ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448271

INVESTIGATING SELF-SUPERVISED DEEP REPRESENTATIONS FOR EEG-BASED AUDITORY ATTENTION DECODING

Karan Thakkar, Jiarui Hai, Mounya Elhilali

Laboratory for Computational Audio Perception, Johns Hopkins University, USA

ABSTRACT

Auditory Attention Decoding (AAD) algorithms play a crucial role in isolating desired sound sources within challenging acoustic environments directly from brain activity. Although recent research has shown promise in AAD using shallow representations such as auditory envelope and spectrogram, there has been limited exploration of deep Self-Supervised (SS) representations on a larger scale. In this study, we undertake a comprehensive investigation into the performance of linear decoders across 12 deep and 2 shallow representations, applied to EEG data from multiple studies spanning 57 subjects and multiple languages. Our experimental results consistently reveal the superiority of deep features for AAD at decoding background speakers, regardless of the datasets and analysis windows. This result indicates possible nonlinear encoding of unattended signals in the brain that are revealed using deep nonlinear features. Additionally, we analyze the impact of different layers of SS representations and window sizes on AAD performance. These findings underscore the potential for enhancing EEG-based AAD systems through the integration of deep feature representations.

Index Terms— auditory attention decoding, electroencephalogram (EEG), self-supervised speech representations

1. INTRODUCTION

Throughout daily life, we all experience the challenges of following a particular conversation in presence of other competing speakers or noise sources. This challenge is particularly pronounced in individuals with hearing impairment and impedes their ability to interact socially. This process relies on our brain's attentional mechanisms in order to hone in on a speaker of interest and render the rest of the acoustic scene to the background. Auditory attention decoding (AAD) is a general framework developed to determine the sound a listener is attending to based on their brain activity; hence holding the potential to improve hearing aids and neuroprosthetics [1]. Various methods to collect brain data have shown promise for AAD, though they come with different trade-offs in terms of invasiveness, portability, resolution, and signal quality. Non-invasive electroencephalography (EEG), for instance, relies on scalp electrodes to capture signals [2]. It is also the most portable and adaptable allowing the user to potentially go about their daily life; though it comes at a cost of lower spatial resolution and signal quality. On the other hand, magnetoencephalography (MEG) offers higher resolution in mapping brain activity patterns, but demands bulky and expensive equipment for its operation [3]. Alternative invasive methods such as electrocorticography (ECoG) implant electrodes directly into the brain, providing unparalleled precision but also exposing the patient to surgical risk.

Across all these techniques, AAD learns a mapping between the complex auditory stimuli entering the ears and the brain activity patterns generated in response [2]. In its simplest form, this mapping can be approximated by a linear function (e.g. regression) by learning correspondence between brain signals and the attended envelope or spectrogram of the foreground speaker [4]. These techniques have shown great promise for AAD given their simplicity and minimal training and data needs. In contrast, deep learning excels at learning hierarchical abstractions, allowing it to identify intricate relationships between auditory inputs and neural representations. Recent work has shown promise in applying mappings learned through CNNs or LSTMs [5, 6]; though by-and-large, these models are data-and computation-hungry and require extensive tuning.

Nevertheless, deep learning representations offer the potential to capture intricate and nuanced connections between auditory stimuli and neural responses, which may elude simpler linear models, particularly in shedding light on how the brain disentangles the representation of foreground (attended) and background (unattended) information. Building on this potential, this study performs a meta-analysis of a range of deep and shallow features, all evaluated within the *same* framework on publicly available datasets, spanning different research subjects and languages. The study addresses the following research questions: **1)** How do deep features fair against shallow features in a direct AAD comparison? **2)** Do abstractions learned through deep features reveal distinctions in how the brain represents foreground and background sensory signals? **3)** How generalizable are deep features trained on one language to other languages when applied to AAD?

