Forward Nonlinear Model for Deep Learning of EEG Auditory Attention Detection in Cocktail Party Problem

FORWARD NONLINEAR MODEL FOR DEEP LEARNING OF EEG AUDITORY ATTENTION DETECTION IN COCKTAIL PARTY PROBLEM

Laura Falaschetti, Michele Alessandrini, Claudio Turchetti
DII - Department of Information Engineering
Università Politecnica delle Marche
via Brecce Bianche,12, 60131 Ancona, Italy
l.falaschetti@staff.univpm.it
m.alessandrini@staff.univpm.it
c.turchetti@staff.univpm.it

ABSTRACT

In a multi-speaker scenario, humans are able to focus on a target speaker, ignoring all other speakers and noise, thus solving the so-called *cocktail-party problem*. However, elderly people and people suffering for hearing loss struggle to listening under these conditions. Recent studies have confirmed that the listener's selective attention to the attended speaker can be decoded using recording of brain activity such as *electroencephalography*, thus opening new opportunities in developing a new generation of neuro-steered hearing aids and hearing prostheses. To this end several algorithms have been developed for solving the so called *auditory attention decoding* problem from electroencephalography on the basis of *neural entrainment* mechanism. The most common approaches in development of auditory attention decoding algorithms are based on linear modeling of the neural entrainment. However, even though these algorithms have shown to be effective in solving cocktail-party problem, they have some inherent limitations. The main objective of this contribution is to show that nonlinear modeling of speech-electroencephalography system ensures the best performance in terms of higher correlation between stimulus and neural response, thus proving the limitations of linear approach. For this purpose the most common linear models for auditory attention decoding are reviewed and a new nonlinear model for auditory attention decoding is proposed. An extensive experimentation using a specific speech-electroencephalography dataset, confirms the superiority of nonlinear modeling in solving the auditory attention decoding problem.

***K*eywords** Cocktail Party Problem, Auditory Attention Detection, Speech entrainment, CCA, EEG, Deep Learning

List of abbreviations cocktail-party problem(CCP), electroencephalography (EEG), auditory attention decoding (AAD), canonical correlation analysis (CCA), kernel canonical correlation analysis (KCCA), deep canonical correlation analysis (DCCA), magnetoencephalography (MEG), electrocorticography (ECoG), magnetoencephalography (MEG).

1 Introduction

The *cocktail party problem* (CPP) is a psychoacoustic phenomenon that refers to the human ability to selectively attend to and recognize one of more competing streams of speech [1] [2]. However, many people with hearing problems struggle to listen under these noisy conditions [3], [4]. Numerous efforts have been dedicated to the CPP in diverse fields, and even though a complete understanding of the cocktail phenomenon is still missing [5], previous studies have provided evidence that this ability of auditory system is closely related to the auditory attention during the process of speech perception in the human brain [6], [7]. Recently, research in neuroscience has showed that the attention to target speech can be decoded from the cerebral cortical responses [8], [9], [10], [11], [12], [13], [14], [15], in particular using recording of brain activity such as *electroencephalography* (EEG) [16]. [17], [18], [19], [20], [21], and in order to guarantee speech intelligibility in a multi-speaker environment hearing devices have been developed by using noise suppression systems [22], [23], [24].
However to be effective it is essential to inform these systems about which speaker to enhance and which other speaker to treat as background noise and to suppress.
These new insights open up opportunities to develop a new generation of neuro-steered hearing aids and hearing prosteshes [25], [26], using *auditory attention decoding* (AAD) techniques that extract the attention-related information directly from the brain. For this purpose several AAD neurorecording modalities can be used to perform AAD, such as EEG [27], [28], [29], [30], [31], *electrocorticography* (ECoG) [5], and *magnetoencephalography* (MEG) [32], [33], nevertheless EEG is the most suitable approach due to high cost of ECoG and the lack of wearability of MEG. Conversely, EEG can be easily integrated with hearing devices, as it is noninvasive, wearable, and relatively cheap technique, [34] [35], [36], [37], [38], [39], [40], [41].
The algorithm used for deciphering human auditory attention from EEG signal are based on the mechanism of *neural entrainment* (or speech entrainment), that consists of some features of the audio heard by a listener, such as the envelope, that are tracked by the brain [12], [42] [43], [44], [45], [46], [47]. The two most common linear approaches in development AAD algorithms based on speech entrainment are i) *encoding* , (or forward modeling), *i.e.* estimating the EEG response from the speech stimulus, and *decoding*, (or backward modeling), *i.e.* estimating the speech stimulus from the EEG response [27], [26], [48], [49], [29]. A third approach combines forward and backward modeling using the *canonical correlation analysis* (CCA), [50], [51], [52], [53], [54].
Even though these algorithms have shown to be effective in solving CCP, linear modeling approach for the study of auditory attention decoding has some inherent limitations, the most severe of which are the followings.
1) It is generally considered that human speech perception corresponds to nonlinear information integration in the brain [55].
2) There is a very low correlation coefficient between the reconstructed speech envelope [4], [27], which clearly shows that linear modeling is not suitable to relate EEG response to speech stimulus.
3) From a mathematical point of view, both forward and backward models, may not be invertible causing an ill-posed problem. Thus, even though regularization techniques can be adopted, the model parameter selection is an empirical process [45].

4) Decoding accuracy depends on the EEG signal duration, *i.e.* higher accuracy is achieved with longer duration, thus establishing a trade-off between accuracy and duration [36].
Due to the above limitations of the linear modeling several nonlinear approaches have recently proposed. *Kernel canonical correlation analysis* (KCCA) [56], [51], [57], [58], [59], is an extension of CCA that finds pairs of nonlinear projections of the two views by maximizing their correlation and aims to overcome the limitations of standard CCA due to the nonlinear relationship between the EEG and audio features. *Deep canonical correlation analysis* (DCCA) is a method to learn complex nonlinear transformation of two views in which parameters of both transformations are jointly learned to maximize the total correlation [60] [61], [62], that can be viewed as an alternative to the nonparametric method of KCCA. More recently, thanks to the widespread diffusion of neural networks, a large amount of solutions based on this paradigm has been suggested, to improve the AAD performance [63], [60] [64], [65], [34], [66], [67], [68], [64].
The contribution of this chapter is twofold, i) firstly a comprehensive review of the principal techniques, both linear and nonlinear, used for AAD is presented; ii) secondly a technique based on a novel forward nonlinear modeling approach for AAD is suggested. As a main result of this contribution, experimental results show that nonlinear modeling ensures the best performance, thus proving the potential limitations of the linear modeling approaches for the study of AAD. The remainder of this chapter is organized as follows. In Section 2 the problem of AAD is formally stated. Section 3 reviews three different linear AAD algorithms, based on backward, forward and mixed backward-forward models. Section 4 dicusses the proposed forward nonlinear (NFL) model for AAD. Section 5 reports some experimental results achieved with the proposed model. Finally, Section 6 concludes this chapter.

