Deep Canonical Correlation Analysis For Decoding The Auditory Brain

Jaswanth Reddy Katthi[1], Sriram Ganapathy[1], Sandeep Kothinti[2], Malcolm Slaney[3]

Abstract— **The process of decoding the auditory brain for an acoustic stimulus involves finding the relationship between the audio input and the brain activity measured in terms of Electroencephalography (EEG) recordings. Prior methods focus on linear analysis methods like Canonical Correlation Analysis (CCA) to establish a relationship. In this paper, we present a deep learning framework that is learned to maximize correlation. For dealing with high levels of noise in EEG data, we employ regularization techniques and experiment with various model architectures. With a paired dataset of audio envelope and EEG, we perform several experiments with deep correlation analysis using forward and backward correlation models. In these experiments, we show that regularized deep CCA is consistently able to outperform the linear models in terms of providing improved correlation (up to 9% absolute improvement in Pearson correlation which is statistically significant). We present an analysis that highlights the benefits of using dropouts for neural network regularization in the deep CCA model.**

Clinical relevance — **The proposed method helps to decode human auditory attention. In the case of overlapping speech from two speakers, decoding the auditory attention provides information about how well the sources are separated in the brain and which of the sources is attended. This can impact cochlear implants that use EEG for decoding attention as well as in development of BCI applications. The correlation method proposed in this work can also be extended to other modalities like visual stimuli.**

I. INTRODUCTION

The two main types of non-invasive methods to record brain activity with high temporal resolution are electroencephalography (EEG) and magnetoencephalography (MEG). These signals tend to be highly noisy (with SNRs below -20dB) [1] as the brain recordings capture all the underlying processes in the brain in addition to the desired stimuli induced effects. One of the most common approaches to alleviate this noise is to analyze the data using the event related potential (ERP) method [2] which involves the averaging of brain potentials for a large number of repetitions of the input stimuli for a given condition. Averaging removes the components that are out-of-phase while preserving components that are related to the stimuli. However, this approach is practical only for short stimuli or a small number of isolated events. As EEG analysis extends to the study of responses to longer and naturalistic stimuli such as speech or music, new approaches to analyze single trial EEG data are in substantial demand. For example, the decoding of the semantic components of the auditory stimulus from the EEG data can provide insight into semantic processing in the brain [3] or highlight the attended speaker in a cocktail party listening environment [4], [5].

The single trial decoding techniques that have been successful thus far assume that the entire system is linear and time-invariant. The earliest methods in this direction approximated the relation between stimulus and response as a convolution with an impulse response referred to as the temporal response function (TRF) [6]. This approach relies on a reverse correlation/system identification framework. The linear TRF model can be a forward model where the model predicts the EEG response from the audio or a backward model where the neural response is used to predict the features of the audio signal. The measure of performance is typically the Pearson correlation between the predicted signal and the true signal. However, using these methods, the average correlation on held-out EEG is typically in the range 0.1-0.2 [6]. The low correlation may be attributed to the fact that EEG signals represent the stimuli effects along with the other brain processes and thus only a fraction of the variance can be explained from the external stimuli.

Canonical correlation analysis (CCA) is a more powerful linear method to project two signals to a domain that maximizes the correlation between the two signals [7], [8]. The method finds a linear transform on each of the signals that minimizes the variability irrelevant to the other signal. Recently, the method has been successfully applied in forward and backward models in auditory EEG analysis using a combination of linear transforms and convolutions [9], [10], [11]. However, the model is still based on linear assumptions.

In this work, we propose to develop a deep CCA model for audio-EEG data. A deep model of CCA was proposed for image data [12], [13] where the model learns deep feature embeddings from each of the datasets which maximize the correlation. This method was shown to be significantly better than the linear CCA for image data under low noise conditions. The main challenge compared to image data [12] is the significant amount of noise present in the EEG signal. We use the dropout strategy to avoid overfitting and also use a leaky-ReLU based non-linear activation. The dropout regularization prevents the model from overfitting. In audio-EEG experiments with the forward and backward model and various CCA configurations, we show that the deep CCA consistently improves over the linear CCA model. A pairwise t-test comparing the proposed method with the linear CCA also reveals that the improvements are statistically significant for all the subjects ($p < 0.01$).