This work was supported by ONR N00014-23-1-2050 and N00014-23-1-2086 and NIH U01AG058532

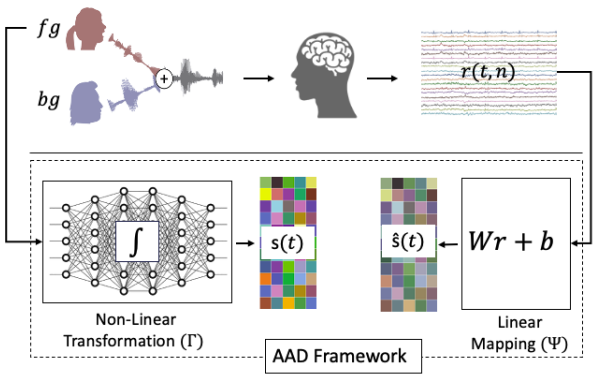

Fig. 1. Auditory Attention Decoding (AAD) framework

2. AAD METHODOLOGY

2.1. General AAD framework

The general AAD framework learns the mapping between an audio signal $a(t)$ and the brain response $r(t,n)$, where t is the index of time and n represents different neural channels. This mapping is often estimated in two steps (Eq.1):

$$a(t) \xrightarrow{\Gamma(.)} s(t) \xleftrightarrow{\Psi(.)} r(t,n) \tag{1}$$

The first step (Γ) projects the audio signal onto a meaningful auditory representation $s(t)$ (see Fig. 1), such as the signal envelope or spectrogram (or mel-spectrogram). These two representations have meaningful links to the patterns that invoke strong responses in cortical networks in the brain that are typically observed in surface electrodes [2, 4]. The second step learns a mapping Ψ between the representation $s(t)$ and neural response $r(t,n)$. The linear AAD scheme employs a regression framework that minimizes the Mean Squared Error (MSE) loss between the ground truth and predicted signals. Alternative models have considered end-to-end schemes to learn nonlinear transformations using CNNs, LSTMs and self-attention networks, therefore accounting for complex relationships and variable interactions between the sensory and brain signals [5, 6, 7]. Though leading to improved performance over the linear formulations, these end-to-end systems have limited generalizability due to data requirements, computational resources and need for extensive tuning and regularization. An alternative method that has been recently considered is to leverage deep methods for the initial mapping $\Gamma(.)$ ending with a linear layer for the second mapping $\Psi(.)$. This approach enables adoption of a wider range of deep methods particularly self-supervised embeddings which can be trained independently on larger datasets. This framework has been recently tested on a small scale in ECoG recordings from 3 subjects [8]. The scalability of this scheme for different embedding features and larger datasets has not been evaluated before.

Table 1. Self-Supervised Model Specifications

Model	Quantized	Stride	# Layers
AlBERT	x	10ms	4
Mockingjay	x	10ms	4
TERA	x	10ms	4
HuBERT	✓	20ms	13
Wav2Vec2.0	✓	20ms	13
WavLM	✓	20ms	13

2.2. Meta Analysis of Sensory Representations

The current study explores a diverse array of Self Supervised (SS) representations, which have recently demonstrated promising results in various audio and speech tasks [9]. Central to the potential of these models is the use of pre-trained transformer architectures, adaptable for different audio and speech applications. The current study explores the benefits of a range of these representations for AAD. Generally, SS mappings can be grouped into 2 broad categories. Models such as ALBERT [10], Mockingjay [11], and TERA [12] focus on reconstructing continuous filter bank features by masking parts of the input along different axes. Conversely, wav2vec2.0 [13] and HuBERT [14] focus on extracting discrete features from time-domain signals and learning to predict these discrete representations from masked audio signals. Notably, while wav2vec2.0 jointly trains the tasks of vector quantization and mask prediction, HuBERT employs an iterative re-clustering and re-training method for discrete representation learning and mask prediction. Based on HuBERT, a more recent development WavLM [15] further extends the pre-training task to both reconstruction and denoising. Table 1 summarizes all the deep representations considered for our meta-analysis. All SS embeddings map onto a 768 dimensional representation hence allowing side-by-side comparisons of different deep features. Additionally, we also consider 'shallow' features that have been widely used in AAD tasks, notably the auditory envelope and mel-spectrogram.

2.3. Decoding Evaluations

In a two-speaker scenario, AAD can be approached from two distinct angles, as suggested by [16]: attended decoding and unattended decoding. Attended decoding focuses on identifying the speaker to whom attention is directed, while unattended decoding aims to ascertain information about the unattended speaker (see Fig. 1). In attended decoding, a trial is considered correctly decoded if the correlation between foreground and predicted foreground is greater than or equal to the correlation between background, and predicted foreground; and the opposite in unattended decoding. Studies suggest that both the attended and unattended speech streams are present in brain signals [17], further emphasizing the significance of these different decoding approaches.

