Deep Correlation Analysis for Audio-EEG Decoding

Jaswanth Reddy Katthi, *Student Member, IEEE*, and Sriram Ganapathy[iD], *Senior Member, IEEE*

Abstract—**The electroencephalography (EEG), which is one of the easiest modes of recording brain activations in a non-invasive manner, is often distorted due to recording artifacts which adversely impacts the stimulus-response analysis. The most prominent techniques thus far attempt to improve the stimulus-response correlations using linear methods. In this paper, we propose a neural network based correlation analysis framework that significantly improves over the linear methods for auditory stimuli. A deep model is proposed for intra-subject audio-EEG analysis based on directly optimizing the correlation loss. Further, a neural network model with a shared encoder architecture is proposed for improving the inter-subject stimulus response correlations. These models attempt to suppress the EEG artifacts while preserving the components related to the stimulus. Several experiments are performed using EEG recordings from subjects listening to speech and music stimuli. In these experiments, we show that the deep models improve the Pearson correlation significantly over the linear methods (average absolute improvements of 7.4% in speech tasks and 29.3% in music tasks). We also analyze the impact of several model parameters on the stimulus-response correlation.**

Index Terms—**Canonical correlation analysis (CCA), multiway CCA (MCCA), deep learning, audio-EEG analysis.**

I. INTRODUCTION

UNDERSTANDING the human brain has been a topic of profound interest in both science and engineering fields. One of the most common methods to perform this analysis is to measure the evoked brain response for a given stimuli and establish a relation between them. The electroencephalography (EEG) constitutes the simplest non-invasive technique to collect brain signals while having sufficient temporal resolution for auditory analysis. Since the EEG recordings involve scalp level measurements, the recordings are significantly impacted by noise [1]. The most popular method for analyzing auditory

invoked EEG signals is the classical event-related potential (ERP) approach [2], [3]. This approach involves averaging the EEG responses in time/frequency domain to suppress the noise in the recordings [4]. However, this approach is limited to isolated stimuli that have to be repeated and is therefore often restrictive for use in the analysis of natural stimuli like speech and music.

One of the first successful attempts in this direction is the temporal response function (TRF) proposed by Lalor and Foxe [5]. The linear TRF model describes the relationship between a stimulus and its response as a linear time-invariant (LTI) system. It can be a linear forward model where the model estimates the EEG response for the stimulus or a backward model where the model predicts the components of the stimulus from the EEG response. The model estimation is performed using linear least squares. The performance of these models is typically validated using the Pearson correlation value between the target signal and the predicted signal [6].

The initial studies used the slowly varying speech envelopes of the stimuli and the corresponding single-trial EEG responses [7], [8]. The analysis can also be extended to speech spectrograms [9], phonemes [10], or semantic features [11].

The Canonical Correlation Analysis (CCA) is an extension of the linear methods for analysis. Here, two signals are projected onto a subspace that maximizes the correlation between them [12]. It determines a set of orthogonal directions on which the two signals are highly correlated. The CCA has been shown to be better than forward and backward TRF models in auditory-EEG analysis recently [13], [14].

For each subject, the stimulus and response representations are defined as "views" of the auditory signal. The stimulus-view represents the audio signal using a temporal envelope and the response-view represents it using the brain responses collected as EEG recordings. The linear CCA can be performed only on two views of the data at a time. In order to aggregate the EEG responses from multiple-views (subjects), multiway CCA (MCCA) or generalized CCA [15]–[17] has been proposed. As all views (EEG responses) represent the same object (audio stimulus), some components are common across the views [18]. The application of multiway CCA for EEG mapping has shown improvements over the intra-subject CCA [19].

In this paper, we explore a deep neural network based architecture for correlating the EEG response with the stimulus features. The deep CCA framework, introduced by Andrew *et al.* [20], had shown promise over the linear CCA for image data. However, the direct application of the deep

CCA to the EEG data is cumbersome as the EEG data is significantly noisy with signal-to-noise ratio (SNR) below -20dB [1]. The dropout strategy [21] alleviates the impact of noise partly. In audio-EEG experiments, we show that the deep CCA consistently improves over the linear CCA model.

We also propose an approach for deep multiway CCA where multiple EEG responses for the same stimuli can be combined in a neural network architecture. For this task, we use a reconstruction approach with a shared hidden representation to derive the deep transform that aligns multiple views. Using this novel approach, we show that the deep MCCA improves over the linear MCCA model [19]. In subsequent analysis, we also illustrate how the deep MCCA can be combined with the deep CCA model for EEG analysis.

The data used in the experiments consists of EEG responses for speech and music listening tasks. The speech task is the same dataset used in the linear CCA work by Cheveigné et al. [13] where subjects listen to the narration of an audiobook. The music dataset used in this work is the Naturalistic Music EEG Dataset - Hindi (NMED-H) [22] which is an open dataset of EEG responses collected for Hindi pop songs.

The remainder of the paper is arranged as follows. Section II highlights prior work done in the domain. Section III describes the background of linear CCA and multiway CCA. Section IV discusses the proposed deep CCA model and the deep MCCA model. Section V details the datasets and experimental setup. Section VI reports the results of the proposed deep models and a comparison with linear models. Section VII presents a discussion on the hyper-parameters. Finally, Section VIII presents the summary of this work.

II. RELATED PRIOR WORK

Machine learning methods for the extraction of information from brain signals like EEG have a significant impact on both understanding and applications like brain-computer interfaces (BCI) [23]. It is therefore of profound importance to transfer the recent advancements in machine learning (for example, deep learning [24]) to improve the models for brain signal decoding and single-trial analysis. One of the first works in this direction involved the use of convolutional neural networks to identify the P300 wave in EEG signals [25]. The recent years have seen the use of deep learning for several brain mapping tasks like computational memory prediction [26], driver's cognitive state prediction [27], and the brain activity reconstruction for visual stimuli [28]. A review of several efforts in decoding brain activity using deep learning techniques is given in Zheng et al. [29].

In auditory tasks, EEG recordings have shown to contain rhythm information in music perception using classifiers based on deep networks [30]. A recent work by Das et al. [31] has shown that auditory attention decoding in the perception of noisy speech can also be improved by deep learning techniques. In multi-speaker cocktail party scenarios, Deckers et al. [32] showed that neural networks are capable of identifying the attended speaker. A Convolutional Neural Network (CNN) based model for EEG-based speech stimulus reconstruction was also proposed by de Taillez et al. [33],

showing that deep learning is a feasible alternative to linear decoding methods. Our work on deep CCA models [34] for intra-subject analysis and inter-subject analysis [35] is extended in this paper with inter-subject models and additional evaluations on music dataset.

III. MATHEMATICAL BACKGROUND

A. Linear Canonical Correlation Analysis

For a dataset with pairs of multi-variates, linear Canonical Correlation Analysis [36] obtains an optimal linear transform for both the views such that the Pearson correlation of the two transformed vectors is maximized. Let $\mathbf{x} \in \mathcal{R}^{\mathcal{D}_1}$ and $\mathbf{y} \in \mathcal{R}^{\mathcal{D}_2}$ be two random vectors that represent two views of the data. Let d be the dimension of the projected subspace. The subspace is determined such that resultant projection vectors are maximally correlated. For example, if $d = 1$, then a pair of transform vectors $\mathbf{u}_1 \in \mathcal{R}^{\mathcal{D}_1}$ and $\mathbf{v}_1 \in \mathcal{R}^{\mathcal{D}_2}$ need to be determined such that $\hat{x} = \mathbf{u}_1^\top \mathbf{x}$ and $\hat{y} = \mathbf{v}_1^\top \mathbf{y}$ are maximally correlated. Mathematically,

$$\rho(\mathbf{u}_1, \mathbf{v}_1) = \frac{\mathbf{u}_1^\top \mathbf{C_{xy}} \mathbf{v}_1}{\sqrt{\mathbf{u}_1^\top \mathbf{C_{xx}} \mathbf{u}_1 \mathbf{v}_1^\top \mathbf{C_{yy}} \mathbf{v}_1}} \quad (1)$$

$$(\mathbf{u}_1^*, \mathbf{v}_1^*) = \arg\max_{\mathbf{u}_1, \mathbf{v}_1} \rho(\mathbf{u}_1, \mathbf{v}_1)$$

where, $\mathbf{C_{xx}}$, $\mathbf{C_{yy}}$ are the auto-correlation matrices of $\mathbf{x}$, $\mathbf{y}$ respectively while $\mathbf{C_{xy}} = \mathbf{E}[(\mathbf{x} - \mu_\mathbf{x})(\mathbf{y} - \mu_\mathbf{y})^\top]$ is the cross correlation matrix. Here, $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ are the mean vectors of $\mathbf{x}$ and $\mathbf{y}$ respectively.

It can be shown that the optimal solution to the transform vectors $\mathbf{u}_1^*$ and $\mathbf{v}_1^*$ is given by the first left and right singular vectors of the matrix

$$\mathbf{T} \triangleq \mathbf{C_{xx}}^{-1/2} \mathbf{C_{xy}} \mathbf{C_{yy}}^{-1/2} \quad (2)$$

respectively. For $d > 1$, the solution is obtained by finding the subsequent singular vectors of $\mathbf{T}$ [20].