The rest of the paper is organized as follow. Sec. II gives the mathematical details of CCA and deep CCA. Sec. III details the proposed model of deep CCA for audio-EEG analysis. The experiments and analysis are presented in Sec. IV. A summary of the work is given in Sec. V.

II. CANONICAL CORRELATION ANALYSIS

A. Linear CCA

For a pair of multi-variate datasets, Canonical Correlation Analysis (CCA) [7] finds the optimal linear transforms that maximize the Pearson correlation between the transformed vectors.

Let, x and y denote $\mathscr{D}_1$ and $\mathscr{D}_2$ dimensional vectors respectively. Let, n denote the dimension of the desired canonical sub-space

This work started at the Telluride Neuromorphic Cognition Engineering Workshop 2019 held during the summer of 2019 and sponsored by the NSF. The work was performed with grants from the Pratiksha Trust Investigator award and the Department of Science and Technology, India grants under the Early Career Research Award (ECR01341).

[1] Jaswanth Reddy Katthi and Sriram Ganapathy are with the Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.

[2] Sandeep Kothinti is with the Department of ECE, Johns Hopkins University, Baltimore, MD, USA.

[3] Malcolm Slaney is with the Google Research, Machine Hearing Group, Mountain View, CA, USA.

which maximizes the correlation between transformed vectors. For example, if $n = 1$, let u_1, v_1 denote the pair of vectors which project x and y respectively into 1-dimensional space. Now, the problem is to find u_1 and v_1 such that the correlation, ρ, between $x' = {u_1}^T x$ and $y' = {v_1}^T y$ is maximized. The problem can be written as maximizing

$$\rho = \frac{{u_1}^T \mathbf{C}_{xy} v_1}{\sqrt{{u_1}^T \mathbf{C}_{xx} u_1 {v_1}^T \mathbf{C}_{yy} v_1}} \tag{1}$$

where, $\mathbf{C}_{xy} = E[(x-\mu_x)(y-\mu_y)^T]$ and $\mathbf{C}_{xx}$, $\mathbf{C}_{yy}$ are the auto-correlation matrices of x, y respectively.

Let, $T \triangleq \mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}$. Then, the solution to the CCA problem ($\mathbf{u}_l^*$ and $\mathbf{v}_l^*$) are given as the first left and right singular vectors of the T matrix and the maximum correlation is the top singular value [12]. This can be extended for $n > 1$ by finding the subsequent singular vectors.

B. Deep CCA

The extension of the linear transformation based CCA analysis to deep transformation learning based CCA was first proposed by Andrew et.al [12]. The two input sets of vectors are passed through a pair of feed-forward connections to undergo a set of non-linear transformations. The outputs of each network are the final representations on which the cross correlation is computed. The neural network is trained to maximize the correlation cost.

Let the non-linear transform performed by the first neural network on x be denoted as $f_1(\cdot)$. Similarly, let the second network transformation on y be denoted as $f_2(\cdot)$. Let θ_1 be the set of all trainable parameters of the first neural network and θ_2 be that of the second network. The deep CCA optimization is

$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1,\theta_2)}{\operatorname{argmax}} \operatorname{corr}\left(f_1\left(x;\theta_1\right), f_2\left(y;\theta_2\right)\right) \tag{2}$$

Let the dimensionality of the outputs of the two neural networks be n and a batch of m examples from each of the (x, y) are used in training. Let, $\mathbf{H_x}, \mathbf{H_y} \in \mathbb{R}^{n\times m}$ be the matrices whose columns are the feed-forward network output from the first and second network respectively.

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

$$\mathbf{C_{xx}} = \frac{1}{m}\bar{\mathbf{H}}_\mathbf{x}\bar{\mathbf{H}}_\mathbf{x}^\mathbf{T} + r_1 I \tag{3}$$

$$\mathbf{C_{yy}} = \frac{1}{m}\bar{\mathbf{H}}_\mathbf{y}\bar{\mathbf{H}}_\mathbf{y}^\mathbf{T} + r_2 I. \tag{4}$$

where $r_1, r_2 > 0$ are the regularization parameters so that the covariance matrices are positive definite and I is the identity matrix. Similarly, the cross correlation is $\mathbf{C_{xy}} = \frac{1}{m}\bar{\mathbf{H}}_\mathbf{x}\bar{\mathbf{H}}_\mathbf{y}^\mathbf{T}$.