Table 2. Dataset Information

Dataset	#Subjects	Duration	Language
FU_18 [18]	18	15 hrs	Dutch
ET_22 [19]	18	6 hrs	English
ZH_20 [20]	21	14 hrs	Mandrian
Total	57	35 hrs	-

3. EXPERIMENTAL SETUP

3.1. Neural Datasets and Preprocessing

The evaluation covers three unique datasets each with a different language: the Fuglsang dataset (FU_18) also commonly known as the DTU dataset [18]; the Etard dataset (ET_22) [19]; and the Zhang dataset (ZH_22) [20], as outlined in Table 2. The trials corresponding to the repetition and single-speaker conditions were discarded from all datasets to avoid any potential leaks in the test set. To ensure a fair comparison across these datasets, we employed a standardized preprocessing pipeline using the MNE python toolbox [21]. The standardized preprocessing involved the following steps: line noise removal, band pass filtering 0-8 Hz, artifact removal, and re-referencing based on average. All EEG channels were included in the modeling and downsampled to 64 Hz for AAD. All auditory stimuli were downsampled to 16KHz before applying different transformations for feature extraction.

3.2. Deep Feature Extraction and Preprocessing

Each deep feature was extracted using the standard S3PRL [9] upstream configuration. Amongst different versions of the models available online, we only selected the model checkpoints that were trained on the LibriSpeech [22] corpus. We analyze two different layer combinations for each SSL model: the Last Layer (LL) and the First-Middle-Last (FML) concatenation representation. Analyzing the FML outputs allowed us to assess whether using information from multiple layers could improve AAD performance. To reduce the high dimensional embedding space of deep representations (768 dimensions), we reduced the embedding space of each layer in a model down to its 20 principal components. Leading to 20 channels for LL and 60 channels for FML configuration after concatenation. Due to different stride values (see Table 1), all features were resampled to 64 Hz. Lastly, all features underwent normalization before training.

3.3. Shallow Feature Extraction and Preprocessing

The speech envelope is estimated using a gamma tone filter bank of 28 filters, covering frequencies from 50 Hz to 5 kHz to capture key auditory features (Bisman et al. [23]). To enhance the envelope first, the absolute value of each filter sample is computed to focus on signal magnitude. Then, these absolute values are exponentiated with a power factor of 0.6 to

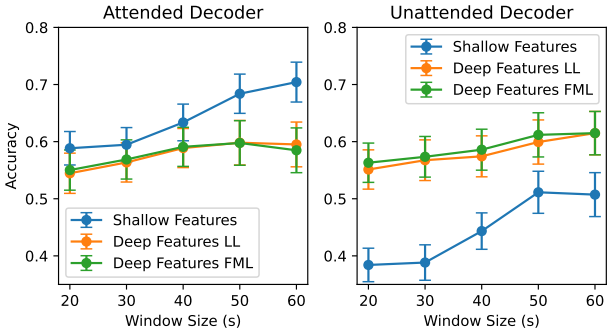

Fig. 2. AAD performance of Attended and Unattended decoders with varying window sizes averaged individually for three groups across all the features.

model the relationship between perceived loudness and intensity. Finally, the algorithm calculates the mean across all 28 filters to obtain the speech stimulus envelope. A 20-bin mel spectrogram is extracted using Librosa's[24] built-in function with a 25 ms window. The choice of 20 bins is to match it along the reduced 20-dimensional space of deep representations. Both the features are downsampled to 64 Hz and normalized for AAD.

3.4. Model Training and Cross Validation

To train the linear decoder, we harnessed the regularized TimeDelayingRidge toolbox available in MNE Python [21]. The multivariate approach employed in this article shares similarities with the mTRF method outlined in the [4]. The primary objective is to establish a linear relationship between the EEG data and the target audio features over an integration window. Ridge regression introduces a regularization term to the standard mean squared error, controlled by the hyperparameter λ. The final solution is obtained by minimizing the MSE loss between the reconstructed feature and the target feature with a given hyperparameter λ. We employed models for each subject, implementing a 90-10 train-test split for each subject's data. For hyperparameter tuning, we conducted cross-validation on each subject, utilizing leave-one-out cross-validation on the training dataset to determine the optimal λ. We choose a time-delayed window of 500ms for model training similar to [16] accounting for the delay in EEG response to the audio stimuli.

4. EXPERIMENTAL RESULTS

4.1. Main Results

AAD performance was measured using accuracy on the subject's test trials over non-overlapping windows. In Table 3, our results across various experiments and datasets demonstrate that shallow features are better when decoding attention using the attended decoder. However, when using the

Table 3. Comparison of Attended and Unattended Decoders' accuracy across varying datasets and features.

