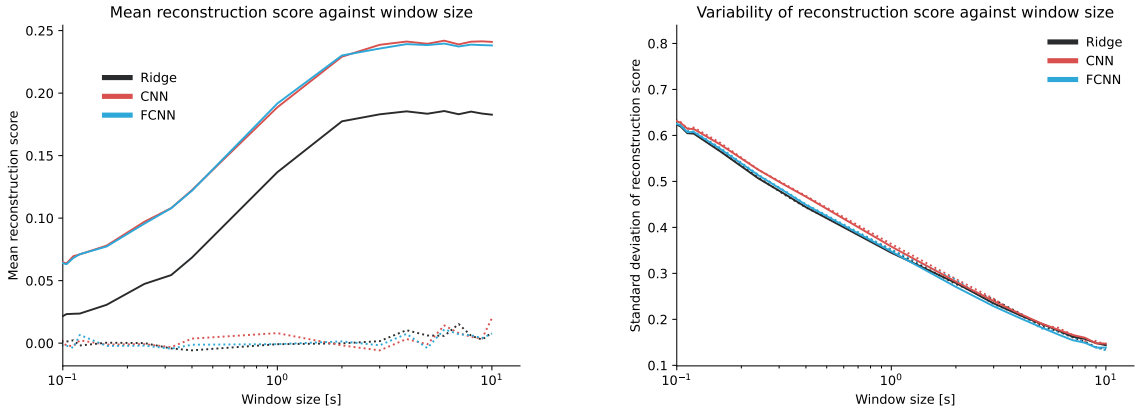
The median values of the null scores are shown in green on Figure 2b. We tested the reconstruction scores for significance with a t-test (single-tailed unpaired t-test, FDR-corrected). Additionally, we compared the reconstruction scores of each pair of models for every participant (unpaired t-tests, FDR corrected). The corrected p-values are reported in Table 1.

To analyze the performance on the population level, we calculated the mean reconstruction score for every participant and model. We then compared the 13 mean reconstruction scores achieved by each model. We found no significant difference between the two DNNs ($p = 0.83$, two-tailed paired t-test). However, both DNNs significantly outperformed the linear model (CNN: $p = 2.2 \times 10^{-5}$; FCNN: $p = 1.6 \times 10^{-4}$; single-tailed t-tests, Bonferroni corrected.)

The mean and standard deviation of the reconstruction score varied with window duration. We determined the dependence of the mean and standard deviation of the reconstruction scores on the duration by performing the analysis procedure with windows of various sizes (ranging from 0.1 s and 10 s). The mean and standard deviation of the reconstruction scores were averaged over all participants (Figure 3). The mean reconstruction scores of both linear and nonlinear models are strongly degraded for window sizes less than 2 s. For window sizes greater than 2 s,

Table 1 — Statistical tests on the data shown in Figure 2b. The first three rows show the p-values obtained when testing for difference between the reconstruction scores (ρ) and null distributions (ρ_0) for each method. These were FDR-corrected independently of the p-values in the next three rows. The next three rows show the p-values obtained by testing for differences between the reconstruction scores of the different techniques.