B. Linear Multiway CCA

The multiway CCA is a linear method that generalizes the linear CCA to multiple (more than two) data-views. It finds a linear transform for each data-view, such that all the projections are maximally correlated to each other [18], [19].

Let $\mathbf{x}_n \in \mathcal{R}^{d_n}$, for $n = 1$ to N, denote the N data-views and $D_N = \sum_{n=1}^{N} d_n$. For the 1-D projection case, let $\mathbf{v}_n \in \mathcal{R}^{d_n}$ denote the transform vector that projects $\mathbf{x}_n$ onto the common subspace. The goal of MCCA is to find the transform vectors $\{\mathbf{v}_n\}_{n=1}^{N}$ such that the inter-set correlation (ISC) is maximized. The ISC is defined as

$$\rho_{\text{ISC}} = \frac{1}{N-1} \frac{r_B}{r_W} \quad (3)$$

where r_B is the between-set covariance and r_W is the within-set covariance. The between-set and within-set covariances are,

$$r_B = \sum_{j=1}^{N} \sum_{k=1, k \neq j}^{N} \mathbf{v}_j^\top \mathbf{R}^{jk} \mathbf{v}_k; \quad r_W = \sum_{j=1}^{N} \mathbf{v}_j^\top \mathbf{R}^{jj} \mathbf{v}_j$$

where $\mathbf{R}^{jk} \in \mathcal{R}^{d_j \times d_k}$ is the cross-covariance matrix between the views $\mathbf{x}_j$ and $\mathbf{x}_k$.

The cross-covariance matrices among all the views form elements of a block matrix $\mathbf{R} \in \mathcal{R}^{D_N \times D_N}$ such that $[\mathbf{R}]_{ij} = \mathbf{R}^{ij}$. By considering only the autocovariance matrices, a block-diagonal matrix $\mathbf{D} \in \mathcal{R}^{D_N \times D_N}$ is formed whose block diagonal entries are the same as that of $\mathbf{R}$. The optimal transform vectors $\{\mathbf{v}_n\}_{n=1}^N$ are obtained by solving the equation [18]:

$$\mathbf{Rv} = \lambda \mathbf{Dv} \qquad (4)$$

The eigenvector $\mathbf{v} \in \mathcal{R}^{D_N \times 1}$ with the maximum eigenvalue is the concatenation of the transform vectors $\{\mathbf{v}_n\}_{n=1}^N$.

For projection onto a higher dimensional subspace d (>1), the transform matrix for each multivariate $\mathbf{x}_n$ becomes $\mathbf{V}_n \in \mathcal{R}^{d_n \times d}$. This involves finding the top d eigenvectors of Equation (4).

C. Understanding CCA for EEG Decoding

The CCA model attempts to find a subspace of brain activity which is maximally correlated with the auditory stimulus. The EEG signal in the form of spatial channels (electrodes) and time-domain lags are used while the time-lagged audio envelopes are used as stimulus features. These two vectors form the data-views for CCA [13]. Specifically, the stimuli features, $\mathbf{x}$, in our experiments represent the time-lagged envelope of the audio signal while the response features, $\mathbf{y}$, represent the EEG recordings. These two features are provided at the same sampling rate and the CCA model attempts to relate them. This is done by finding the linear transforms, $\mathbf{u}_1^*$ and $\mathbf{v}_1^*$, that transform the stimulus and the response respectively, in such a way as to maximize the correlation.

In this case, the components of CCA can also be regarded as spatio-temporal filters applied on the EEG data and modulation filters on the audio envelope. The multiple CCA components of the audio signal correspond to FIR filtered envelopes that are orthogonal. Similarly, the CCA components of the EEG signal represent spatio-spectrally filtered projections that are orthogonal. The CCA was used recently for auditory and audio-visual EEG analysis by Dmochowski *et al.* [37]. The use of CCA in forward and backward models with time lags has shown additional improvements in correlation values [13].

When multiple subjects are presented with the same stimulus, the functional similarity is expected to generate similar responses [38]. However, the position or orientation of the sources with respect to the electrodes may be different for different subjects and therefore, the direct mapping of the EEG responses for the same stimulus from different subjects is cumbersome. The multi-CCA attempts to align the data from each subject to a common representation that makes it possible to compare across subjects. This is achieved by deriving spatial filters that are specific to each subject [39].

For the MCCA model, the N subjects' response features (EEG recordings), $\mathbf{x}_n$ for $n = 1$ to N and the common stimulus features (audio envelope) $\mathbf{x}_{N+1}$, are provided as inputs. Now, the model provides a linear transform for each

of them, $\mathbf{v}_n^*$ for $n = 1$ to $N+1$, such that the projected representations $(\mathbf{v}_n^{*\top} \mathbf{x}_n)$ are highly correlated to each other.

IV. PROPOSED ANALYSIS APPROACH

A. Deep CCA

A deep CCA model finds a pair of optimal non-linear transforms for the two views of the dataset through a pair of neural networks such that the two new projections are highly correlated [20]. As before, let random vectors $\mathbf{x} \in \mathcal{R}^{\mathcal{D}_1}$ and $\mathbf{y} \in \mathcal{R}^{\mathcal{D}_2}$, denote the data-views. Let $f_1(\cdot)$ and $f_2(\cdot)$ denote the non-linear functions realized by the neural networks operating on $\mathbf{x}$ and $\mathbf{y}$ respectively. Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent the trainable parameters of f_1 and f_2 respectively. We find their optimal values by solving the following optimization problem:

$$(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) = \underset{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\arg\max} \; \rho \left(f_1\left(\mathbf{x}; \boldsymbol{\theta}_1\right), f_2\left(\mathbf{y}; \boldsymbol{\theta}_2\right) \right) \qquad (5)$$

where ρ corresponds to the cross correlation coefficient.

For a batch of m examples from each of the $(\mathbf{x}, \mathbf{y})$ data-views, let $\mathbf{H_x} \in \mathbb{R}^{d \times m}$ denote the matrix whose columns are the neural network outputs $f_1(\cdot)$. Similarly, let $\mathbf{H_y} \in \mathbb{R}^{d \times m}$ denote the outputs from the second network $f_2(\cdot)$.

Let, $\bar{\mathbf{H}}_x = \mathbf{H_x} - \frac{1}{m}\mathbf{H_x}\mathbf{1}$ and similarly, $\bar{\mathbf{H}}_y = \mathbf{H_y} - \frac{1}{m}\mathbf{H_y}\mathbf{1}$ denote the centred data matrices, where $\mathbf{1}$ is an all-1 matrix of dimension $m \times m$. The covariance matrices of the feed-forward network outputs are given by,

$$\hat{\mathbf{C}_{xx}} = \frac{1}{m-1}\bar{\mathbf{H}}_x\bar{\mathbf{H}}_x{}^\top; \quad \hat{\mathbf{C}_{yy}} = \frac{1}{m-1}\bar{\mathbf{H}}_y\bar{\mathbf{H}}_y{}^\top$$
$$\hat{\mathbf{C}_{xy}} = \frac{1}{m-1}\bar{\mathbf{H}}_x\bar{\mathbf{H}}_y{}^\top$$

Further, let

$$\mathbf{T}_H \triangleq \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\hat{\mathbf{C}}_{\mathbf{xy}}\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2} \qquad (6)$$

Let $\mathbf{T}_H = \mathbf{UDV}^\top$ denote the SVD of $\mathbf{T}_H$. It can be shown that optimization[1] of Equation (5) is given by,

$$\frac{\partial \rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial \mathbf{H_x}} = \frac{1}{m-1}\left(2\nabla_{\mathbf{xx}}\bar{\mathbf{H}}_x + \nabla_{\mathbf{xy}}\bar{\mathbf{H}}_y\right) \qquad (7)$$

where

$$\nabla_{\mathbf{xy}} = \hat{\mathbf{C}_{xx}}^{-1/2}\mathbf{U}\mathbf{V}^\top\hat{\mathbf{C}_{yy}}^{-1/2} \qquad (8)$$
$$\nabla_{\mathbf{xx}} = -\frac{1}{2}\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{U}\mathbf{D}\mathbf{U}^\top\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2} \qquad (9)$$

Similar expression can be obtained for gradient with respect to $\mathbf{H_y}$. These gradients are backpropagated to learn the optimal model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Note that the gradient ascent, for $j = 1, 2$, is performed as

$$\boldsymbol{\theta}_j{}^{t+1} = \boldsymbol{\theta}_j{}^t + \eta \frac{\partial \rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial \boldsymbol{\theta}_j{}^t} \qquad (10)$$

where the η is the learning rate for the gradient ascent.

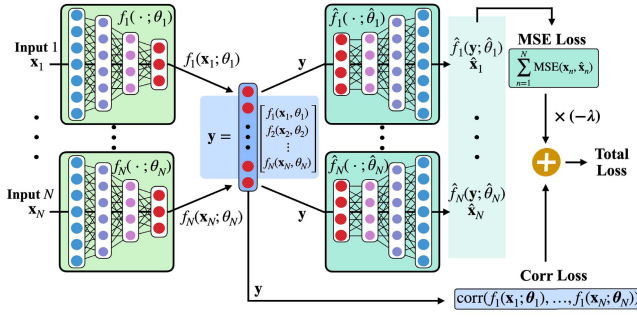[1]A detailed derivation of the gradients is given in Appendix A.