Feature	Attended Decoder				Unattended Decoder			
	FU_18	ET_22	ZH_20	Avg	FU_18	ET_22	ZH_20	Avg
Envelope	0.65 ± 0.36	0.94 ± 0.23	0.55 ± 0.25	0.69 ± 0.34	0.48 ± 0.34	0.55 ± 0.51	0.48 ± 0.25	0.49 ± 0.36
Spectrogram	0.70 ± 0.32	0.89 ± 0.32	0.60 ± 0.27	$\mathbf{0.71 \pm 0.33}$	0.53 ± 0.36	0.56 ± 0.51	0.46 ± 0.26	0.51 ± 0.37
Albert_ll	0.73 ± 0.30	0.50 ± 0.51	0.58 ± 0.27	0.65 ± 0.35	0.69 ± 0.34	0.61 ± 0.50	0.57 ± 0.21	0.64 ± 0.35
Mockingjay_ll	0.74 ± 0.35	0.44 ± 0.51	0.46 ± 0.30	0.62 ± 0.39	0.69 ± 0.36	0.50 ± 0.51	0.54 ± 0.27	0.62 ± 0.38
Tera_ll	0.77 ± 0.32	0.33 ± 0.49	0.48 ± 0.26	0.62 ± 0.38	0.74 ± 0.30	0.78 ± 0.43	0.59 ± 0.23	$\mathbf{0.71 \pm 0.32}$
Hubert_ll	0.72 ± 0.30	0.33 ± 0.49	0.45 ± 0.27	0.58 ± 0.37	0.60 ± 0.34	0.50 ± 0.51	0.56 ± 0.24	0.57 ± 0.36
Wav2Vec2_ll	0.61 ± 0.35	0.50 ± 0.51	0.47 ± 0.24	0.55 ± 0.36	0.63 ± 0.38	0.50 ± 0.51	0.59 ± 0.30	0.59 ± 0.37
WavLM_ll	0.57 ± 0.34	0.50 ± 0.51	0.43 ± 0.28	0.52 ± 0.37	0.61 ± 0.36	0.33 ± 0.49	0.53 ± 0.29	0.53 ± 0.38
Albert_fml	0.68 ± 0.34	0.39 ± 0.50	0.50 ± 0.21	0.58 ± 0.37	0.69 ± 0.35	0.67 ± 0.49	0.59 ± 0.27	0.66 ± 0.36
Mockingjay_fml	0.68 ± 0.39	0.39 ± 0.50	0.55 ± 0.25	0.59 ± 0.40	0.62 ± 0.39	0.50 ± 0.51	0.54 ± 0.31	0.57 ± 0.40
Tera_fml	0.68 ± 0.35	0.39 ± 0.50	0.55 ± 0.28	0.59 ± 0.38	0.73 ± 0.33	0.67 ± 0.49	0.60 ± 0.27	0.69 ± 0.34
Hubert_fml	0.70 ± 0.33	0.44 ± 0.51	0.49 ± 0.26	0.60 ± 0.37	0.64 ± 0.37	0.44 ± 0.51	0.55 ± 0.24	0.58 ± 0.38
Wav2Vec2_fml	0.67 ± 0.32	0.39 ± 0.50	0.49 ± 0.20	0.57 ± 0.35	0.64 ± 0.36	0.50 ± 0.51	0.59 ± 0.26	0.60 ± 0.37
WavLM_fml	0.65 ± 0.35	0.39 ± 0.50	0.48 ± 0.25	0.56 ± 0.37	0.64 ± 0.36	0.33 ± 0.49	0.63 ± 0.26	0.57 ± 0.38

unattended decoder the deep representations are better. Notably, the TERA [12] model consistently outperforms others across all datasets. This suggests that deep features have an advantage in capturing and decoding unattended signals in the brain, implying potential nonlinear encoding of auditory information in EEG similar to that of the deep features. Nonetheless, the TERA features stand out as best performer for the FU_18 dataset. Despite deep features being pretrained on English data, our features exhibit no bias towards English language datasets, indicating their adaptability for cross-linguistic analysis in brain signal decoding. We do not observe any consistent improvements in concatenating representations from the layers of the model across features when using linear decoders. Overall, there is no one feature that works best across the attended and unattended decoders using linear decoders on EEG data.