Alternative Hypothesis	Participant												
	1	2	3	4	5	6	7	8	9	10	11	12	13
$\rho^{\text{ridge}} > \rho_0^{\text{ridge}}$	**2.94e-05**	**2.77e-04**	**1.02e-03**	**1.48e-09**	**7.86e-13**	**1.23e-06**	**1.89e-08**	**1.61e-09**	**2.05e-13**	**1.10e-13**	**2.01e-19**	**2.57e-19**	**3.79e-19**
$\rho^{\text{CNN}} > \rho_0^{\text{CNN}}$	**1.08e-09**	**1.48e-09**	**3.04e-09**	**8.09e-19**	**2.08e-13**	**1.50e-12**	**2.30e-12**	**2.19e-16**	**2.64e-12**	**3.73e-21**	**7.60e-27**	**7.56e-21**	**1.69e-31**
$\rho^{\text{FCNN}} > \rho_0^{\text{FCNN}}$	**2.09e-14**	**3.90e-11**	**7.92e-10**	**1.04e-15**	**1.71e-18**	**3.55e-11**	**1.03e-17**	**3.18e-16**	**4.31e-13**	**3.04e-28**	**2.97e-26**	**1.66e-22**	**1.35e-20**
$\rho^{\text{CNN}} \neq \rho^{\text{FCNN}}$	5.12e-01	5.12e-01	9.11e-01	9.97e-01	3.91e-01	9.97e-01	9.97e-01	9.97e-01	7.90e-01	9.11e-01	7.90e-01	9.11e-01	5.12e-01
$\rho^{\text{CNN}} > \rho^{\text{ridge}}$	**3.18e-02**	**2.09e-02**	1.63e-01	**1.82e-02**	**3.67e-02**	1.99e-01	2.97e-01	2.84e-01	3.67e-01	**1.71e-03**	**2.34e-02**	5.12e-01	**3.67e-02**
$\rho^{\text{FCNN}} > \rho^{\text{ridge}}$	**2.43e-03**	1.00e-01	1.00e-01	**1.82e-02**	**2.43e-03**	1.96e-01	2.97e-01	2.62e-01	5.12e-01	**7.66e-04**	**7.86e-03**	6.11e-01	1.99e-01



(a) (b)

Figure 3 — The mean (a) and standard deviation (b) of the reconstruction score (Pearson's correlation coefficient), averaged over all participants from Dataset 1, is plotted against the size of the correlation window used during evaluation. For window sizes less than 2 s in duration, the mean reconstruction score was considerably degraded. The variability of the reconstruction score increases sharply with decreasing window size, but is similar for all three methods. The dotted lines show the mean and standard deviation of the null reconstruction scores for the three methods.

the mean reconstruction score for each DNN was around 38% above that of the linear model. The mean of the set of 13 standard deviations was very similar for all three methods across all window sizes.

3.2 Subject-independent models

To test whether the models generalise between participants, we left one participant's data out of the training procedure, and instead trained each of the models on the data from the 12 remaining participants in Dataset 1. We repeated this process 13 times, leaving out a different participant each time. In this way, we pre-trained 13 subject-independent models and applied them to data from the unseen participant. For comparison, we also trained population models using training data from all of the participants, and applied these to distinct test data (recorded from the same 13 participants). Our results are summarised in Figure 4.

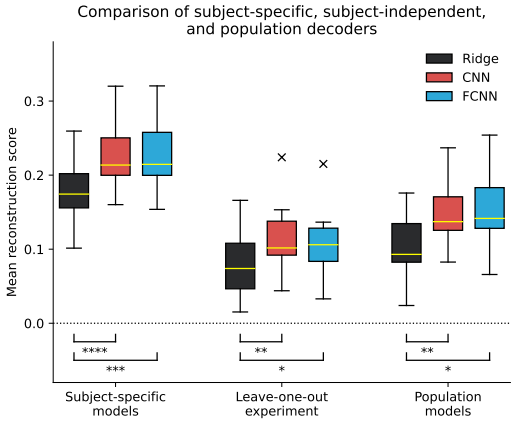


Figure 4 — Comparison between subject-specific and subject-independent decoders applied to clean speech (Dataset 1). The boxplots represent the spread of the mean reconstruction scores achieved for each participant. The first group (subject-specific models) shows the mean reconstruction scores achieved by subject-specific decoders applied to Dataset 1. The second group (leave-one-subject-out experiment) shows the mean reconstruction scores achieved by the subject-independent decoders described in Section 3.2. For the third group (population models), population models were trained with Dataset 1 and subsequently applied them to data recorded from individual subjects in Dataset 2.

On the subject level, the use of the linear subject-independent models resulted in significant mean reconstruction scores for 9 of the 13 participants (single-tailed unpaired t-test, FDR corrected). Both the subject-independent CNN and FCNN yielded significant reconstruction scores for 12 participants. For each participant, we compared the use of each pair of subject-independent models using unpaired t-tests (FDR corrected). We found that no subject-independent model significantly outperformed either of the other subject-independent models for any of the participants. On the population level, however, both subject-independent DNNs significantly outperformed the subject-independent linear models (CNN: $p = 3.1 \times 10^{-4}$; FCNN: $p = 0.01$; single-tailed t-tests, Bonferroni corrected). There was no significant difference between the subject-independent DNNs on the population level.