Fig. 1. The proposed deep multiway CCA Model. N inputs are provided to N encoders. All N outputs are provided to the correlation loss and all N decoders. The decoders' outputs are provided to the reconstruction (MSE) loss. The sum of the correlation loss and negative of the reconstruction loss becomes the cost function.

B. Deep Multiway CCA

For each of the N data-views $\{\mathbf{x}_n \in \mathcal{R}^{d_n}\}_{n=1}^N$, the goal of deep MCCA is to derive optimal non-linear transforms such that the transformed vectors are highly correlated. Let $f_n(\cdot)$, for $n = 1$ to N, represent neural networks with trainable parameters $\boldsymbol{\theta}_n$ operating on $\mathbf{x}_n$.

The N neural networks are trained to maximize the inter-set correlations defined as:

$$\rho_{\text{Total}} = \sum_{j=1}^N \sum_{k=1, k \neq j}^N \rho\left(f_j\left(\mathbf{x}_j; \boldsymbol{\theta}_j\right), f_k\left(\mathbf{x}_k; \boldsymbol{\theta}_k\right)\right) \quad (11)$$

Comparing with Equation (3), the correlation cost here (ρ_{Total}) is the summation of pairwise correlation coefficients. While the two definitions are related, the cost based on sum of pairwise correlations is more suitable for gradient based optimization. The parameters are obtained as:

$$\left(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*\right) = \underset{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)}{\text{argmax}} \; \rho_{\text{Total}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \quad (12)$$

The backpropagation for each network is similar to the deep CCA model, as described in Equation (7).

The proposed model (shown in Figure 1) has multiple autoencoders sharing encoded representations (N autoencoders for N dataviews respectively). Each view $\mathbf{x}_n$ forward propagates through the encoder part of its autoencoder, $f_n(\boldsymbol{\theta}_n, \cdot)$. All the encoded representations, $f_n(\mathbf{x}_n; \boldsymbol{\theta}_n)$ are concatenated (denoted as $\mathbf{y}$), and propagated to the decoders, $\hat{f}_n(\hat{\boldsymbol{\theta}}_n, \cdot)$. This shared encoder-decoder model allows the learning of data-view specific transforms that align the views.

The model is trained to maximize the joint cost function of correlation and negative of the mean square error (MSE) in reconstruction. This cost function is given as,

$$E = \rho_{\text{Total}} - \lambda \sum_{n=1}^N \text{MSE}\left(\mathbf{x}_n, \hat{f}_n(\mathbf{y}; \hat{\boldsymbol{\theta}}_n)\right) \quad (13)$$

where ρ_{Total} is defined by Equation (11) and MSE$(\cdot)$ is the average squared reconstruction loss. The parameter λ controls the trade-off between maximizing the correlation metric and minimizing the MSE in learning the model parameters.

The model is trained using the N data-views $\{\mathbf{x}_n\}_{n=1}^N$ with the cost metric defined. Note that the correlation loss is

independent of the decoder parameters $\hat{\boldsymbol{\theta}}_n$ while the MSE$(\cdot)$ is a function of both the encoder parameters $\boldsymbol{\theta}_n$ and decoder parameters $\hat{\boldsymbol{\theta}}_n$. Once the model is trained, each data-view $\mathbf{x}_n$ is projected using the encoder $f_n(\mathbf{x}_n; \boldsymbol{\theta}_n)$.

V. AUDIO-EEG SETUP

A. Datasets

We experiment our methods on two datasets. The first one is a dataset of EEG responses for speech stimuli, recorded by Liberto et al. [10]. The second dataset is NMED-H [22]. It contains EEG recordings for a music listening task.

1) Speech-EEG Dataset: The EEG recordings were collected using 128 channels, when the subjects were listening to a male speaker reading snippets of a novel.[2] A Biosemi system, sampled at 512 Hz, was used to collect the EEG data. We perform the same preprocessing steps as described in Cheveigné et al. [13]. Specifically, the EEG data are downsampled to 64 Hz and processed using noise suppression software [40]. A band-pass filtering with a passband in the range of $0.1-12$ Hz is applied to the EEG data. At the stimulus side, the speech envelopes sampled at $44, 100$ Hz, are squared and smoothed by a convolution with a square window. Finally, the stimuli data are downsampled to 64 Hz with a cubic-root compression.

2) Music-EEG Dataset: The NMED-H [41] is an open source dataset containing EEG responses to naturalistic music - 4 versions (normal, time-reversed, phase-scrambled, and shuffled) of 4 full-length "Bollywood" songs, each approximately of 4.5 minutes long. The last three stimuli versions were chosen to manipulate the temporal features at varying degrees, while the aggregate frequency content of each stimulus was same. The shuffled version imposes minimal temporal disruption whereas the phase-scrambled versions were considerably distorted [41].

In the phase-scrambled subset, three stimuli files were not considered in the analysis as the features had discontinuities. The EEG recordings were recorded from 125 electrodes at 1 kHz. Each recording is filtered between 0.3-50 Hz and downsampled to 125 Hz. We use the "Clean EEG" recordings which are cleaned and aggregated on a per-stimulus, per-listen basis. More details on data acquisition and preprocessing are given in Kaneshiro et al. [42], Kaneshiro [41].

The stimuli features are extracted as described in Gang et al. [43]. The acoustic features are extracted using the music information retrieval (MIR) toolbox, Version 1.7.2 [44]. From the stimuli, 20 features are extracted in 25ms analysis windows with a 50% overlap between frames [45], [46]. The 20 features are discussed in the Section IX-B as an appendix. The principal component analysis (PCA) is performed to obtain a 1D representation (PC1) on these 20 extracted features. The two individual features, root mean square (RMS) and spectral flux, along with the PC1 are chosen to obtain a 3D representation for the stimuli. The EEG responses are re-sampled to the sampling rate of the acoustic features (80 Hz).

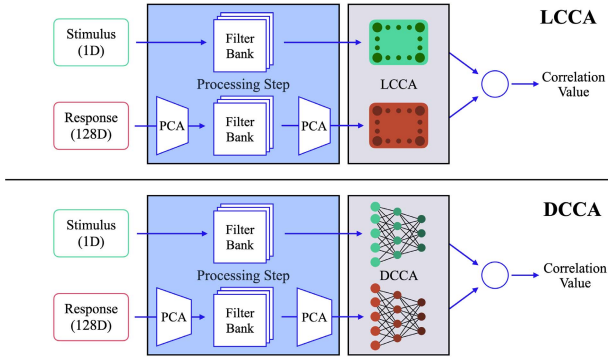[2]The speech-EEG dataset is available open source at https://datadryad.org/stash/dataset/doi:10.5061/dryad.070jc

Fig. 2.   The LCCA and DCCA methods. The stimulus features are low-pass filtered envelopes.
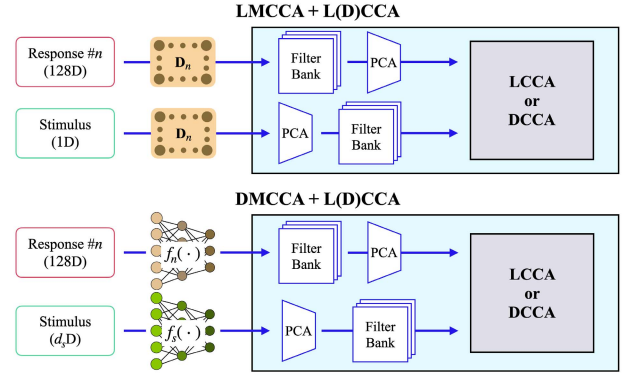

Fig. 3.   Configuration of the four systems - linear multiway CCA with linear CCA (LMLC), linear multiway CCA with deep CCA (LMDC), deep multiway CCA with linear CCA (DMLC) and deep multiway CCA with deep CCA (DMDC) methods.

B. CCA Methods

In all our experiments, the linear CCA (LCCA) [13] and linear MCCA (LMCCA) [19] analysis act as the baseline setup for comparison with the deep CCA (DCCA) and the deep multiway CCA (DMCCA) methods. For the multi-subject EEG analysis, the outputs from the multiway CCA are further processed with CCA (either LCCA or DCCA).

1) LCCA: On the 1D preprocessed stimuli data, a dyadic bank of 21 FIR band-pass filters is applied that contains filters that are approximately uniformly distributed on a logarithmic scale [13]. At the response end, a PCA is applied to the EEG data that transforms the 128D (or 125D for music data) EEG data to 60D. The filterbank is applied on these 60D EEG data to yield 1260D data. A second PCA is applied that transforms them to 139D subspace. Now, the 21D stimuli data and the 139D EEG data are projected onto common subspace using CCA transforms. The data are processed (the choice of PCA and the dimensions after each stage) as proposed by Cheveigné *et al.* [13]. The combination of PCA and filterbank acts as a spatio-temporal filter on the data.

2) DCCA: The neural networks used in DCCA model have a 2 hidden layer architecture, for each view, with 2038 and 1608 units for the first and second layers respectively followed by a d dimensional output layer. The data are processed similar to the LCCA method, and the final 21D and 139D representations are input to a deep CCA model. Figure 2 describes the LCCA and DCCA methods.