Let, $T \triangleq \mathbf{C_{xx}}^{-1/2}\mathbf{C_{xy}}\mathbf{C_{yy}}^{-1/2}$. It can be shown [12] that the gradient of $\operatorname{corr}\left(\mathbf{H_x}, \mathbf{H_y}\right)$ is given by,

$$\frac{\partial \operatorname{corr}\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial \mathbf{H_x}} = \frac{1}{m-1}\left(2\nabla_{xx}\bar{\mathbf{H}}_\mathbf{x} + \nabla_{xy}\bar{\mathbf{H}}_\mathbf{y}\right) \tag{5}$$

where

$$\nabla_{xy} = \mathbf{C_{xx}}^{-1/2} U V' \mathbf{C_{yy}}^{-1/2}, \tag{6}$$

$$\nabla_{xx} = -\frac{1}{2}\mathbf{C_{xx}}^{-1/2} U D U' \mathbf{C_{xx}}^{-1/2} \tag{7}$$

where U,V and D are obtained from the singular value decomposition of T as $T = UDV'$. Similar expression can be obtained for gradient with respect to $\mathbf{H_y}$. These gradients are backpropagated to learn the model parameters θ_1 and θ_2 of the two neural networks $f_1(.)$ and $f_2(.)$.

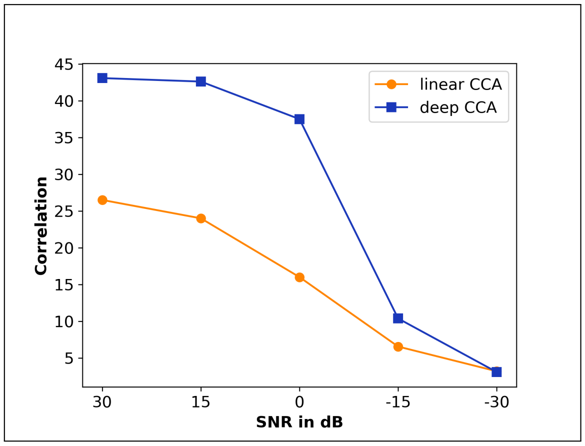

Fig. 1. Comparison between linear and Deep CCA for different SNR on the MNIST task. The number of dimensions at the output is 50.

The original Deep CCA work [12] showed that the correlation between left and right halves of MNIST hand-written digit images can be increased significantly by using the deep CCA model over the linear CCA model. In this work, we analyze the impact of noise on correlation training. Specifically, we apply the linear CCA and deep CCA with varying amounts of noise on the left half of the image (similar to the presence of noise in EEG recordings) and measure the performance. This result in shown in Fig. 1. Here, we use $n = 50$ dimensions at the output of linear CCA/deep CCA. As seen here, when the models are trained with noise, the correlation drops significantly for both the models in the presence of noise. In particular, the deep models do not have an advantage over the linear models under noisy conditions below -15dB. The difficulty in modeling noisy data proves to be challenging for the DCCA model when it is applied to EEG data. This is partly alleviated with dropout regularization.

III. Audio-EEG DCCA

The linear CCA analysis described by de Cheveigné [9] forms the baseline for this work. We use the same stimuli response data collected by Liberto et. al. [14]. Specifically, the EEG recordings from 128 channels are recorded when subjects are listening to a male speaker reading snippets of a novel. A Biosemi system was used for EEG data collection that was sampled at 512 Hz. We use 20 speech excerpts, each of duration approximately 3 minutes presented diotically via headphones. The EEG data were downsampled to 64 Hz. It was further processed using de-trending and de-noising using noise tools software [15]. The data were processed with band-pass filtering between $0.1-12$ Hz. The stimulus data was obtained from audio sampled at 44100 Hz. The audio envelope was obtained by a squaring and smoothing operation by convolution with a square window and downsampled to 64 Hz. The envelope was further compressed to the power $1/3$. In all our experiments, we perform DCCA projection to one dimension and compare with the linear CCA projection to one dimension. More details about the EEG pre-processing and the audio envelope extraction are described by de Cheveigné [9].

The configuration of DCCA models that are tested in this work are shown in Fig. 2.