2 Auditory attention detection on cocktail party problem

There are many studies showing that the temporal envelope of the audio heard by a listener is tracked by the brain [12]. This mechanism called *neural entrainment* to continuous speech has been demonstrated using magnetoencephalography (MEG) [69], [70], electroencephalography (EEG) [71] and electrocorticography (ECoG) [72]. With reference to the problem of attentional selection in a cocktail party environment, early research in this area [7], [4], [5] has shown the brain activity to be sensitive to auditory attention, by assuming a mapping from the amplitude envelope of speech to EEG. Several recent studies [27], [4], [73], [13] have used recorded EEG data to estimate the input stimulus using a mapping from the neural data back to the stimulus, $i.e.$ in the reverse direction. With reference to this stimulus reconstruction method, it has been demonstrated that the cortical tracking of the attended sound is increased compared to the tracking of the unattended sounds.Following the above considerations, the problem of auditory attention detection can be formally stated as follows. With reference to Fig. 1, there are two competing speakers whose speech signals are $u_1(t)$ and $u_2(t)$, and a listener whose auditory attention is focused on one of the two speakers, called the attended speaker, corresponding to the unknown speech signal $u_a(t) \in \{u_1(t), u_2(t)\}$. For simplicity we only refer to two speakers alone, even though all the results presented here can be extended, without loss of generality, to more speakers. By denoting with $z(t, c)$ the EEG-signal at channel c, we assume $z(t, c) = f(u_a(t), t)$, i.e. the brain signal is only dependent on the attended speech signal $u_a(t)$. Then the problem is to detect the unknown signal $u_a(t)$, given the EEG-signal $z(t, c)$.

A crucial aspect in facing this problem is to derive a proper model of neural entrainment. System identification [74] is a well known technique that mathematically models a function that describes

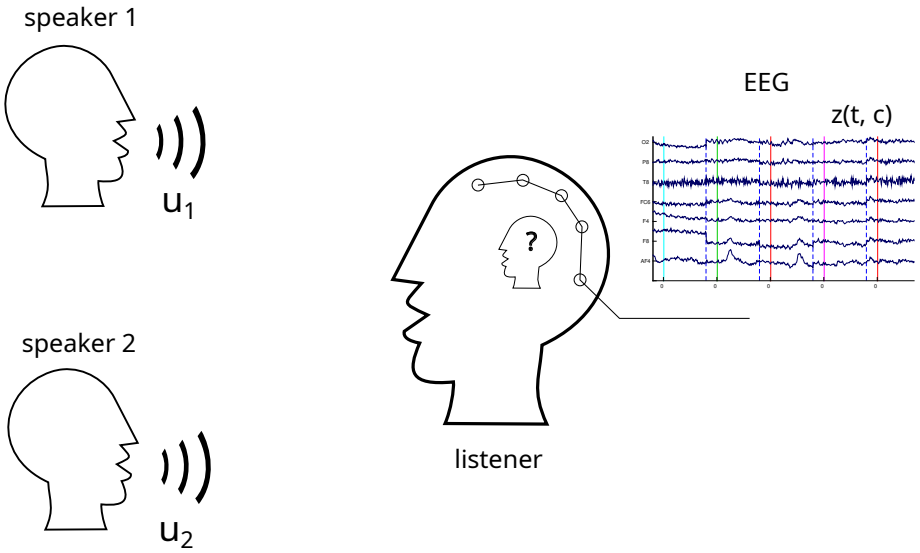

Figure 1: The cocktail party problem: given EEG segments $z(t, c)$ and two stimuli, classify which one of the input stimuli corresponds to the EEG.

the way a particular property of the stimulus is mapped onto a neural response. This approach is known as *forward modeling* and it has been used to study how speech is encoded in human brain activity [69], [75], [71] and to model response functions describing the linear mapping between properties of natural speech (such as the envelope or the spectrogram) and neural response [76], [32]. A complimentary way to investigate how stimulus features are encoded in neural response is modeling this mapping in the reverse direction, giving rise to the so called *backward modeling* [27], [73], [13].

3 Linear models for AAD

Following previously discussed classification modeling, the aim of this section is to review three different linear algorithms based on *backward* (or decoding), *forward* (or encoding) and mixed *backward-forward* (or decoding-encoding) models [44], [45], [46]. Decoding approach attempts to extract the sound from the neural response, while encoding approach attempts to predict neural responses (EEG) given the sound stimulus. Mixed decoding-encoding approach combines both decoding and encoding modeling.

Backward model In an EEG sensory system where the output is monitored by C recording channels, $z(t, c), c = 1 : C$, we assume the speech stimulus $u(t)$ can be reconstructed as [45]

$$u(t) = \sum_{c=1}^{C} \sum_{\tau=0}^{L-1} z(t + \tau, c) d(\tau, c), \tag{1}$$

where the spatio-temporal filter or decoder $d(\tau, c)$ represents the linear mapping from the neural response $z(t, c)$, back to the stimulus $u(t)$. Since the decoder maps backwards in time, this filter is an anti-causal filter that acts on L post-stimulus time laps. For a generic channel c and the time lag index τ ranging from zero to $L - 1$, we define the vector

$$\mathbf{e}_c(t) = [z(t,c), z(t+1,c), \ldots, z(t+L-1,c)]^T \in R^{L\times 1}, \;\; c = 1:C \tag{2}$$

and, collecting all the decoder coefficients for all time lags, the vector

$$\mathbf{d}_c = [d(0,c), d(1,c), \ldots, d(L-1,c)]^T \in R^{L\times 1}, \;\; c = 1:C \tag{3}$$

Thus (1) can be rewritten in a compact form as

$$u(t) = \mathbf{d}^T \mathbf{e}(t) \tag{4}$$

where $\mathbf{d} = [{\mathbf{d}_1}^T, \ldots, {\mathbf{d}_C}^T]^T \in R^{LC\times 1}$, and $\mathbf{e}(t) = [\mathbf{e}_1(t)^T, \ldots \mathbf{e}_C(t)^T]^T \in R^{LC\times 1}$.

Eq. (4) represents the decoder model in vector form and it is used to decode the attended speech from the knowledge of the two competing speech signals and the neural response.