3) LMCCA: The preprocessed EEG responses from N subjects and a time-lagged version of their common stimuli (d_sD), are provided to a linear MCCA model to obtain the denoised representations for each subject's EEG response. Now, each subject's denoised EEG data and their corresponding stimuli can be provided, separately, to the LCCA and DCCA methods to obtain final representations. This is performed as proposed in Cheveigné *et al.* [19].

4) DMCCA: The preprocessed EEG responses, along with the common stimuli are provided to a deep MCCA model to obtain a dD denoised representation for each EEG response.

The architecture of the DMCCA model is shown in Figure 1. The encoder has two hidden layers of 60 units each and an output layer of d units. The decoding part has two hidden layers of 60 and 110 units respectively.

The d_s and d are hyperparameters and the best values are selected by varying them, for both the variants of MCCA. More details are discussed in Section VII-A. The outputs from the linear MCCA model are of 128D for speech (125D for music) dataset. For both the MCCA methods, the denoised responses are provided to the filterbank followed by a PCA to generate 139D vectors. The dD stimuli obtained are projected onto a 1D subspace using PCA, followed by the filterbank resulting in a 21D data. These steps make sure that the inputs to the CCA transforms are equivalent in both versions of MCCA.

For intra-subject analysis, the LCCA/DCCA are performed on each subject's data separately. For inter-subject analysis, the LMCCA/DMCCA are performed on data from multiple subjects data followed by a subject-specific LCCA/DCCA method. Thus, we have four combinations

1) **LMLC**: LMCCA + LCCA
2) **LMDC**: LMCCA + DCCA
3) **DMLC**: DMCCA + LCCA
4) **DMDC**: DMCCA + DCCA

Figure 3 shows the four combinations of the MCCA denoising followed by CCA analysis for each subject.

C. Experimental Setup

For the speech dataset, the methods are performed on the preprocessed 1D stimuli envelopes and 128D EEG responses. For NMED-H, along with the 1D stimuli envelopes, each dimension of the 3D preprocessed stimuli is also used with the 125D clean EEG recordings.

From the speech dataset, stimulus-response data of 8 subjects are considered randomly to perform the experiments. The data from each subject contains 20 sessions with approximately 160s of recording in each session. For all the methods, 20 fold validation experiments are performed where 18 sessions are used for training, one session for validation and the remaining session for testing. Given a sampling rate of 64 Hz, the approximate number of instances for the linear/deep model training per subject is about 200k.

The NMED-H dataset contains recordings from 48 subjects and 16 stimuli. The subjects were divided into 16 groups of

12 subjects with each subject appearing in 4 groups. Each group is presented with 2 trials of a stimulus which results in each subject listening to 2 trials of 4 different stimuli. In our analysis, we have used all the data available. All the 12 subjects that were available for each stimulus were used in the inter-subject analyses and the intra-subject analysis.

We split the data into $90 - 5 - 5$ for training, validation and test respectively. It results in about 155k samples for training and 8.5k samples for testing and validation, for each subject in the CCA experiments. Similarly, we use 38k samples for training and 2k samples for testing and validation, for each subject per stimulus in the MCCA experiments.

A leaky ReLU activation function with a negative slope coefficient of 0.1 is used in the DCCA model and the encoder part of the DMCCA model. A linear activation function is used at the output layer of the decoder sections in the DMCCA model.[3] Further, dropout regularization [21], [47] is incorporated in the deep models training to avoid over-fitting in the noisy conditions.

1) Performance Metric: The primary metric used is the Pearson correlation between the transformed EEG and audio signals on the held-out test set. For each subject, the LCCA and DCCA methods performance is measured by the correlation coefficient (ρ) of the two final representations. The LMCCA tries to maximize the ρ_{ISC}, and the DMCCA tries to maximize the ρ_{Total}. For overall results, instead of direct averaging of Pearson correlation values which is mathematically incorrect, we perform a z-score based averaging implemented in the Statsoft software [48].

We also use a secondary performance metric based on classification of aligned versus misaligned EEG-audio segments [13]. Here, fixed-length segments of EEG and audio signals that are randomly located are correlated using the models. If the audio and EEG segments are aligned, the model is expected to generate a higher correlation score than when the two signals are misaligned. The correlation scores are analyzed using the sensitivity index (Cohen's d' statistic).

Let the means of the matched and mismatched segments' correlation coefficients be μ_1 and μ_2 respectively. Let, σ_1^2 and σ_2^2 be their respective variances. The Cohen's d' is:

$$d' = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}} \tag{14}$$

VI. RESULTS

The results comparing the linear and deep CCA models for intra-subject and inter-subject experiments on the speech-EEG dataset are given in Table I and Table II respectively. The results for the music-EEG dataset are shown in Table III and Table IV. Pairwise one-tailed t-tests are performed on the pairs of LCCA-DCCA (Table I and III) and LMLC-DMDC (Table II and IV).

[3]The implementation is available at https://github.com/iiscleap/deep-cca-for-audio-EEG

TABLE I
PEARSON CORRELATION VALUES FOR LCCA AND DCCA METHODS ON THE SPEECH - EEG DATASET. A PAIRED T-TEST WITH {P-VALUE} AND [T-VALUE] ARE INDICATED. STATISTICALLY SIGNIFICANT ($p < 0.025$) FOR 4 OUT OF 8 SUBJECTS AND THE OVERALL AGGREGATE

Models	LCCA	DCCA	t-test
Sub1	0.220	**0.275**	{0.01}[2.3]
Sub2	0.258	**0.316**	{0.01}[2.5]
Sub3	0.175	**0.213**	{0.051}[1.7]
Sub4	0.316	**0.403**	{1e-3}[3.3]
Sub5	0.307	**0.338**	{0.03}[1.9]
Sub6	0.315	**0.354**	{0.049}[1.7]
Sub7	0.260	**0.292**	{0.06}[1.5]
Sub8	0.183	**0.232**	{0.02}[2.2]
Overall	0.255	**0.304**	{1e-6}[4.8]

TABLE II
PEARSON CORRELATION VALUES FOR LMLC, LMDC, DMLC AND DMDC. A STATISTICAL SIGNIFICANCE TEST (T-TEST) BETWEEN LMLC AND DMDC ({P-VALUE} [T-VALUE]) IS ALSO REPORTED. THE IMPROVEMENTS ARE STATISTICALLY SIGNIFICANT ($p < 0.025$) FOR 5 OUT OF 8 SUBJECTS AND THE OVERALL AGGREGATE

Models	LMLC	LMDC	DMLC	DMDC	t-test
Sub1	0.262	0.271	0.375	**0.377**	{9.1e-7}[5.6]
Sub2	0.289	0.325	0.367	**0.374**	{5.1e-4}[3.6]
Sub3	0.160	0.177	0.258	**0.259**	{6.3e-5}[4.2]
Sub4	0.310	**0.378**	0.341	0.361	{3.6e-2}[1.8]
Sub5	0.309	0.354	0.389	**0.392**	{8.5e-5}[4.2]
Sub6	0.327	0.342	0.416	**0.420**	{4.6e-5}[4.4]
Sub7	0.275	0.289	**0.310**	0.310	{4.4e-2}[1.7]
Sub8	0.221	0.245	0.259	**0.272**	{2.8e-2}[2.0]
Overall	0.270	0.299	0.339	**0.344**	{9e-14}[7.7]

A. Speech-EEG Dataset Results

For speech-EEG experiments, the 20 cross-validation results (correlation values) for the 20 folds for all the subjects are considered for a pairwise t-test. As seen in Table I, the DCCA improves over the LCCA for all the subjects. The average absolute improvements for DCCA over the LCCA in terms of correlation value is 5%. The improvements are also statistically significant ($p < 0.025$)[4] for 4 out of 8 subjects and the overall aggregate.

The comparison of various inter-subject experiments on the speech-EEG dataset is shown in Table II. Here, the inter-subject alignment using linear method (LMLC) improved over the linear intra-subject correlation model (LCCA) on all subjects except subject 3 and 4. The inter-subject alignment using deep learning (DMLC/DMDC) improves the correlation scores compared to the intra-subject scores reported in Table I for all cases except subject 4. Further, the deep models consistently improve over the linear counterparts. In particular, the deep multiway CCA approach improves over the linear multiway CCA by an absolute correlation value of 7.4% on the average. The improvements are found to be statistically significant ($p < 0.025$) for 5 out of 8 subjects. The overall aggregate score is found to be statistically significant as well.