1) **DCCA1** : The 40 time delayed audio envelope samples and the 40 principal components of the 128 channel EEG are

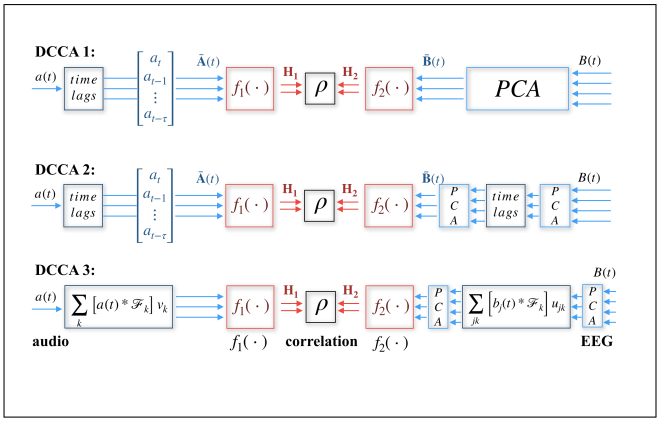

Fig. 2. In DCCA1 model, the time lagged stimulus audio is provided to the Deep CCA module whereas EEG response is provided after performing PCA, whose outputs' correlation is computed. In DCCA2 model, the outputs of the PCA on the EEG response side go through a set of time lags and one more PCA before being provided to the Deep CCA module. In DCCA3 model, both the audio inputs and the EEG outputs go through a filterbank of FIR filters rather than delays.

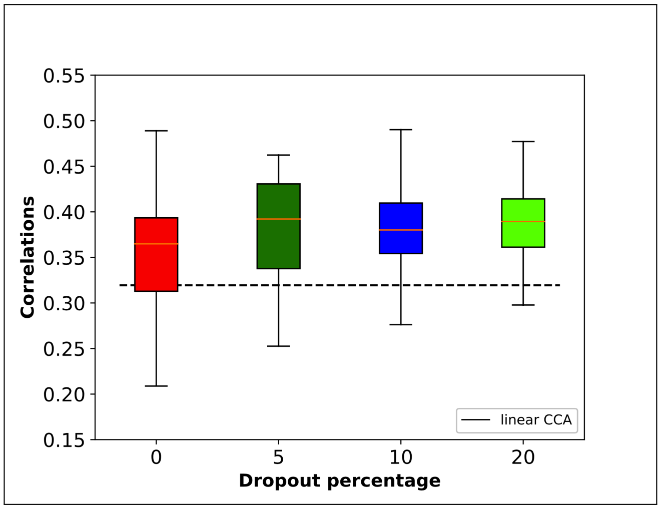

Fig. 3. For subject 4, DCCA3 model (averaged) correlations as a function of the dropout regularization in the neural network. The horizontal dotted line is of the linear CCA3 model. The correlations are averaged over all the 20 validation sessions.

provided to the Deep CCA network as inputs.

2) **DCCA2** : In this case, the 128 channel EEG is processed with a PCA transformation of 60 dimensions. Then, 10 time lags of 60D EEG data yielding 600 dimensions is further processed with another PCA transformation to generate 80 dimensional response data. The 80 time delayed audio envelope samples (similar to DCCA1) are used for the stimulus side.
3) **DCCA3** : For this model, a filter-bank of 21 FIR band-pass filters whose characteristics like centre frequency, bandwidth and duration of impulse response are uniformly distributed on a logarithmic scale. The 21 D output of the audio envelope after passing through the filter-bank is used as the representation from the stimulus side. The 60 D PCA output of the original 128 channel EEG through the 21 filter filterbank gives 1260 dimensional response. This high dimensional vector is processed with another PCA transformation to 139 dimensional output. We use a wavelet based filter-bank for this implementation.

All the above model configurations are homologus to the linear CCA models proposed by de Cheveigné et. al. [9]. We try two neural network architectures for the deep CCA models. The first architecture [12] contains a 2 hidden layer network for each of the stimulus side (audio envelope) and the response side (EEG), with 2038 and 1608 units for the first and second layers respectively followed by a 1 dimensional output layer. The second architecture is similar to the one proposed by Wang et. al. [13] and it contains 4 hidden layers, each of 1024 units. We use a leaky ReLU activation function, with a negative slope coefficient of 0.1, at the output of the deep CCA model [16].