The subject-independent decoders yielded scores which were approximately 50% below those of the subject-specific decoders. The population decoders performed better than the subject-independent decoders, but worse than the subject-specific decoders.

3.3 Performance of subject-specific models in different listening conditions

For real-world applications, a decoder needs to perform well across a range of listening conditions. We therefore trained subject-specific decoders using Dataset 2, which consisted of EEG recorded under a number of different listening conditions. The spread of mean reconstruction scores in each test condition are shown in Figure 5. We compared the sets of mean reconstruction scores achieved by each pair of models within each listening condition using two-tailed paired t-tests (FDR corrected). There was no significant difference between the DNNs in any listening condition. However, both DNNs significantly outperformed the linear model at reconstructing the attended speech stream in the competing-speakers conditions, as well as in the background babble noise condition. The DNNs performed similarly to the linear models at reconstructing the envelope of clean speech in a foreign language, as well as at reconstructing the unattended speech envelope in the competing-speakers conditions.

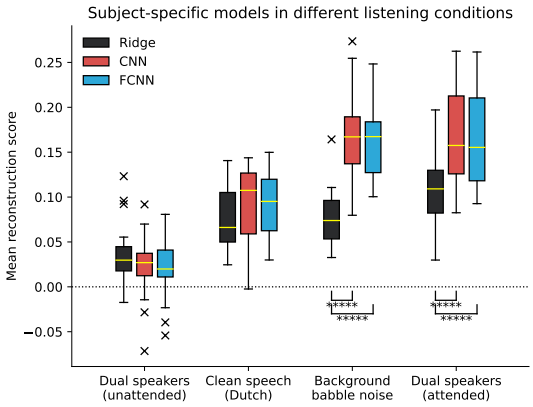


Figure 5 — The spread of mean reconstruction scores for the participants in Dataset 2 when subject-specific decoders were applied in different listening conditions.

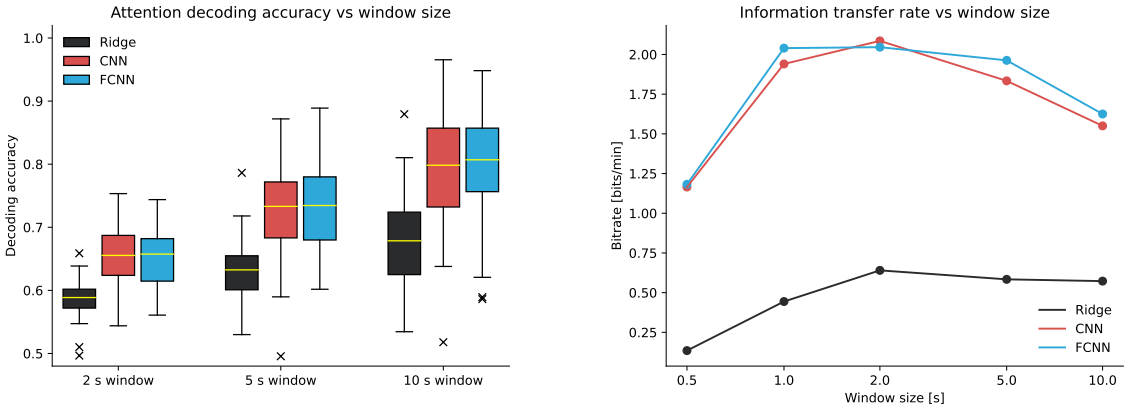
3.4 Attention decoding performance

For our final case study, we investigated whether the subject-specific decoders described in Section 3.3 could actually be used for auditory attention decoding in the competing-speaker scenarios. We compared the reconstruction score (correlation coefficient) for the attended and unattended speakers in each window, and calculated the proportion of windows for which the reconstruction of the attended envelope was greater than that of the unattended envelope (the binary classification accuracy). The decoding accuracies for three different window durations (2 s, 5 s, 10 s) are shown in Figure 6. Both DNNs offered clear accuracy improvements over the linear model across all three window durations.