[4]Compensating for multiple comparisons on two datasets, the significance threshold ($\alpha = 0.05$) is divided by 2 to obtain a threshold of 0.025 as per the Bonferroni correction method [49].
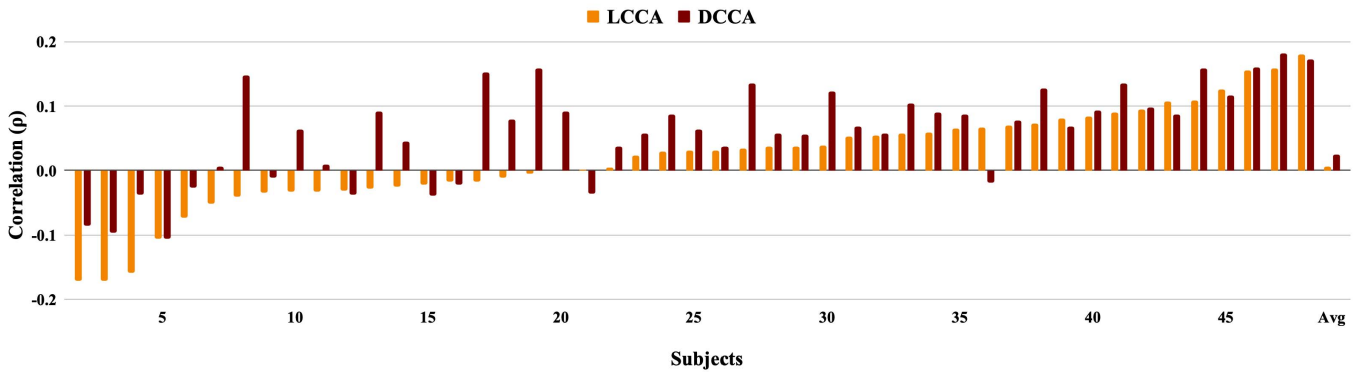
Fig. 4. Comparison between LCCA and DCCA methods for PC1 stimuli features, of the 48 subjects from the NMED-H dataset, arranged in the increasing order of the LCCA correlation values. The last column shows the average of the 48 subjects.

TABLE III

AVERAGE CORRELATION VALUES FOR 48 SUBJECTS FROM THE NMED-H DATASET IN INTER-SUBJECT ANALYSIS. A STATISTICAL SIGNIFICANCE TEST (T-TEST) BETWEEN LMLC AND DMDC METHODS (INDICATED AS {P-VALUE} [T-VALUE]) IS ALSO REPORTED. THE IMPROVEMENTS ARE STATISTICALLY SIGNIFICANT ($p < 0.025$) FOR ALL EXPERIMENTS

Stimulus feature	Normal			Reversed			Phase-Scrambled			Shuffled		
CCA Model	LCCA	DCCA	[t-test]	LCCA	DCCA	[t-test]	LCCA	DCCA	[t-test]	LCCA	DCCA	[t-test]
Envelope	0.007	**0.118**	{1e-4}[3.7]	-0.003	**0.117**	{3e-5}[4.1]	-0.052	**0.095**	{4e-7}[5.3]	-0.013	**0.134**	{3e-5}[4.2]
PC1	-0.020	**0.077**	{9e-4}[3.2]	0.012	**0.105**	{1e-3}[3.0]	-0.016	**0.072**	{1e-3}[3.1]	0.030	**0.135**	{1e-3}[3.0]
RMS	-0.004	**0.087**	{2e-4}[3.6]	0.008	**0.101**	{3e-4}[3.5]	-0.042	**0.091**	{4e-7}[5.2]	-0.025	**0.100**	{3e-5}[4.2]
Spectral Flux	0.008	**0.102**	{3e-4}[3.5]	-0.004	**0.113**	{9e-6}[4.5]	-0.034	**0.107**	{3e-7}[5.3]	0.005	**0.123**	{3e-5}[4.2]

TABLE IV

AVERAGE CORRELATION VALUES FOR 48 SUBJECTS FROM THE NMED-H DATASET IN INTER-SUBJECT ANALYSIS. A STATISTICAL SIGNIFICANCE TEST (T-TEST) BETWEEN LMLC AND DMDC METHODS (INDICATED AS {P-VALUE} [T-VALUE]) IS ALSO REPORTED. THE IMPROVEMENTS ARE STATISTICALLY SIGNIFICANT ($p < 0.025$) FOR ALL EXPERIMENTS

Stimulus feature	Normal					Reversed				
MCCA Model	LMLC	LMDC	DMLC	DMDC	t-test	LMLC	LMDC	DMLC	DMDC	t-test
Envelope	0.076	0.146	0.344	**0.349**	{3e-26}[14]	0.062	0.099	0.299	**0.384**	{1e-37}[21]
PC1	-0.007	0.102	**0.384**	0.321	{1e-26}[14]	0.030	0.159	**0.360**	0.323	{2e-19}[11]
RMS	0.001	0.114	**0.341**	0.246	{3e-13}[8.5]	0.042	0.135	**0.318**	0.220	{1e-08}[6.2]
Spectral Flux	0.017	0.110	0.341	**0.343**	{2e-24}[13]	0.053	0.170	**0.340**	0.321	{1e-16}[10]

Stimulus feature	Phase-Scrambled					Shuffled				
MCCA Model	LMLC	LMDC	DMLC	DMDC	t-test	LMLC	LMDC	DMLC	DMDC	t-test
Envelope	0.042	0.092	**0.312**	0.299	{8e-26}[14]	0.077	0.132	**0.341**	0.333	{4e-19}[11]
PC1	0.012	0.166	0.262	**0.389**	{6e-21}[12]	0.051	0.149	0.345	**0.347**	{5e-21}[12]
RMS	0.020	0.108	0.176	**0.397**	{2e-07}[7.4]	0.051	0.156	0.327	**0.345**	{1e-19}[11]
Spectral Flux	0.038	0.207	0.340	**0.390**	{3e-22}[12]	0.061	0.145	0.294	**0.322**	{8e-15}[9.2]

B. Music-EEG Dataset Results

For LCCA/DCCA methods, the average correlation values for the 48 subjects in the NMED-H dataset is reported in Table III. The results are reported for different music conditions - normal, shuffled, time-reversed and phase-scrambled; and stimuli features - envelope, PC1, RMS and spectral flux. The pair-wise t-test on the NMED-H dataset shows that all improvements obtained by the deep versions are statistically significant ($p < 0.025$). The performance of LCCA and DCCA methods on the PC1 features of 48 subjects from the NMED-H dataset is shown in Figure 6. The average absolute improvements are about 11% for the DCCA over LCCA method. For inter-subject analysis, the Table IV shows that the DMDC improves over the LMLC method with an average absolute improvement of 29.3%.

C. Statistical Analysis

In order to measure the significance of the improved correlations of our proposed deep models over the baseline system, we have performed two statistical tests: a) one-tailed pairwise t-test and b) Cohen's d'. The pairwise t-test is performed to objectively quantify the difference in the distribution of the correlation scores from the two methods. Given that the same hypothesis (LCCA versus DCCA on intra-subject analysis or LMLC versus DMDC on inter-subject analysis) is tested on two different datasets (speech-EEG and music-EEG), a compensation is required for multiple comparisons. We use the Bonferroni correction [49]. Thus, a p-value $0.05/2 = 0.025$ is used to check if the improvements in the correlation are statistically significant on each dataset. The pairwise t-test results comparing the linear and deep models are reported for speech-EEG (Table I, II) and music-EEG (Table III, IV).
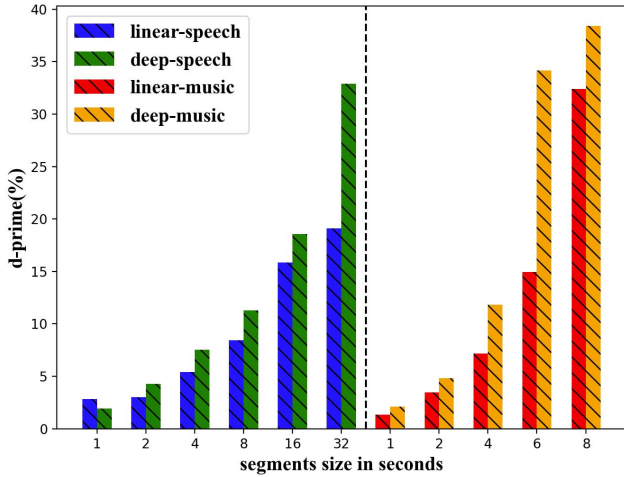
Fig. 5. The d' metric for both the datasets. linear-speech and linear-music correspond to the d' values of the final representations from LMLC models for the speech and music datasets. Similarly, deep-speech and deep-music correspond to DMLC models.

As mentioned in Section V-C, a classification metric is also performed where audio-EEG segments are classified as aligned or misaligned based on the Pearson correlation measure. The second statistical test, Cohen's d' [50], is an effect size used to indicate the standardised difference between two classes (in our case, these classes are aligned and mis-aligned audio-EEG pairs). The d' metric quantifies the model's ability to match the corresponding stimulus-response pair based on the correlation value, ρ. The test data is divided into N segments of t seconds each, and the correlation coefficient values ρ are calculated for the linear the deep methods for both aligned and mis-aligned segments (speech/audio-EEG pairs). The Cohen's d' metric measures the model's ability to separate aligned versus mis-aligned pairs. The LMLC/DMLC methods are used on the audio-EEG segments and the correlation values are computed. Using the correlation score from the respective models, the Cohen's d' is computed on the correlation score. The d' statistics are presented in Figure 5. This is performed separately for the speech-EEG and music-EEG datasets. As seen in Figure 5, the deep model improves over the linear model in all the cases except for 1 second segments in speech-EEG data. In longer segments, considerable improvements in the d' statistic are observed for the deep models.