Forward model Forward model is also referred to as temporal response function (TRF) or encoding model since it describes how the brain system encodes information. In this model it is commonly assumed [45] that the neural response $z(t,c)$ sampled at times $t = 1:T$ and at channel $c = 1:C$, is the convolution of the speech stimulus $u(t)$, with an unknown TRF $h(t,c)$

$$z(t,c) = \sum_{\tau=0}^{L-1} h(\tau,c)u(t-\tau) + n(t,c), \tag{5}$$

which can be rewritten as

$$z(t,c) = \mathbf{h}_c^T \mathbf{u}(t) + n(t,c), \;\; c = 1:C \tag{6}$$

where $\mathbf{h}_c = [h(0,c), h(1,c), \ldots, h(L-1,c)]^T \in R^{L\times 1}$, $\mathbf{u}(t) = [u(t), u(t-1), \ldots, u(t-L+1)]^T \in R^{L\times 1}$ and $n(t,c)$ is the residual response at each channel not dependent on the speech stimulus. By defining the matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_C]^T \in R^{C\times L}$, $\mathbf{z}(t) = [z(t,1), \ldots, z(t,C)]^T \in R^{C\times 1}$, and $\mathbf{n}(t) = [n(t,1), \ldots, n(t,C)]^T \in R^{C\times 1}$, (6) becomes

$$\mathbf{z}(t) = \mathbf{H}\mathbf{u}(t) + \mathbf{n}(t) \tag{7}$$

Eq. (7) represents the encoder model in vector form and it is used to encode the attended speech from the knowledge of the two competing speech signals and the neural response.

3.1 Training the models

Estimation of Backward model The decoder $\mathbf{d}$ is pretrained to optimally reconstruct the attended speech signal from the EEG data. Assuming that there are T samples available, then we can form the vector of stimulus $\mathbf{u} = [u(1), \ldots, u(T)]^T \in R^{T\times 1}$ and the matrix of response $\mathbf{E} = [\mathbf{e}(1), \ldots, \mathbf{e}(1)]^T \in R^{T\times LC}$, so that (4) becomes $\mathbf{u} = \mathbf{E}\mathbf{d}$. Thus the estimation of decoder vector $\mathbf{d}$ in a least square (LS) formulation is equivalent to the minimum MSE

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \|\mathbf{u} - \mathbf{E}\mathbf{d}\|_F^2 \tag{8}$$

More generally, assuming the availability of N training segments of T time samples, $\{\mathbf{u}^{(i)}, \mathbf{E}^{(i)}, i = 1 : N\}$, thus (8) can be rewritten as

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \sum_{i=1}^{N} \|\mathbf{u}^{(i)} - \mathbf{E}^{(i)}\mathbf{d}\|_F^2 \tag{9}$$

which, by defining $\hat{\mathbf{u}} = [\mathbf{u}^{(1)^T}, \ldots, \mathbf{u}^{(N)^T}]^T \in R^{NT\times 1}$ and $\hat{\mathbf{E}} = \begin{pmatrix} \mathbf{E}^{(1)} \\ \ldots \\ \mathbf{E}^{(N)} \end{pmatrix} \in R^{NT\times LC}$, becomes

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \|\hat{\mathbf{u}} - \hat{\mathbf{E}}\mathbf{d}\|_F^2 \tag{10}$$

.

The solution of (10) represents the estimation of decoder $\mathbf{d}$ and is given by

$$\hat{\mathbf{d}} = (R_{zz})^{-1} r_{zu} \tag{11}$$

with

$$R_{zz} = \mathbf{E}^T\mathbf{E} \in R^{LC\times LC}, \tag{12}$$

the estimated EEG autocorrelation matrix, and

$$r_{zu} = \mathbf{E}^T\hat{\mathbf{u}} \in R^{LC\times 1} \tag{13}$$

the estimated cross correlation vector between the EEG and the attended speech.

Estimation of Forward model In order to estimate the TRF $h(t, \tau)$, (6) is rewritten as

$$z(t, c) = \mathbf{u}^T(t)\mathbf{h}_c + n(t, c), \;\; c = 1 : C \tag{14}$$

and for T time samples we have

$$\mathbf{z}_c = \mathbf{U}\mathbf{h}_c + \mathbf{n}_c, \tag{15}$$

where $\mathbf{z}_c = [z(1, c), \ldots, z(T, c)]^T \in R^{T\times 1}$, $\mathbf{U} = [\mathbf{u}(1), \ldots, \mathbf{u}(T)]^T \in R^{T\times L}$, and $\mathbf{n}_c = [n(1, c), \ldots, n(T, c)]^T \in R^{T\times 1}$. Defining $\mathbf{z} = [\mathbf{z}_1{}^T, \ldots, \mathbf{z}_c{}^T]^T \in R^{TC\times 1}$,
$\mathbf{h} = [\mathbf{h}_1{}^T, \ldots, \mathbf{h}_C{}^T]^T \in R^{LC\times 1}$ and the block diagonal matrix $\mathcal{U} = diag(\mathbf{U}, \ldots, \mathbf{U}) \in R^{TC\times LC}$, (15) becomes

$$\mathbf{z} = \mathcal{U}\mathbf{h} + \mathbf{n}. \tag{16}$$

Using the unconstrained least square method, the estimate $\hat{\mathbf{h}}$ minimizes the norm

$$\epsilon = \|\mathbf{z} - (\mathcal{U}\mathbf{h} + \mathbf{n})\|^2. \tag{17}$$

Assuming $\mathcal{U}\mathbf{h}$ and $\mathbf{n}$ are orthogonal, if $\mathcal{U}^T\mathcal{U}$ is nonsingular, a solution of (17) is given by

$$\hat{\mathbf{h}} = \mathcal{U}^{+}\mathbf{z} \tag{18}$$

where $\mathcal{U}^{+} = (\mathcal{U}^T\mathcal{U})^{-1}\mathcal{U}^T$ is the pseudoinverse of $\mathcal{U}$.

Jointly estimation of forward and backward model The forward and backward models can be jointly estimated by canonical correlation analysis (CCA), in such a way their outputs are maximally correlated. Before proceeding with the CCA-based scheme for model training, a brief summary of the CCA algorithm is given in the following. Let us consider two $r.v.'s$ x, y such that $Ex = Ey = 0$ and two directions w_x,w_y. then we can define the one-dimensional $r.v.'s$

$$\alpha_x = w_x^T x \tag{19}$$

$$\alpha_y = w_y^T y \tag{20}$$

The objective of CCA is to determine w_x,w_y such that the correlation between α_x and α_y is maximum. This problem corresponds to maximize the Pearson coefficient

$$\max_{w_x,w_y} \rho = \max_{w_x,w_y} \frac{E\{\alpha_x\alpha_y\}}{\sqrt{E\{\alpha_x^2\}}\sqrt{E\{\alpha_y^2\}}} \tag{21}$$

which can be rewritten as

$$\max_{w_x,w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}} \tag{22}$$

and C_{xy}, C_{xx}, C_{yy} are entries of the covariance matrix C of vector $(x^T, y^T)^T$

$$C = E\{\begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}^T\} = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \tag{23}$$

It is worth to notice that (22) is not affected by rescaling w_x and/or w_y, and since rescaling is arbitrary, thus solving (22) is equivalent to the problem

$$\max_{w_x,w_y} w_x^T C_{xy} w_y \quad s.t. \quad w_x^T C_{xx} w_x = 1, w_y^T C_{yy} w_y = 1 \tag{24}$$