Following (de Taillez, Kollmeier, & Meyer, 2020), we also calculated effective bitrates for attention decoding using different window sizes. The bitrate is related to the time-rate of correct classifications, and was calculated according to Equation 2. We found that a window size of 2 s maximises the decoding performance of the CNN as well as that of the linear model (Figure 6b). A window size of 1 s was marginally more suitable for the FCNN. Both DNNs achieved much higher bitrates than the linear model.

4 Discussion

We have investigated the performance of two types of DNNs at estimating the speech envelope from EEG recordings. The performance of each DNN was compared to that of a standard linear model. A comprehensive evaluation has shown that the two DNNs can achieve very similar performances when reconstructing the envelope of clean speech, whilst exceeding that of the linear model by about 38%. The advantage of using the DNNs over the linear models remained even when the models were applied to subjects whose EEG data had not been seen during training. Importantly, the DNN architectures and hyperparameters have been shown to generalise to a distinct data set, and subject-specific models have been applied effectively to data in which speech was presented in different types of noise. Our results have demonstrated that DNNs can robustly enhance the decoding of speech features from EEG recordings.



(a) (b)

Figure 6 — Comparison of the attention decoding accuracies of the three methods, across three different window sizes (a). The information transfer rate for the three methods, which quantifies the tradeoff between decoding accuracy and window duration (b).

4.1 Similarities between the two DNNs

The scores achieved by the two DNNs investigated in this work were remarkably close. In fact, the outputs of the two neural networks were themselves highly correlated, which suggests that the two DNNs had learned to represent very similar functions. Owing to local neuron connectivity, the CNN required far fewer parameters than the FCNN: about four thousand versus twelve million, respectively. The linear model required about three thousand parameters, which is comparable to the number of parameters in the CNN. The CNN may therefore be a preferable, lightweight alternative to the FCNN for practical applications. Future investigation may reveal effective architectures which are even more lightweight, for example by removing or 'pruning' neurons which are of low importance (Frankle & Carbin, 2019; Zhu & Gupta, 2017). Additionally, prior information about this signal processing problem could be exploited by imposing inductive biases on a DNN (Bronstein et al., 2021). For example, the neural response to the speech envelope is well characterised in the literature, and the known spatial arrangement of the EEG sensors may be leveraged.

4.2 Subject-specific decoders

The effect of broadening the EEG frequency band from 0.5-8 Hz to 0.5-32 Hz had no discernible impact on the performance of the DNNs. De Taillez *et al.* found that the use of broadband EEG improved the attention decoding accuracy of their DNN when applied in a competing-speaker scenario (de Taillez, Kollmeier, & Meyer, 2020). Whilst it could be true that more information is encoded by higher frequency oscillations during selective attention in competing-speaker scenarios, it could also be argued that the noise introduced through the higher frequency bands assisted the optimisation procedure during DNN training in that study (via noise injection).

We used the 0.5-8 Hz EEG frequency band for subsequent analyses. Both of the DNNs as well as the linear modelling method achieved significant reconstruction scores for all participants. On the population level, the improvement offered by the DNNs was statistically significant. The

overall performance of the DNNs was around 38% greater than that of the linear model. Even on the subject level, the CNN (FCNN) offered a statistically significant performance increase compared against the linear model for 7 (5) participants.

We found that for windows smaller than around 2 s in duration, the reconstruction accuracy of all three methods was severely degraded. The latency of a real-world decoder which is based on the correlation method may therefore be limited to this timescale, unless techniques such as state-space models are employed (Miran et al., 2018). Indeed, we found that a window size of about 2 s maximises the information transfer rate of the correlation-based auditory attention decoding algorithm. The variability in the reconstruction score followed similar power-law dependencies on window size for all three methods. This finding contrasts with a previous study which found that the reconstruction score of a DNN similar to the FCNN used in this work was much more variable than that of a linear model, when applied in a competing-speaker scenario (Aroudi, de Taillez, & Doclo, 2020).