VII. DISCUSSION

A. Impact of Hyperparameters

In this section, we analyze the impact of the hyperparameters involved in the deep CCA/MCCA models on the correlation metric. The parameters analyzed are: model architecture, dropout percentage, regularization parameter in DMCCA, and the number of output dimensions in DMCCA. We use a learning rate of $1e-3$ and a batch size of 2048 on experiments where these parameters are not mentioned. Further, the number of time-lags used in the stimulus input is also varied to understand its impact. For initializing the models, we start with multiple random seeds and choose the one which gives

the best correlation on the validation data before the model training. Unless specified otherwise, the values of parameters d_s and λ are fixed to 60 and 0.1. The value of d is fixed to 1 and 10 for all the DCCA and DMCCA models respectively.

1) Dropout: For the speech-EEG dataset, we experiment with dropout percentage from $0-20\%$ in the DCCA/DMCCA model. The correlation values obtained by DMLC and DCCA methods are shown in Figure 6 (A). When there is no dropout, there is a tendency for the model to overfit. A similar effect is seen in the DCCA model as well.

2) Batchsize: The effect of the batch size is analyzed for the DCCA model. The average correlation values of 6 subjects from the speech dataset, for 20 cross-validation trials is reported in Figure 6 (B). Given the noisy nature of the data, we find that the higher batchsizes (compared to typical choices of few hundreds in supervised classification setting) are found to improve the final correlation value. The optimal batch size on the validation data is 2048.

3) DCCA Output Dimension: Just like the CCA model where multiple canonical components dimensions can be derived from the data, the DCCA model also can be trained for multiple output dimensions. The comparison of the linear and deep CCA models for 5 canonical correlation components is shown in Figure 6 (C) for each subject in the speech-EEG dataset. As seen here, the DCCA model improves over the linear CCA model consistently for all the subjects.

4) DMCCA Output Dimension: We have varied the number of output dimensions in the DMCCA model from 10 to 128. The performance, average of the final correlation values of all the subjects, of the DMLC model is presented in the Figure 6 (D) on the speech-EEG dataset. The best performance is achieved for DMCCA outputs of 10 dimensions.

5) Context Size of Stimuli: The time-lag applied to the stimulus (d_s) in the DMCCA model also plays an important role. It is varied from 10 to 110. For 6 subjects from the speech-EEG dataset, the DMLC model is trained and tested and the average correlation of all the subjects is shown in the Figure 6 (E). A stimulus time-lag of 80 gives the best performance.

6) Regularization Parameter: The DMCCA model incorporates the MSE regularization loss in addition to the inter-set correlation loss. The effect of the regularization parameter (λ) is studied in Figure 6 (F), by varying it from 0 to 1000. The results show that the regularization with $\lambda = 10$ gives the best performance.

B. Model Architecture

We also experimented with various architecture choices for the DCCA model. The experiments with varying the number of hidden layers (L) from 2 to 5 and number of units (n_L) in each layer is shown in Figure 7. As seen in this plot, increasing the number of layers degrades the correlation, as the model tends to over-fit. This trend may also be attributed to the lack of sufficient audio-EEG data for each subject. We hypothesize that, with more training data, the deeper models may further improve over linear models as well as the shallow models.
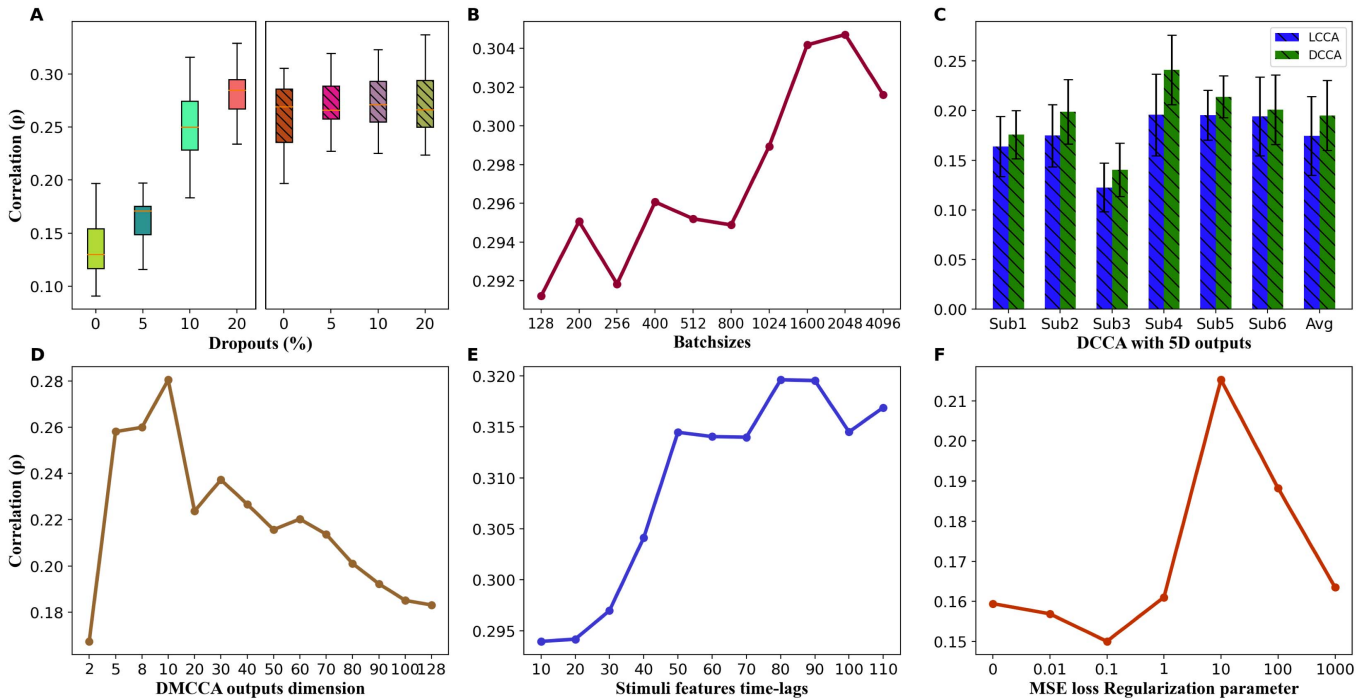
Fig. 6.   **A**. Effect of Dropout percentage on the final correlations of DMLC method and DCCA method. **B**. Effect of varying batchsizes on a DCCA model with output as 1D. **C**. Comparison between the correlation per dimension of the final representations from linear and deep CCA models with outputs of 5D. **D**. Comparing the performance of DMLC with varying output dimensionality **E**. Impact of choice of time-lags for the stimulus, given as input to the DMLC method. **F**. Effect of the MSE regularization parameter in the DMLC method.
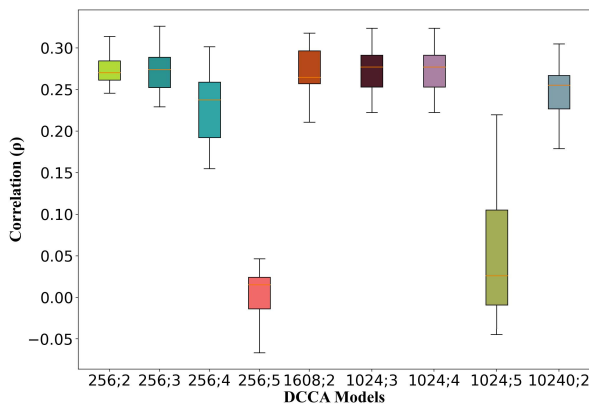

Fig. 7.   Different architectures explored in DCCA models. The x-axis denotes "number of units per layer; number of layers."

C. Impact of Improved Correlations

The EEG recordings capture various unrelated brain processes along with the response to the stimuli. Thus, only a fraction of the variance in the EEG signal can be explained by its stimulus. This results in low correlation values for many of the linear methods proposed in the past. In this paper, we have explored the application of deep models whereby consistent improvements in correlations are illustrated. Many applications based on EEG would benefit significantly with the improvements in stimulus-response correlations. For example, the improved correlations will help the EEG enabled auditory assistance device (e.g. hearing aid) as suggested by Cheveigné *et al.* [13]. In music information retrieval, the

performance improvement in EEG decoding systems will enable the understanding of the perceptual attributes of music. Throughout the study, we have pursued simple features like envelope. However, audio signals are encoded in several other dimensions like pitch, rhythm, zero-crossings, phase, semantic/ linguistic features etc. The exploration of the model with additional features may further throw light on the encoding of these dimensions in the EEG signals. Further, the techniques proposed in this work would be applicable to other brain signals like MEG, ECoG and fMRI signals as well.

VIII. SUMMARY

In this paper, we have proposed extensions to models that uncover the stimulus-response relationships between auditory signals like speech and music and their EEG responses. The models advance linear correlation methods and are proposed for improving single trial analyses. The models pose the problem of finding the optimal transforms that need to be applied to the stimulus and response in a deep learning framework which enables the learning of these transforms using established optimization methods. Using the proposed deep models, we show that the correlations between the stimulus and the response can be significantly improved over the linear methods. Further, the applicability of the proposed models is separately analyzed for speech and music EEG tasks.