This is a typical quadratic problem that can be written in Lagrangian form as

$$L(\lambda_x, \lambda_y, w_x, w_y) = w_x^T C_{xy} w_y - \frac{\lambda_x}{2}(w_x^T C_{xx} w_x - 1) - \frac{\lambda_y}{2}(w_y^T C_{yy} w_y - 1) \tag{25}$$

and maximizing (25) is equivalent to the equations

$$\frac{dL}{dw_x} = C_{xy} w_y - \lambda_x C_{xx} w_x = 0 \tag{26}$$

$$\frac{dL}{dw_y} = C_{yx}w_x - \lambda_y C_{yy}w_y = 0 \tag{27}$$

Multiplying the first equation for ${w_x}^T$ and the second equation for ${w_y}^T$ and then subtracting the second equation from the first one, we have

$$0 = \lambda_y w_y^T C_{yy} w_y - \lambda_x w_x^T C_{xx} w_x = \lambda_y - \lambda_x \tag{28}$$

where the constraints in (24) are used. Assuming $\lambda = \lambda_y = \lambda_x$ and C_{yy} is invertible, the equation (27) gives

$$w_y = \frac{C_{yy}^{-1} C_{yx} w_x}{\lambda} \tag{29}$$

and substituting in (27) yields

$$C_{xy}C_{yy}^{-1}C_{yx}w_x = \lambda^2 C_{xx}w_x \tag{30}$$

Finally assuming C_{xx} is invertible it results

$$C_{xx}^{-1}C_{xy}C_{yy}^{-1}C_{yx}w_x = \lambda^2 w_x \tag{31}$$

Similarly from (26) we have

$$C_{yy}^{-1}C_{yx}C_{xx}^{-1}C_{xy}w_x = \lambda^2 w_y \tag{32}$$

The canonical coefficients w_x, w_y are the eigenvectors of matrices $C_{xx}^{-1}C_{xy}C_{yy}^{-1}C_{yx}$ and $C_{yy}^{-1}C_{yx}C_{xx}^{-1}C_{xy}$ respectively, with the same eigenvalue λ^2.
Now, let us proceed with the training problem by CCA that combines a decoding model $w_z \in R^{LC\times 1}$ and an encoding model $w_u \in R^{L\times 1}$. Decoding model

$$\alpha_z = w_z^T \mathbf{Z}(t) \tag{33}$$

is applied to the EEG signal $\mathbf{Z}(t) = [\mathbf{z}(t), \mathbf{z}(t-1), \ldots, \mathbf{z}(t-L+1)]^T \in R^{LC\times 1}$, where $\mathbf{z}(t) = [z(t,1), \ldots, z(t,C)]^T \in R^{C\times 1}$, with the channel index c ranging from 1 to C, and the time lag ranging from 0 to $(L-1)$. Encoding model

$$\alpha_u = w_u^T \mathbf{u}(t) \tag{34}$$

acts on the speech signal $\mathbf{u}(t) = [u(t), u(t-1), \ldots, u(t-L+1)]^T \in R^{L\times 1}$, L being the number of filter taps of the encoding filter. Thus, forward and backward models are jointly estimated in such a way their outputs are maximally correlated

$$\max_{w_z, w_u} \frac{E\{\alpha_z \alpha_u\}}{\sqrt{E\{\alpha_z^2\}}\sqrt{E\{\alpha_u^2\}}} = \max_{w_z, w_u} \frac{w_z^T C_{zu} w_u}{\sqrt{w_z^T C_{zz} w_z w_u^T C_{uu} w_u}}, \tag{35}$$

that corresponds to solving eqs.(31) and (32).

3.2 Decoding the attended speech

Once the decoder has been trained, the reconstructed speech is correlated with the speech of all speakers, after which the one with the highest Pearson coefficient is identified as the attended speaker. Three different schemes correspond to the models previously discussed.

A) With reference to the backward model, given the estimated decoder $\hat{\mathbf{d}}$ and the T_{test} time samples of an EEG response $\mathbf{E}^{(test)} \in R^{T_{test} \times LC}$ of a subject listening to one out of two competing speakers with speech stimuli $\mathbf{u}_1{}^{(test)}$ and $\mathbf{u}_2{}^{(test)}$, a decision about the auditory attention to the listener can be made by:
1) reconstructing the attended speech from EEG

$$\hat{\mathbf{u}}^{(test)} = \mathbf{E}^{(test)} \hat{\mathbf{d}} \tag{36}$$

2) computing the Pearson correlation coefficients

$$\rho(\hat{\mathbf{u}}^{(test)}, \mathbf{u}_1{}^{(test)}), \;\; \rho(\hat{\mathbf{u}}^{(test)}, \mathbf{u}_2{}^{(test)}) \tag{37}$$

between the reconstructed responses to stimuli and the EEG signal. where $\rho(.,.)$ is defined as

$$\rho = \frac{E\{x^T y\}}{\sqrt{E\{x^T x\}}\sqrt{E\{y^T y\}}} \tag{38}$$

The speaker corresponding to the highest coefficient is identified as the attended speaker.

B) By referring to the forward model, once the decoder $\hat{\mathbf{h}}$ has been estimated and the EEG response $\mathbf{z}^{(test)}$ is given, then the Pearson coefficients

$$\rho(\mathbf{z}^{(test)}, \mathcal{U}_1^T \hat{\mathbf{h}}), \;\; \rho(\mathbf{z}^{(test)}, \mathcal{U}_2^T \hat{\mathbf{h}}) \tag{39}$$

are computed, and the highest coefficient identifies the attended speaker. A comparative study of the ability of different estimation methods in backward and forward models, to classify attended speakers from multi-channel EEG data, is reported in [77].

C) Similarly, in the case of jointly estimation of forward and backward models, given the estimated decoders w_z, w_u and the T time samples of an EEG response $\mathbf{Z}(t)^{(test)} \in R^{TC \times 1}$, the Pearson coefficients

$$\rho(w_z^T \mathbf{Z}(t)^{(test)}, w_u^T \mathbf{u}_1(t)^{(test)}), \;\; \rho(w_z^T \mathbf{Z}(t)^{(test)}, w_u^T \mathbf{u}_2(t)^{(test)}) \tag{40}$$

are computed, and the speaker with the highest coefficient is identified as the attended speaker [78].