IV. Results

The first architecture [12] is consistently found to provide better results, when all the remaining hyperparameters are same for both the architectures [12] [13]. So, all the following results are from the first architecture. In this work, the leaky ReLU was found to be empirically better than other output activation functions like linear and sigmoid. We also incorporate dropout regularization [17], [18] in the deep CCA model training to avoid over-fitting in the noisy conditions. With varying amounts of dropout regularization, we show that the deep CCA model can outperform the linear model in the noisy conditions.

Each subject recorded 20 sessions with approximately 160 seconds of audio recording in each session. All the results are obtained for 20 fold validation experiments in which one of the sessions is held out as the test data while the 19 other sessions are used in training the model (both the linear models as well as the DCCA models). Given a sampling rate of 64 Hz, the approximate number of instances for the model training is about $19 * 64 * 160 \approx 200k$. This set of training instances were further split randomly into training and validation with a $90 - 10$ split.

The first experiment compared the impact of dropout on the model training. The first set of experiments is performed on the 20 sessions from one subject in the pool. Given that the deep models are prone to over-fitting, particularly with the high levels of noise in EEG data, we find that using dropout in the model training provides significant boost in the correlation performance (Fig. 3). The best average correlation (among all the 20 sessions) is obtained for a dropout of 5%. All the subsequent experiments use dropout in the model training.

All three DCCA models are compared with their linear counterpart and these results are shown in Fig. 4. As seen here, among all the linear models, the CCA3 model gives the highest correlation (similar to the analysis by de Cheveigné [9]). Comparing linear CCA versus DCCA for all the three configurations, it is seen that the DCCA outperforms the linear model consistently for all the sessions. The best correlation results are obtained for DCCA3 model (the average correlation over all the 20 sessions is about 0.4). The absolute improvement in correlation over the best linear model for the DCCA3 model is about 9% in terms of the Pearson correlation.

The comparison of the average correlation for the linear CCA3 model and the DCCA3 model is shown in Fig. 5 for all 6 subjects. Note that, for each subject 20 linear and deep CCA models were trained with leave one out training and the average correlation on the held out session is used for that subject. As seen here, the DCCA model consistently improves over the linear CCA model in all the evaluations. The absolute improvements in terms of % correlation ranges from $2 - 9$ % for these subjects. We also tested

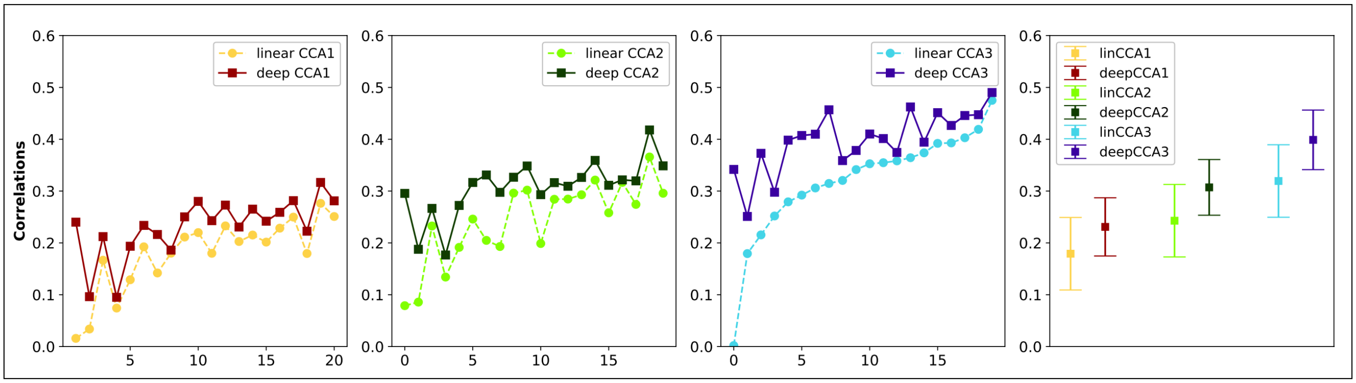

Fig. 4. Comparison of linear and deep CCA models with CCA1, CCA2 and CCA3 configurations, for subject 4. The session indices (x-axis) are arranged in the non-decreasing order of the correlations obtained for the linear CCA3 model The first three plots show the results for 20 sessions and the last plot shows the average over all the 20 sessions. A dropout value of 0.1 was used in these experiments.