4.3 Subject-independent decoders

All three subject-independent decoders yielded reconstruction scores that were significantly different from the null distribution for the majority of the participants (9 for the linear model; 11 for both the CNN and the FCNN). We found that all three subject-independent decoders performed very similarly on the subject level. However, on the population level, the subject-independent DNNs both significantly outperformed the subject-independent linear model.

The subject-independent decoders performed significantly worse than their subject-specific counterparts (the performance decrease was around 50% for all three methods). Whilst a performance penalty is to be expected when subject-independent information is unavailable, a penalty of this magnitude may imply that some subject-specific information is required for real-world applications.

4.4 Application to EEG recorded under different listening conditions

We trained linear and nonlinear subject-specific decoders using a mixture of clean speech and speech-in-babble-noise conditions. We found that the DNNs outperformed the linear model by a considerable margin when reconstructing the envelope of an attended speaker in competing-speaker scenarios, as well as in background babble noise. All three methods performed very similarly at the task of reconstructing the unattended speaker in the competing-speaker scenarios.

The three methods also performed very similarly at reconstructing the envelope of clean speech in foreign Dutch. The comprehension score in this listening condition was 0%, and it is has been shown that cortical speech tracking in the delta band is modulated by the speech comprehension level (Etard & Reichenbach, 2019). Since very low comprehension levels were not represented in the training data, this may explain why the DNNs did not perform as well in this listening condition.

4.5 Attention decoding performance

Finally, we decoded auditory attention in competing-speaker scenarios using the subject-specific decoders that were trained with Dataset 2. It was found that the use of DNNs was advantageous for this purpose, as was shown in (de Taillez, Kollmeier, & Meyer, 2020). We also replicated the

finding that a short window length of about 2 s was optimal for real-time applications. However, the bitrates that were achieved by the DNNs were somewhat lower than those reported in (de Taillez, Kollmeier, & Meyer, 2020): this might be explained by the fact the authors trained their DNN using EEG recorded in a competing-speaker scenario, which was the same listening condition as was used for evaluation. Nevertheless, our study provides conclusive evidence that DNNs can be used for enhanced and robust decoding of selective attention in competing-speaker scenarios.

Acknowledgements

Michael Thornton was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1)

Code availability

Supporting Python code is available at https://github.com/Mike-boop/mldecoders. This package contains all the functions used for data preprocessing, model training, and analysis.

References

Aroudi, A., de Taillez, T., & Doclo, S. (2020). Improving auditory attention decoding performance of linear and non-linear methods using state-space model. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8703–8707. https://doi.org/10.1109/ICASSP40776.2020.9053149

Bleichner, M. G., Mirkovic, B., & Debener, S. (2016). Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. *Journal of Neural Engineering*, *13*(6), 066004. https://doi.org/10.1088/1741-2560/13/6/066004

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. https://doi.org/10.48550/arXiv.2104.13478

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., Mesgarani, N., Quatieri, T. F., & Smalt, C. J. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Scientific Reports*, *9*(1), 11538. https://doi.org/10.1038/s41598-019-47795-0

de Taillez, T., Kollmeier, B., & Meyer, B. T. (2020). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*, *51*(5), 1234–1241. https://doi.org/https://doi.org/10.1111/ejn.13790

Dozat, T. (2016). Incorporating Nesterov momentum into Adam. *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*.