4 Forward Nonlinear model (FNL) for AAD

CCA model can be easily generalized by finding a pair of optimal nonlinear transforms for the outputs of both forward and backward models, such that the two new projections are maximally correlated. To this end let $f_z(\mathbf{Z}(t), \theta_z)$ and $f_u(\mathbf{u}(t), \theta_u)$ two nonlinear functions that generalize the transforms (33) and (34)

$$\alpha_z = f_z(\mathbf{Z}(t), \theta_z) \tag{41}$$

$$\alpha_u = f_u(\mathbf{u}(t), \theta_u) \tag{42}$$

where θ_z, θ_u are unknown parameters to be determined by training stage. Thus their optimal values are obtained by solving the following optimization problem

$$(\hat{\theta_z}, \hat{\theta_u}) = arg \max_{\theta_z, \theta_u} \rho(f_z(\mathbf{Z}(t), \theta_z), f_u(\mathbf{u}(t), \theta_u)) \tag{43}$$

where ρ represents the Pearson coefficient. A very efficient approach to face this problem is to use two neural networks to implement the nonlinear functions (41) and (42), and this algorithm is called deep CCA [60], [61], [62]. The main drawbacks of this approach is that the two nonlinear functions f_z and f_u have not a physical meaning and that it requires two neural networks to be trained. The objective of this section is to describe an approach that is more adherent to the physical model of the audio-EEG system and it requires only one neural network to be trained. Here we assume the neural response in each EEG channel can be predicted from the speech stimulus via a forward nonlinear (FNL) model. The model is obtained from (7) by simply replacing the linear operator $\mathbf{Hu}(t)$ with a nonlinear operator $T(\mathbf{u}(t))$, thus we have

$$\mathbf{z}(t) = T(\mathbf{u}(t)) + \mathbf{n}(t) \tag{44}$$

where $T(.)$ is in general a nonlinear dynamical transformation, explicity depending on time t. Now, let us refer to a frame $\mathbf{u} = [u(1), \ldots, u(T)]^T \in R^{T \times 1}$ of the speech stimulus, thus the response obtained from (44) is given by

$$\mathbf{z} = T(\mathbf{u}) + \mathbf{n} = \mathbf{y} + \mathbf{n} \tag{45}$$

where $\mathbf{z} = [\mathbf{z}(1)^T, \ldots, \mathbf{z}(T)^T]^T \in R^{TC \times 1}$, $\mathbf{n} = [\mathbf{n}(1), \ldots, \mathbf{n}(T)]^T \in R^{TC \times 1}$, $\mathbf{y} = T(\mathbf{u}) \in R^{TC \times 1}$. Assuming $\mathbf{u}$ is a random signal whose covariance matrix is $R_{uu} = E\{\mathbf{u}\mathbf{u}^T\}$, thus R_{uu} can be decomposed as

$$R_{uu} = \Psi \Gamma \Psi^T \tag{46}$$

where ψ is a unitary matrix. Due to the orthogonality property of Ψ, the following representation for $\mathbf{u}$

$$\mathbf{u} = \Psi \mathbf{x} \tag{47}$$

$$\mathbf{x} = \Psi^T \mathbf{u} \tag{48}$$

holds.

Similarly for $\mathbf{y}$ it results

$$\mathbf{y} = \Phi \mathbf{k} \tag{49}$$

$$\mathbf{k} = \Phi^T \mathbf{y} \tag{50}$$

where Φ derives from the decomposition of $R_{yy} = E\{\mathbf{y}\mathbf{y}^T\} = \Phi \Lambda \Phi^T$. From (50) we have

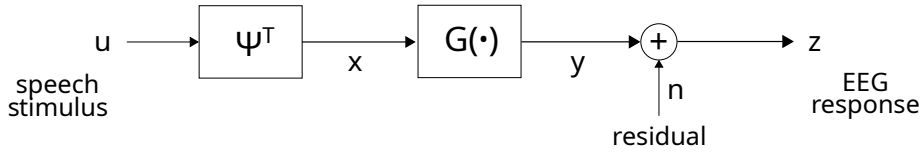

Figure 2: Scheme representing the FNL model of audio-EEG system.

$$\mathbf{k} = \Phi^T \mathbf{y} = \Phi^T T(\Psi \mathbf{x}) = \Phi^T T(\Psi \mathbf{x}) \tag{51}$$

Defining $f(\mathbf{x}) = \Phi^T T(\Psi \mathbf{x})$, (51) is equivalent to

$$\mathbf{k} = f(\mathbf{x}) \tag{52}$$

where $f(\mathbf{x})$ is a nonlinear function of $\mathbf{x}$. Thus (45) can be rewritten as

$$\mathbf{z} = \Phi f(\mathbf{x}) + \mathbf{n} = G(\mathbf{x}) + \mathbf{n} \tag{53}$$

where $G(\mathbf{x}) = \Phi f(\mathbf{x})$. Finally the FNL model of audio-EEG system is given by

$$\mathbf{z} = G(\Psi^T \mathbf{u}) + \mathbf{n} \tag{54}$$

that is depicted in the scheme of Fig. 2.

4.1 Training the FNL model

The nonlinear function $G(.)$ in (54) represents the encoder of the auditory model, and the objective of training is to identify this input-output mapping. Due to nonlinearity this function can be fruitfully estimated by a neural network $\hat{G}(\mathbf{x}; \theta)$, with θ parameter vector to be estimated during training stage. The optimal solution θ can be chosen such that $\mathbf{z}$ and $\hat{G}(\mathbf{x}; \theta)$ are maximally correlated, thus the parameter θ is the solution of the optimal problem

$$\max_{\theta} \rho(\mathbf{z}, \hat{G}(\mathbf{x}; \theta)) \tag{55}$$

Eq.(55) corresponds to the training of neural network $\hat{G}(\mathbf{x}; \theta)$, and with reference to the train dataset $\Omega = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}, i = 1 : N\}$, it becomes

$$\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \{\rho(\mathbf{z}^{(i)}, \hat{G}(\mathbf{x}^{(i)}; \theta))\} \tag{56}$$

4.2 Decoding the attended speech

Given the estimated decoder $\hat{G}$ and the T time samples of an EEG response $\mathbf{z}^{(test)} \in R^{TC \times 1}$ of a subject listening to one out of two competing speakers with speech stimuli ${\mathbf{u}_1}^{(test)}$ and ${\mathbf{u}_2}^{(test)}$ a decision about the auditory attention to the listener can be made by:
1) reconstructing the responses to stimuli

$$\hat{G}(\Psi^T \mathbf{u}_1^{(test)}, \theta) \tag{57}$$

$$\hat{G}(\Psi^T \mathbf{u}_2^{(test)}, \theta) \tag{58}$$

2) computing the Pearson correlation coefficients

$$\rho(\mathbf{z}^{(test)}, \hat{G}(\Psi^T \mathbf{u}_1^{(test)}, \theta)), \quad \rho(\mathbf{z}^{(test)}, \hat{G}(\Psi^T \mathbf{u}_2^{(test)}, \theta)) \tag{59}$$

between the reconstructed responses to stimuli and the EEG signal. The speaker corresponding to the highest coefficient is identified as the attended speaker.