APPENDIX A
DERIVATION OF THE DEEP CCA MODEL GRADIENTS

The matrix $\mathbf{T}_H$ is defined as $\mathbf{T}_H \triangleq \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2} \hat{\mathbf{C}}_{\mathbf{xy}} \hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}$ and its SVD decomposition is denoted as $\mathbf{T}_H = \mathbf{UDV}^\top$. The objective

function that needs to be maximized is

$$\rho\left(\mathbf{H_x}, \mathbf{H_y}\right) = \|\mathbf{T}_H\|_{\mathrm{tr}} = \mathrm{Trace}\left(\mathbf{T}_H^\top \mathbf{T}_H\right)^{1/2}$$

where $\|\cdot\|_{\mathrm{tr}}$ denotes the trace norm of the matrix.

We can derive the partial derivative of the objective function with respect to the matrix $\hat{\mathbf{C}}_{\mathbf{xx}}$, $\nabla_{\mathbf{xx}}$, as following:

$$[\nabla_{\mathbf{xx}}]^{ij} = \frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}\right]^{ij}}$$

$$= \sum_{kl}\frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial\left[\mathbf{T}_H^\top \mathbf{T}_H\right]^{kl}} \cdot \frac{\partial\left[\mathbf{T}_H^\top \mathbf{T}_H\right]^{kl}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{ij}}$$

$$= \sum_{kl}\left[\frac{1}{2}\left(\mathbf{T}_H^\top \mathbf{T}_H\right)^{-1/2}\right]^{kl} \cdot \frac{\partial\left[\mathbf{T}_H^\top \mathbf{T}_H\right]^{kl}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{ij}} \quad (15)$$

where,

$$\frac{\partial\left[\mathbf{T}_H^\top \mathbf{T}_H\right]^{kl}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{ij}}$$

$$= \sum_{pq}\frac{\partial\left[\mathbf{T}_H^\top \mathbf{T}_H\right]^{kl}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{pq}}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{pq}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}\right]^{ij}}$$

$$= \sum_{pq}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{C}}_{\mathbf{yx}}\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\hat{\mathbf{C}}_{\mathbf{xy}}\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\right]^{kl}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{pq}}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{pq}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}\right]^{ij}}$$

$$= -\sum_{pq}\left[\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{C}}_{\mathbf{yx}}\right]^{kp}\left[\hat{\mathbf{C}}_{\mathbf{xy}}\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\right]^{ql}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{pi}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{jq}$$

$$= -\left[\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{C}}_{\mathbf{yx}}\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\right]^{ki}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1}\hat{\mathbf{C}}_{\mathbf{xy}}\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\right]^{jl}$$

$$= -\left[\mathbf{T}_H^\top \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{ki}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{T}_H\right]^{jl} \quad (16)$$

Now, Equation (15) can be rewritten as:

$$[\nabla_{\mathbf{xx}}]^{ij}$$

$$= -\frac{1}{2}\sum_{kl}\left[\mathbf{T}_H^\top \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{ki}\left[\left(\mathbf{T}_H^\top \mathbf{T}_H\right)^{-1/2}\right]^{kl}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{T}_H\right]^{jl}$$

$$= -\frac{1}{2}\sum_{kl}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{T}_H\right]^{ik}\left[\left(\mathbf{T}_H^\top \mathbf{T}_H\right)^{-1/2}\right]^{kl}\left[\mathbf{T}_H^\top \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{lj}$$

$$= -\frac{1}{2}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{T}_H\left(\mathbf{T}_H^\top \mathbf{T}_H\right)^{-1/2}\mathbf{T}_H^\top \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{ij}$$

$$= -\frac{1}{2}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{U}\mathbf{D}\mathbf{V}^\top\left(\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^\top\right)\mathbf{V}\mathbf{D}\mathbf{U}^\top\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{ij}$$

$$= -\frac{1}{2}\left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{U}\mathbf{D}\mathbf{U}^\top\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{ij} \quad (17)$$

Similarly, the partial derivative of the objective function with respect to the matrix $\hat{\mathbf{C}}_{\mathbf{xy}}$, $\nabla_{\mathbf{xy}}$, can be derived as:

$$[\nabla_{\mathbf{xy}}]^{ij} = \frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{ij}}$$

$$= \sum_{kl}\frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial\left[\mathbf{T}_H\right]^{kl}} \cdot \frac{\partial\left[\mathbf{T}_H\right]^{kl}}{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{ij}}$$

$$= \sum_{kl}\left[\mathbf{U}\mathbf{V}^\top\right]^{kl} \cdot \left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\right]^{ki}\left[\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\right]^{jl}$$

$$= \left[\hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2}\mathbf{U}\mathbf{V}^\top\hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2}\right]^{ij} \quad (18)$$

Now, the Equation (7) can be derived as following. Let us denote the gradient of the objective function $\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)$ with respect to $\mathbf{H_x}$ as:

$$\frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial[\mathbf{H_x}]^{ij}} = \sum_{kl}[\nabla_{\mathbf{xx}}]^{kl}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}\right]^{kl}}{\partial[\mathbf{H_x}]^{ij}} + \sum_{kl}[\nabla_{\mathbf{xy}}]^{kl}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{kl}}{\partial[\mathbf{H_x}]^{ij}} \quad (19)$$

Each term from the right hand side can be derived as:

$$\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}\right]^{ab}}{\partial[\mathbf{H_x}]^{ij}}$$

$$= \frac{1}{m-1}\frac{\partial}{\partial[\mathbf{H_x}]^{ij}}\left(\left(\left[\bar{\mathbf{H}}_\mathbf{x}\right]^{a*}\right)^\top\left[\bar{\mathbf{H}}_\mathbf{x}\right]^{b*}\right)$$

$$= \begin{cases} \dfrac{2}{m-1}\left([\mathbf{H_x}]^{ij} - \dfrac{1}{m}\sum_k[\mathbf{H_x}]^{ik}\right) & \text{if } a=i, \ b=i \\[2ex] \dfrac{1}{m-1}\left([\mathbf{H_x}]^{bj} - \dfrac{1}{m}\sum_k[\mathbf{H_x}]^{bk}\right) & \text{if } a=i, \ b\neq i \\[2ex] \dfrac{1}{m-1}\left([\mathbf{H_x}]^{aj} - \dfrac{1}{m}\sum_k[\mathbf{H_x}]^{ak}\right) & \text{if } a\neq i, \ b=i \\[2ex] 0 & \text{if } a\neq i, \ b\neq i \end{cases}$$

$$= \frac{1}{m-1}\left(1_{\{a=i\}}\left[\bar{\mathbf{H}}_\mathbf{x}\right]^{bj} + 1_{\{b=i\}}\left[\bar{\mathbf{H}}_\mathbf{x}\right]^{aj}\right) \quad (20)$$

And,

$$\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{ab}}{\partial[\mathbf{H_x}]^{ij}} = \frac{1}{m-1}1_{\{a=i\}}\left([\mathbf{H_y}]^{bj} - \frac{1}{m}\sum_k[\mathbf{H_y}]^{bk}\right)$$

$$= \frac{1}{m-1}1_{\{a=i\}}\left[\bar{\mathbf{H}}_\mathbf{y}\right]^{bj} \quad (21)$$

Hence, equation (19) can be rewritten as:

$$\frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial[\mathbf{H_x}]^{ij}}$$

$$= \sum_{kl}[\nabla_{\mathbf{xx}}]^{kl}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xx}}\right]^{kl}}{\partial[\mathbf{H_x}]^{ij}} + \sum_{kl}[\nabla_{\mathbf{xy}}]^{kl}\frac{\partial\left[\hat{\mathbf{C}}_{\mathbf{xy}}\right]^{kl}}{\partial[\mathbf{H_x}]^{ij}}$$

$$= \frac{1}{m-1}\left(\sum_l\nabla_{\mathbf{xx}}^{il}\bar{H}_\mathbf{x}^{lj} + \sum_k\nabla_{\mathbf{xx}}^{ki}\bar{H}_\mathbf{x}^{kj} + \sum_l\nabla_{\mathbf{xy}}^{il}\bar{H}_\mathbf{y}^{lj}\right)$$

$$= \frac{1}{m-1}\left(\left[\nabla_{\mathbf{xx}}\bar{\mathbf{H}}_\mathbf{x}\right]^{ij} + \left[\nabla_{\mathbf{xx}}^\top\bar{\mathbf{H}}_\mathbf{x}\right]^{ij} + \left[\nabla_{\mathbf{xy}}\bar{\mathbf{H}}_\mathbf{y}\right]^{ij}\right)$$

$$= \frac{1}{m-1}\left(2\left[\nabla_{\mathbf{xx}}\bar{\mathbf{H}}_\mathbf{x}\right]^{ij} + \left[\nabla_{\mathbf{xy}}\bar{\mathbf{H}}_\mathbf{y}\right]^{ij}\right) \quad (22)$$

Therefore,

$$\frac{\partial\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial\mathbf{H_x}} = \frac{1}{m-1}\left(2\nabla_{\mathbf{xx}}\bar{\mathbf{H}}_x + \nabla_{\mathbf{xy}}\bar{\mathbf{H}}_y\right) \quad (23)$$

APPENDIX B
ACOUSTIC FEATURES USED FOR NMED-H DATASET

The 20 acoustic features are extracted from the NMED-H Dataset as discussed in the baseline work by Gang *et al.* [43]. The features are extracted using the MIR toolbox provided by Lartillot and Toiviainen [44]. The 20 acoustic features are calculated as following.