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *The Journal of Neuroscience*, *39*(29), 5750–5759. https://doi.org/10.1523/jneurosci.1828-18.2019

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., & Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering*, *14*(3), 036020. https://doi.org/10.1088/1741-2552/aa66dd

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*. https://doi.org/10.48550/arXiv.1803.03635

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigne, A., Lalor, E., Meyer, B. T., Miran, S., Francart, T., & Bertrand, A. (2021). Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, *38*(4), 89–102. https://doi.org/10.1109/MSP.2021.3075932

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(267), 1–13. https://doi.org/10.3389/fnins.2013.00267

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, *37*, 448–456. https://doi.org/10.48550/arXiv.1502.03167

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Lesica, N. A. (2018). Why do hearing aids fail to restore normal auditory perception? *Trends in Neurosciences*, *41*(4), 174–185. https://doi.org/https://doi.org/10.1016/j.tins.2018.01.008

Looney, D., Park, C., Xia, Y., Kidmose, P., Ungstrup, M., & Mandic, D. P. (2010). Towards estimating selective auditory attention from EEG using a novel time-frequency-synchronisation framework. *Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN)*, 1–5. https://doi.org/10.1109/IJCNN.2010.5596618

Mandic, D. P., & Chambers, J. A. (2001). *Recurrent neural networks for prediction*. John Wiley & Sons, Ltd. https://doi.org/10.1002/047084535x

McFarland, D. J., Sarnacki, W. A., & Wolpaw, J. R. (2003). Brain–computer interface (BCI) operation: Optimizing information transfer rates. *Biological Psychology*, *63*(3), 237–251. https://doi.org/10.1016/s0301-0511(03)00073-5

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., & Babadi, B. (2018). Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach. *Frontiers in Neuroscience*, *12*, 262. https://doi.org/10.3389/fnins.2018.00262

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1912.01703

Santurkar, S., Tsipras, D., Ilyas, A., & Mądry, A. (2018). How does batch normalization help optimization? *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2488–2498. https://doi.org/10.48550/arXiv.1805.11604

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, 92–96.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient object localization using convolutional networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 648–656. https://doi.org/10.1109/CVPR.2015.7298664

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. https://doi.org/10.1162/jocn_a_01467

Wolpaw, J., Ramoser, H., McFarland, D., & Pfurtscheller, G. (1998). EEG-based communication: Improved accuracy by response verification. *IEEE Transactions on Rehabilitation Engineering*, *6*(3), 326–333. https://doi.org/10.1109/86.712231

Zhu, M., & Gupta, S. (2017). To prune, or not to prune: Exploring the efficacy of pruning for model compression. *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*. https://doi.org/10.48550/arXiv.1710.01878

Comments

Comment from Sarah Knight: This is by no means my specialist subject (I'm a behavioural psychologist), so I'm afraid I can't comment on the technical aspects of the paper. However, I have a couple of (probably very naïve) questions!

First, why did the DNNs outperform the linear model by such a margin when it came to reconstructing the attended speaker in the speech-in-noise (competing talker + babble) conditions? Is this related to the fact that the EEG signal captured the process of selective attention?

Second, does the lower limit of a 2s window for reasonable reconstruction accuracy severely limit the likely usefulness of these techniques for real-world applications?

Thanks for the opportunity to read this work!

> Thanks for your comment!
> I think the DNNs worked so well in the attended speaker + babble condition because this was represented in the training data (the training data consisted of clean speech, and two speech-in-babble-noise conditions - low SNR and high SNR. the network was evaluated on a medium SNR condition).
> The other conditions were not represented in the training data. I think the DNNs worked so well in the attended + distracting speaker conditions because a) perhaps the EEG response to the attended speaker is not so different in the competing-speakers scenarios and the speech-in-babble noise scenarios and b) it is possible that the speech-in-babble noise conditions also elicit similar attentional effects in the EEG response to the attended speech stream in the two-speakers condition.
> As you suggest, shorter windows would improve the responsiveness of a decoder to switches in auditory attention (e.g. from one speaker to another). So I would say shorter windows are better, with two caveats (I will try to find references when I find some time):
>
> 1. Quickly-adapting hearing aids don't appear to offer much benefit to users with lower cognitive function - and if I remember correctly, adapting too quickly may even be detrimental for the user.
> 2. Some work on closed-loop hearing devices of the type suggested in the paper has already been carried out in the literature, and the main complaint from users was actually to do with the relatively low AAD accuracy - it appears that rather higher accuracies will be required if this is going to work well for users. So in some sense the latency of the decoder may actually be a secondary consideration.