5 A deep learning approach for the estimation of FNL model

The aim of this Section is to present a novel framework to derive the nonlinear model $G(.)$ in (54) of the speech-EEG system. The model is based on a neural network with a loss function specifically designed to maximize the correlation between EEG and response to speech stimulus.

5.1 Dataset

The Speech-EEG dataset [79] is a collection of five open-source speech-listening datasets (Cocktail Party Dataset, N400 Dataset, Natural Speech - Reverse Dataset, Natural Speech Dataset, and Speech in Noise Dataset) derived from the study conducted in [80]. This study focuses on the electrophysiological correlates of continuous natural speech understanding and analyzes EEG data involving time-reversed speech, cocktail party attention, and audiovisual speech in noise, demonstrating that the human brain response is very sensitive to whether or not subjects understood the speech they heard.

For the experimentation described in Section 5.4, the chosen dataset is the Natural Speech Dataset, which has been widely used in literature to implement both linear CCA [81] and deep CCA [60, 82] models. The dataset includes:

- EEG traces of 19 subjects engaged in 20 trials of an experiment involving listening to a single audio book.
- 20 stimulus files containing the audio of a male speaker reading snippets of a novel and the associated envelope.

During all trials, 128-channel EEG data were collected at a rate of 512 Hz using an ActiveTwo system from BioSemi. The provided dataset has already undergone several offline preprocessing steps: it was band-pass filtered between 1 and 8 Hz, down-sampled to 128 Hz, and re-referenced to the average of the mastoid channels.

5.2 Data preprocessing

Instead of working on the time-based original signals, the proposed method operates on the input and output spectra achieved through the discrete Karhunen–Loève transform (DKLT) algorithm, as reported in eqs. (47)-(50). This process allows the audio and EEG data can be expressed in an optimal basis as sets of static features. The algorithm, along with other preprocessing operations, will be discussed in the following subsections.

5.2.1 CCA

Generally, datasets comprising EEG responses to specific stimuli exhibit distinct properties for inputs and outputs. Input stimuli, such as audio signals in our case, typically have low dimensionality but a higher sampling frequency. In contrast, outputs, i.e. EEG recordings, have a lower frequency but a substantial number of components that corresponds to the number of EEG tracks. To compute the correlation between these two components, specifically the Pearson correlation coefficient, it is desirable to represent both the input and output as one-dimensional vectors of the same size.

For this reason, the following preprocessing steps have been applied to input and output signals.

Regarding the input audio signals, the envelope of the audio was extracted, as this approach has been demonstrated in the literature to better model the dependency between auditory input and EEG output [60, 80, 83–85]. Additionally, a notable advantage is that the envelope can be computed at a lower sampling frequency, aligning it with the frequency of the EEG signal.

Concerning the EEG data, the data dimensionality was reduced by computing the linear CCA between the audio envelope and the EEG and keeping only one component. As a result, a decomposition for both inputs and outputs is achieved, rendering both sets of data mono-dimensional and maximally correlated. Consequently, this enables an initial input and output with improved correlation compared to the original data. In the original dataset, high levels of noise and the inherent complexity of cerebral activity hinder the identification of meaningful correlations between individual EEG tracks and input audio.

5.2.2 Data windowing

Usually, in common datasets used for EEG experimentation, input-output pairs for every subject are limited in number and heterogeneous in size. In order to have a suitable input set for the neural network, a common technique consists in dividing the input data in (possibly overlapping) windows of fixed size along the time axis.

Given an original mono-dimensional signal of size n, the windowing process results in an $N \times w$ matrix, with

$$N = 1 + (n - w)/(w - v) \tag{60}$$

being w the window size and v the number of overlapping samples between subsequent windows. The parameters w and v are empirically chosen, and in this work they have been set to 256 and 128, respectively, for both input and output signals.

5.2.3 Feature extraction

The input and output matrices obtained as a result of previous step, are transformed by the discrete Karhunen–Loève transform (DKLT) (eqs. (48)-(50)), to extract a set of features that better identifies the informative contents of data. A well-known property of DKLT is that is able to represent data in a suitable basis as a time-independent feature vector. Operatively, the DLKT uses the singular value decomposition (SVD) of a matrix X to solve (47) and the companion equation for basis Φ :

$$X = USV^T \tag{61}$$

being S the diagonal matrix of singular values, and U, V the eigenvector matrices. From that, X can be transformed to a new matrix in the feature space as

$$K = V^T X \tag{62}$$

It's also customary to apply a principal component analysys (PCA) to truncate the resulting DKLT representation to a subset of the most significant components. This step is performed by keeping only a limited number of eigenvectors, corresponding to the largest singular values. This process helps in keeping the computational complexity low, while often improving the quality of the network, since only the most significant parts of the signal are kept, removing possible noise components or other unrelated artifacts. In this work the DKLT has been empirically truncated to the 50 most representative components.

5.3 Neural network and loss function

In order to estimate the response to speech stimuli $\hat{G}(\Psi^T \mathbf{u}, \theta)$, a multilayer perceptron (MLP) neural network with a custom loss function was developed. A multilayer perceptron is a widely used artificial neural network architecture, in the fields of machine learning and deep learning for several tasks like classification, regression and pattern recognition. An MLP is composed by three or more layers of interconnected nodes (neurons), organized in a hierarchical structure. The first layer is usually an input layer, that receives the input data and each neuron in this layer corresponds to a feature of the input data. Then one or more hidden layers are inserted between the input and output of the network. Each element or neuron in this layer applies a linear combination to the outputs of the previous layer and successively a non-linear activation function. The purpose of hidden layers is to identify complex patterns within the data. Finally, an output layer generates the final network outputs, with a number of neurons depending on the specific task.

The architecture developed in this work is depicted in Figure 3, with a detailed description of each layers reported in Table 1: it is composed of four fully-connected layers (Dense), the first and the last ones representing the input and output, respectively. Dropout layers, which discard a random portion of the data at different parts of the processing chain, were introduced to avoid overfitting phenomenon, caused by a too closely adjustment of network parameters to the training data, thus reducing the testing performance. The networks were trained for 1000 epochs on the Speech-EEG dataset (Natural Speech dataset) by using an Adam optimizer with the custom loss function described below.

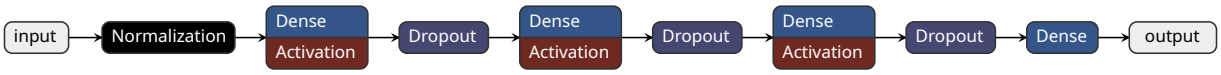

Figure 3: Neural network architecture.

Table 1: Neural network summary.