Let, $M_t[f]$ represent the magnitude of Discrete Fourier transform (DFT) of a given audio signal $(m_t(n))$ at frame instant t and frequency bin f. Let, the number of frequency bins be F.

1) **Zero Crossing Rate**: It represents the number of sign changes of the audio signal $m_t(n)$.
2) **Spectral centroid**: The spectral centroid is the first order moment of the DFT given as:

$$C_t = \frac{\sum_{f=1}^{F} f M_t[f]}{\sum_{f=1}^{F} M_t[f]} \qquad (24)$$

3) **High/Low Energy Ratio**: It represents the ratio of the highest to the lowest magnitudes in $M_t[f]$.
4) **Spectral Spread**: It represents the standard deviation of the $M_t[f]$ in the frequency domain.
5) **Spectral Roll-off**: At a given instant, the roll-off R_t is measured as the frequency below which 85% of the magnitude of the Fourier transform is concentrated.

$$\sum_{f=1}^{R_t} M_t[f] = 0.85 \sum_{f=1}^{F} M_t[f] \qquad (25)$$

6) **Spectral Entropy**: It is measured as the relative Shannon entropy of the normalized magnitude $M_t[f]$.
7) **Spectral Flatness**: It represents whether the magnitude distribution in the frequency domain is smooth or spiky. It is measured as the ratio between the geometric mean and the arithmetic mean of the magnitudes in the frequency domain at each instant.

$$F_t = \frac{\sqrt[F]{\pi_{f=1}^{F} M_t[f]}}{\frac{1}{F} \sum_{f=1}^{F} M_t[f]} \qquad (26)$$

8) **Roughness**: It tries to measure the sensory distance related to the beating phenomenon when a pair of sinusoidal signals are close in frequency. It is estimated by obtaining the peaks of the DFT $M_t[f]$, and taking the average of the distances between all pairs of peaks.
9) **RMS energy**: It measures the root mean square value of the magnitudes in the frequency domain, $M_t[f]$, at a given instant.
10) **Broadband Spectral Flux**: It measures the Euclidean distance between the normalized DFT and its previous instant $M_{t-1}[f]$.
11) **Spectral flux for** 10 **octave-wide sub-bands**: It is measured by dividing the frequency domain into 10 bins and calculating the absolute differences of each bin for the two time instants (after normalizing the DFT).

$$S_t[f] = (M_{t-1}[f] - M_t[f])^2 \qquad (27)$$

More details about these features are available in the primer of the music information retrieval (MIR) toolbox by Lartillot and Toiviainen [44].

REFERENCES

[1] S. Sanei and J. A. Chambers, *EEG Signal Processing*. Hoboken, NJ, USA: Wiley, 2007.
[2] R. Galambos and G. C. Sheatz, "An electroencephalograph study of classical conditioning," *Amer. J. Physiol.-Legacy Content*, vol. 203, no. 1, pp. 173–184, 1962.
[3] A. Soman, C. R. Madhavan, K. Sarkar, and S. Ganapathy, "An EEG study on the brain representations in language learning," *Biomed. Phys. Eng. Exp.*, vol. 5, no. 2, Feb. 2019, Art. no. 025041.
[4] S. J. Luck, *An Introduction to Event-Related Potential Technique*. Cambridge, MA, USA: MIT Press, 2014.
[5] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *Eur. J. Neurosci.*, vol. 31, no. 1, pp. 189–193, Jan. 2010.
[6] 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 Hum. Neurosci.*, vol. 10, p. 604, Nov. 2016.
[7] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving precise temporal processing properties of the auditory system using continuous stimuli," *J. Neurophysiol.*, vol. 102, no. 1, pp. 349–359, Jul. 2009.
[8] N. Ding and J. Z. Simon, "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech," *J. Neurosci.*, vol. 33, no. 13, pp. 5728–5735, Mar. 2013.
[9] N. Ding and Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 29, pp. 11854–11859, 2012.
[10] G. M. di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biol.*, vol. 25, no. 19, pp. 2457–2465, 2015.
[11] 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 Biol.*, vol. 28, no. 5, pp. 803–809, 2018.
[12] B. Thompson, *Canonical Correlation Analysis: Uses Interpretation*, vol. 47. Newbury Park, CA, USA: Sage, 1984.
[13] A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.
[14] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers Neurosci.*, vol. 13, p. 153, Dec. 2019.
[15] N. M. Correa, T. Eichele, T. Adalı, Y.-O. Li, and V. D. Calhoun, "Multiset canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," *NeuroImage*, vol. 50, no. 4, pp. 1438–1445, 2010.
[16] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and A. M.-C. So, "Scalable and flexible multiview MAX-VAR canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4150–4165, Aug. 2017.
[17] Q. Zhang, J. P. Borst, R. E. Kass, and J. R. Anderson, "Inter-subject alignment of MEG datasets in a common representational space," *Hum. Brain Mapping*, vol. 38, no. 9, pp. 4287–4301, 2017.
[18] L. C Parra, "Multi-set canonical correlation analysis simply explained," 2018, *arXiv:1802.03759*.
[19] A. de Cheveigné *et al.*, "Multiway canonical correlation analysis of brain data," *NeuroImage*, vol. 186, pp. 728–740, Feb. 2019.
[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[22] B. Kaneshiro, D. Nguyen, J. Dmochowski, A. Norcia, and J. Berger, "Naturalistic Music EEG Dataset-Hindi (NMED-H)," in *Stanford Digital Repository*. Stanford, CA, USA: Stanford Digit. Repository, 2016.

[23] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[24] Y. Bengio, I. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2016.

[25] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, Mar. 2011.

[26] X. Sun, C. Qian, Z. Chen, Z. Wu, B. Luo, and G. Pan, "Remembered or forgotten?—An EEG-based computational prediction approach," *PLoS ONE*, vol. 11, no. 12, Dec. 2016, Art. no. e0167497.

[27] M. Hajinoroozi, Z. Mao, T.-P. Jung, C.-T. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Process., Image Commun.*, vol. 47, pp. 549–555, Sep. 2016.

[28] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2vec: Deep representation learning for clinical temporal data," *Neurocomputing*, vol. 324, pp. 31–42, Jan. 2019.

[29] X. Zheng, W. Chen, M. Li, T. Zhang, Y. You, and Y. Jiang, "Decoding human brain activity with deep learning," *Biomed. Signal Process. Control*, vol. 56, Jan. 2020, Art. no. 101730.

[30] S. Stober, D. J. Cameron, and J. A. Grahn, "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1449–1457.

[31] N. Das, J. Zegers, H. Van Hamme, T. Francart, and A. Bertrand, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 046039.

[32] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *BioRxiv*, vol. 2018, Apr. 2018, Art. no. . 475673.

[33] T. Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, no. 5, pp. 1234–1241, Mar. 2020.

[34] J. R. Katthi, S. Ganapathy, S. Kothinti, and M. Slaney, "Deep canonical correlation analysis for decoding the auditory brain," in *Proc. IEEE Int. Conf. Eng. Med. Biol. (EMBC)*, Oct. 2020, pp. 3505–3508.

[35] J. R. Katthi and S. Ganapathy, "Deep multiway canonical correlation analysis for multi-subject EEG normalization," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 1245–1249.

[36] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs Statistics*. New York, NY, USA: Springer, 1992, pp. 162–190.

[37] 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, Oct. 2018.

[38] K. Lankinen, J. Saari, R. Hari, and M. Koskinen, "Intersubject consistency of cortical MEG signals during movie viewing," *NeuroImage*, vol. 92, pp. 217–224, May 2014.

[39] J. V. Haxby *et al.*, "A common, high-dimensional model of the representational space in human ventral temporal cortex," *Neuron*, vol. 72, no. 2, pp. 404–416, Oct. 2011.

[40] A. de Cheveigné and D. Arzounian, "Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data," *NeuroImage*, vol. 172, pp. 903–912, 2018.

[41] 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, Jul. 2020, Art. no. 116559.

[42] B. B. Kaneshiro, "Toward an objective neurophysiological measure of musical engagement," Ph.D. dissertation, Center Comput. Res. Music Acoust., Stanford Univ., Stanford, CA, USA, 2016.

[43] N. Gang, B. Kaneshiro, J. Berger, and J. P. Dmochowski, "Decoding neurally relevant musical features using canonical correlation analysis," in *Proc. ISMIR*, 2017, pp. 131–138.

[44] O. Lartillot and P. Toiviainen, "A MATLAB toolbox for musical feature extraction from audio," in *Proc. Int. Conf. Digit. Audio Effects*. Bordeaux, U.K., 2007, pp. 237–244.

[45] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[46] V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico, "Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm," *NeuroImage*, vol. 59, no. 4, pp. 3677–3689, 2012.

[47] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.

[48] I. StatSoft, "Statistica (data analysis software system), version 6," *Tulsa*, vol. 150, pp. 91–94, Dec. 2001.

[49] M. Aickin and H. Gensler, "Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods," *Amer. J. Public Health*, vol. 86, no. 5, pp. 726–728, 1996.

[50] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York, NY, USA: Academic, 2013.