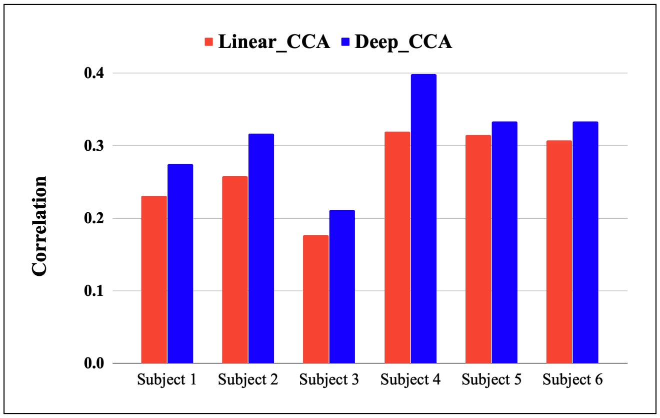

Fig. 5. Comparison of linear CCA3 and the DCCA3 configuration for different subjects.

the statistical significance of the improvements in correlations for the DCCA model over the linear CCA method using a pairwise t-test on each subject. The improvement in correlation values were found to be statistically significant for all the subjects ($p < 0.01$). These results highlight that the deep methods have the potential to become the de-facto standard in auditory attention decoding.

V. Summary

In this paper, we used a deep CCA model to decode auditory EEG activity. The audio stimuli features are the time-lagged envelopes while the EEG response features are the PCA transformed EEG recordings. The deep CCA model performs a non-linear mapping of the stimuli/response where the correlation is maximized. We experimented with several configurations of the deep CCA model and show that the model consistently outperforms the linear counterpart. In summary, this work shows that the deep correlation methods constitute a useful method for analyzing complex relationships between stimulus and EEG recordings. In future, we plan to extend this approach to other auditory decoding tasks like reconstructing the spectrogram from EEG and to understand the encoding of various acoustic features in the EEG signal.

References

[1] Saeid Sanei and Jonathon A Chambers, "EEG signal processing," *Wiley Online Library*, 2007.

[2] Michael GH Coles and Michael D Rugg, *Event-related brain potentials: An introduction.*, Oxford University Press, 1995.

[3] Michael P Broderick, Andrew J Anderson, and Edmund C Lalor, "Semantic context enhances the early auditory encoding of natural speech," *Journal of Neuroscience*, vol. 39, no. 38, pp. 7564–7575, 2019.

[4] Bojana Mirkovic, Martin G Bleichner, Maarten De Vos, and Stefan Debener, "Target speaker detection with concealed eeg around the ear," *Frontiers in neuroscience*, vol. 10, pp. 349, 2016.

[5] Adam Bednar and Edmund C Lalor, "Where is the cocktail party? decoding locations of attended and unattended moving sound sources using eeg," *NeuroImage*, p. 116283, 2019.

[6] Edmund C Lalor, Alan J Power, Richard B Reilly, and John J Foxe, "Resolving precise temporal processing properties of the auditory system using continuous stimuli," *Journal of neurophysiology*, vol. 102, no. 1, pp. 349–359, 2009.

[7] Harold Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190. Springer, 1992.

[8] Bruce Thompson, *Canonical correlation analysis: Uses and interpretation*, Number 47. Sage, 1984.

[9] Alain de Cheveigné, Daniel DE Wong, Giovanni M Di Liberto, Jens Hjortkjaer, Malcolm Slaney, and Edmund Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[10] Alain de Cheveigne, Giovanni M Di Liberto, Dorothee Arzounian, Daniel DE Wong, Jens Hjortkjaer, Søren Fuglsang, and Lucas C Parra, "Multiway canonical correlation analysis of brain data," *NeuroImage*, vol. 186, pp. 728–740, 2019.

[11] Emina Alickovic, Thomas Lunner, Fredrik Gustafsson, and Lennart Ljung, "A tutorial on auditory attention identification methods," *Frontiers in neuroscience*, vol. 13, 2019.

[12] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[13] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[14] Giovanni M Di Liberto, James A O'Sullivan, and Edmund C Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

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

[16] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[17] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.