Layer (type)	Output Shape	Parameters
Normalization	(None, 50)	101
dense (Dense)	(None, 1000)	51000
dropout (Dropout)	(None, 1000)	0
dense_1 (Dense)	(None, 1000)	1001000
dropout_1 (Dropout)	(None, 1000)	0
dense_2 (Dense)	(None, 500)	500500
dropout_2 (Dropout)	(None, 500)	0
dense_3 (Dense)	(None, 50)	25050

Optimization of a neural network is driven by a loss function $\lambda(\mathbf{z}, \hat{G}(\mathbf{x};\theta))$, evaluated at every step of the training. The training process must minimize such function, usually so that $\hat{G}$ tends towards $\mathbf{z}$ according to a desired criterion. This in turn results in optimal values for the network parameters.

Accordingly to (55), the loss function has been suitably coded to maximize the correlation between $\mathbf{z}$ and $\hat{G}(\mathbf{x};\theta)$, as:

$$\lambda(\mathbf{z}, \hat{G}(\mathbf{x};\theta)) = 1 - |\rho(\mathbf{z}, \hat{G}(\mathbf{x};\theta))| \tag{63}$$

5.4 Experimental results

In this section we compare the linear CCA modeling

$$\max_{w_z, w_u} \rho(w_z^T \mathbf{Z}(t), w_u^T \mathbf{u}(t)), \tag{64}$$

that maximizes the Pearson correlation coefficient between the transformed vectors $w_z^T \mathbf{Z}(t), w_u^T \mathbf{u}(t)$, with the FNL modeling

$$\max_{\theta} \rho(\mathbf{z}, \hat{G}(\Psi^T \mathbf{u}, \theta)) \tag{65}$$

that maximizes the Pearson correlation coefficient between the estimated response to stimulus $\hat{G}(\Psi^T \mathbf{u})$ and the EEG signal $\mathbf{z}$. To estimate $\hat{G}(\Psi^T \mathbf{u}, \theta)$ an MLP with a custom loss was developed as described in Section 5.3.

The neural network was implemented using TensorFlow v. 2.15.0 and Keras v .15.0 on a computer equipped with an Intel Core i7-6800K CPU, 32 GiB of RAM, and a GeForce GTX 1080 GPU.

To test the effectiveness of the network, for every subject a cross-testing strategy was adopted: a different train and testing session is performed for every pair of input–output, i.e. stimulus and corresponding EEG signal, where the given pair is used as testing data, while the remaining data are used for training and validation. Finally, the average results were computed for the given subject, as reported in Table 2. The values in the second column refer to the average linear CCA method computed on the two original data views, i.e, the envelope of the original wave file and the EEG for each subjects, while the Pearson correlation coefficients reported in the third column were computed using (65) for each subject and then applying the average. As you can see, the proposed nonlinear method outperforms linear CCA in all the subjects considered in the experiments.

Table 2: Average Pearson Correlation coefficient between EEG signal $\mathbf{z}$ and the estimated response to stimulus $\hat{G}(.)$ of the proposed nonlinear model, compared with that of linear CCA, computed for all subjects of the Speech-EEG dataset, using cross-validation for testing.

Subject	Linear CCA	Proposed method
1	0.0381	0.1019
2	0.0354	0.0939
3	0.0445	0.1247
4	0.0449	0.1271
5	0.0452	0.1197
6	0.0321	0.0730
7	0.0292	0.0778
8	0.0382	0.0898
9	0.0387	0.0843
10	0.0575	0.1143
11	0.0329	0.0866
12	0.0312	0.1041
13	0.0381	0.1161
14	0.0398	0.0955
15	0.0356	0.1031
16	0.0357	0.0962
17	0.0395	0.1028
18	0.0361	0.0844
19	0.0364	0.0779

6 Concluding remarks

Linear modeling of the neural entrainment is the common approach in the development of AAD algorithms. However, even though these algorithms have shown to be effective in solving CCP, they have some inherent limitations. This chapter presents a thorough discussion of both linear and nonlinear modeling for AAD and suggests a technique based on a novel forward nonlinear modeling approach for AAD. Experimental results confirm that nonlinear modeling outperforms linear modeling, in terms of correlation between speech stimulus and neural response, thus highlighting the potential limitations of linear approach in developing a new generation of neuro-steered hearing aids for CCP.

References

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] L. L. Cunningham and D. L. Tucci, "Hearing loss in adults," *New England Journal of Medicine*, vol. 377, no. 25, pp. 2465–2473, 2017.

[4] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[5] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[6] E. C. Bluvas and T. Q. Gentner, "Attention to natural auditory signals," *Hearing research*, vol. 305, pp. 10–18, 2013.

[7] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party"," *Journal of Neuroscience*, vol. 30, no. 2, pp. 620–628, 2010.

[8] A. Simon, G. Loquet, J. Østergaard, and S. Bech, "Cortical auditory attention decoding during music and speech listening," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[9] M. Geravanchizadeh and S. B. Gavgani, "Selective auditory attention detection based on effective connectivity by single-trial EEG," *Journal of neural engineering*, vol. 17, no. 2, p. 026021, 2020.

[10] L. Wang, E. X. Wu, and F. Chen, "Robust EEG-based decoding of auditory attention with high-rms-level speech segments in noisy conditions," *Frontiers in human neuroscience*, vol. 14, p. 557534, 2020.

[11] S. Crottaz-Herbette and V. Menon, "Where and when the anterior cingulate cortex modulates attentional response: combined fmri and erp evidence," *Journal of cognitive neuroscience*, vol. 18, no. 5, pp. 766–780, 2006.

[12] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, p. 311, 2014.

[13] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon *et al.*, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[14] V. Menon and S. Crottaz-Herbette, "Combined EEG and fmri studies of human brain function," *International review of neurobiology*, vol. 66, pp. 291–321, 2005.

[15] E. S. Sussman, "Auditory scene analysis: an attention perspective," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 10, pp. 2989–3000, 2017.

[16] M. Haghighi, M. Moghadamfalahi, M. Akcakaya, and D. Erdogmus, "EEG-assisted modulation of sound sources in the auditory scene," *Biomedical signal processing and control*, vol. 39, pp. 263–270, 2018.

[17] G. Bajwa, M. Fazeen, and R. Dantu, "Detecting driver distraction using stimuli-response EEG analysis," *arXiv preprint arXiv:1904.09100*, 2019.

[18] I. Choi, S. Rajaram, L. A. Varghese, and B. G. Shinn-Cunningham, "Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography," *Frontiers in human neuroscience*, vol. 7, p. 115, 2013.

[19] D. J. Lee, H. Jung, and P. Loui, "Attention modulates electrophysiological responses to simultaneous music and language syntax processing," *Brain sciences*, vol. 9, no. 11, p. 305, 2019.

[20] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach," *Frontiers in neuroscience*, vol. 12, p. 262, 2018.

[21] M. Scherg, J. Vajsar, and T. W. Picton, "A source analysis of the late human auditory evoked potentials," *Journal of cognitive neuroscience*, vol. 1, no. 4, pp. 336–355, 1989.

[22] H. Dillon, *Hearing aids*. Thieme Medical Publishers, 2012.

[23] S. Doclo and M. Moonen, "Gsvd-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on signal processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[24] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[25] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2016.

[26] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, 2020.

[27] J. A. O'sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[28] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, 2015.

[29] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2016.

[30] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of neural engineering*, vol. 15, no. 6, p. 066017, 2018.

[31] S. Zhao, J. Han, X. Jiang, H. Huang, H. Liu, J. Lv, L. Guo, and T. Liu, "Decoding auditory saliency from brain activity patterns during free listening to naturalistic audio excerpts," *Neuroinformatics*, vol. 16, pp. 309–324, 2018.

[32] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[33] S. Akram, J. Z. Simon, and B. Babadi, "Dynamic estimation of the auditory temporal response function from meg in competing-speaker environments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1896–1905, 2016.

[34] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[35] A. J. Casson, "Wearable EEG and beyond," *Biomedical engineering letters*, vol. 9, no. 1, pp. 53–71, 2019.

[36] S. Geirnaert, T. Francart, and A. Bertrand, "An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 307–317, 2019.

[37] ——, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1557–1568, 2020.

[38] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'," *Journal of neural engineering*, vol. 11, no. 4, p. 046015, 2014.

[39] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification," *Journal of neural engineering*, vol. 11, no. 2, p. 026009, 2014.

[40] Z. Xu, Y. Bai, R. Zhao, H. Hu, G. Ni, and D. Ming, "Decoding selective auditory attention with EEG using a transformer model," *Methods*, vol. 204, pp. 410–417, 2022.

[41] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLoS computational biology*, vol. 17, no. 9, p. e1009358, 2021.

[42] G. Cantisani, S. Essid, and G. Richard, "EEG-based decoding of auditory attention to a target instrument in polyphonic music," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 80–84.

[43] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160101, 2017.

[44] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers in neuroscience*, p. 153, 2019.

[45] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[46] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and decoding models in cognitive electrophysiology," *Frontiers in systems neuroscience*, vol. 11, p. 61, 2017.

[47] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand, "Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario," in *2015 37th annual international conference of the ieee engineering in medicine and biology society (embc)*. IEEE, 2015, pp. 5155–5158.

[48] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Auditory attention decoding with EEG recordings using noisy acoustic reference signals," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 694–698.

[49] S. Geirnaert, T. Francart, and A. Bertrand, "Time-adaptive unsupervised auditory attention decoding using EEG-based stimulus reconstruction," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3767–3778, 2022.

[50] X. Zhuang, Z. Yang, and D. Cordes, "A technical review of canonical correlation analysis for neuroscience applications," *Human Brain Mapping*, vol. 41, no. 13, pp. 3807–3833, 2020.

[51] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[52] D. Weenink, "Canonical correlation analysis," in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, vol. 25. University of Amsterdam Amsterdam, 2003, pp. 81–99.

[53] X. Yang, W. Liu, W. Liu, and D. Tao, "A survey on canonical correlation analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2349–2368, 2019.

[54] A. de Cheveigné, G. M. Di Liberto, D. Arzounian, D. D. Wong, J. Hjortkjær, S. Fuglsang, and L. C. Parra, "Multiway canonical correlation analysis of brain data," *neuroimage*, vol. 186, pp. 728–740, 2019.

[55] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Network: Computation in neural systems*, vol. 18, no. 3, pp. 191–212, 2007.

[56] S. Akaho, "A kernel method for canonical correlation analysis," *arXiv preprint cs/0609071*, 2006.

[57] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International journal of neural systems*, vol. 10, no. 05, pp. 365–377, 2000.

[58] R. Sawata, T. Ogawa, and M. Haseyama, "Novel audio feature projection using kdlpcca-based correlation with EEG features for favorite music classification," *IEEE transactions on affective computing*, vol. 10, no. 3, pp. 430–444, 2017.

[59] T. Melzer, M. Reiter, and H. Bischof, "Nonlinear feature extraction using generalized canonical correlation analysis," in *International Conference on Artificial Neural Networks*. Springer, 2001, pp. 353–360.

[60] J. R. Katthi and S. Ganapathy, "Deep correlation analysis for audio-EEG decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2742–2753, 2021.

[61] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[62] A. Aroudi, T. De Taillez, and S. Doclo, "Improving auditory attention decoding performance of linear and non-linear methods using state-space model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8703–8707.

[63] S. Cai, P. Li, E. Su, Q. Liu, and L. Xie, "A neural-inspired architecture for EEG-based auditory attention detection," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 668–676, 2022.

[64] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, p. e56481, 2021.

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

[66] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, p. 11538, 2019.

[67] Z. Fu, B. Wang, X. Wu, and J. Chen, "Auditory attention decoding from EEG using convolutional recurrent neural network," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 970–974.

[68] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[69] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13 367–13 372, 2001.

[70] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, 2007.

[71] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and hearing*, vol. 29, no. 2, pp. 139–157, 2008.

[72] K. V. Nourski, R. A. Reale, H. Oya, H. Kawasaki, C. K. Kovach, H. Chen, M. A. Howard, and J. F. Brugge, "Temporal envelope of time-compressed speech represented in the human auditory cortex," *Journal of Neuroscience*, vol. 29, no. 49, pp. 15 564–15 574, 2009.

[73] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, p. e1001251, 2012.

[74] V. Z. Marmarelis, *Nonlinear dynamic modeling of physiological systems*. John Wiley & Sons, 2004, vol. 10.

[75] D. A. Abrams, T. Nicol, S. Zecker, and N. Kraus, "Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech," *Journal of Neuroscience*, vol. 28, no. 15, pp. 3958–3965, 2008.

[76] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European journal of neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[77] D. D. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. De Cheveigne, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers in neuroscience*, vol. 12, p. 531, 2018.

[78] J. P. Dmochowski, J. J. Ki, P. DeGuzman, P. Sajda, and L. C. Parra, "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity," *NeuroImage*, vol. 180, pp. 134–146, 2018.

[79] "Data from: Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," https://doi.org/10.5061/dryad.070jc, accessed: 2023-07-24.

[80] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809.e3, 2018.

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

[82] J. R. Katthi, S. Ganapathy, S. Kothinti, and M. Slaney, "Deep canonical correlation analysis for decoding the auditory brain," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 3505–3508.

[83] P. Pandey, N. Ahmad, K. P. Miyapuram, and D. Lomas, "Predicting dominant beat frequency from brain responses while listening to music," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 3058–3064.

[84] B. Kaneshiro, D. T. Nguyen, A. M. Norcia, J. P. Dmochowski, and J. Berger, "Natural music evokes correlated EEG responses reflecting temporal structure and beat," *NeuroImage*, vol. 214, p. 116559, 2020.

[85] G. Di Liberto, J. O'Sullivan, and E. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.