4.2. Effect of Window Size

To further explore the factors influencing AAD performance, we investigated the impact of different window sizes on the decoding accuracy. We varied the window sizes while keeping the representation type constant and analyzed the resulting performance differences. Our analysis revealed that window size had a significant influence on AAD performance. Specifically, larger window sizes tended to yield higher decoding accuracy, suggesting that a broader temporal context enhances the ability to isolate desired sound sources (see Fig. 2). In addition, we also observe that the linear decoding accuracy of the unattended decoder for deep features is consistently better across all window sizes. And there exists no significant improvement in adding information from different layers of the model across different window sizes.

4.3. Investigating the Model Weights

Examining the normalized energy of linear model weights across various time delays (0-500ms) in our study revealed insights into the temporal dynamics of AAD. The energy distribution of initial time delays exhibited higher weighting indicating a marker for auditory attention at the start. However, the weightage decreases from 200 to 300 ms and then peaks higher indicating the discrimination between attended and unattended is maximal at a time delay beyond 300ms time delay. This observation is dominant across many features as observed in Fig. 3.

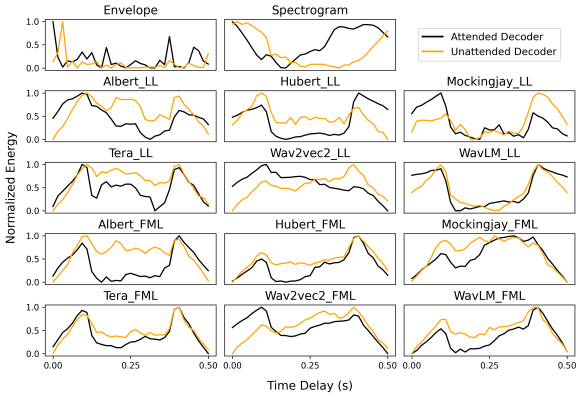

Fig. 3. Normalized energy plots of model weights when averaged across channels.

5. CONCLUSION

The study offers a meta-analysis of different mapping of audio stimuli on large EEG data. By leveraging a final linear layer, the analysis provides a controlled comparison across shallow and deep embeddings and reveals the possible value of nonlinear mappings in unraveling different mechanisms of encoding foreground and background information in the brain. As this technology continues to evolve, these findings open exciting possibilities for the exploration of new features learned by large deep neural networks for improving AAD.

6. REFERENCES

[1] Simon Van Eyndhoven, Tom Francart, and Alexander Bertrand, "Eeg-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2017.

[2] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[3] Sahar Akram, Alessandro Presacco, Jonathan Z Simon, Shihab A Shamma, and Behtash Babadi, "Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, pp. 906–917, 2016.

[4] Michael J Crosse, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, pp. 604, 2016.

[5] I. Kuruvila, J. Muncke, E. Fischer, and U. Hoppe, "Extracting the auditory attention in a dual-speaker scenario from eeg using a joint cnn-lstm model," *Frontiers in Physiology*, vol. 12, pp. 700655, August 2021.

[6] B Accou, J Vanthornhout, HV Hamme, and T Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network," *Scientific Reports*, vol. 13, no. 1, pp. 812, Jan 2023.

[7] Yun Lu, Mingjiang Wang, Longxin Yao, Hongcai Shen, Wanqing Wu, Qiquan Zhang, Lu Zhang, Moran Chen, Hao Liu, Rongchao Peng, et al., "Auditory attention decoding from electroencephalography based on long short-term memory networks," *Biomedical Signal Processing and Control*, vol. 70, pp. 102966, 2021.

[8] Cong Han, Vishal Choudhari, Yinghao Aaron Li, and Nima Mesgarani, "Improved decoding of attentional selection in multi-talker environments with self-supervised learned speech representation," *arXiv preprint arXiv:2302.05756*, 2023.

[9] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[11] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[12] Andy T Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, 2017.

[17] Krishna Chaitanya Puvvada and Jonathan Z. Simon, "Cortical representations of speech in a multitalker auditory scene," *The Journal of Neuroscience*, vol. 37, pp. 9189 – 9196, 2017.

[18] Søren A. Fuglsang, Daniel D.E. Wong, and Jens Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018.

[19] Octave Etard and Tobias Reichenbach, "EEG Dataset for 'Decoding of selective attention to continuous speech from the human auditory brainstem response' and 'Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise'.," Sept. 2022.

[20] Yuanming Zhang, Ziyan Yuan, and Jing Lu, "Auditory Attention Detection Dataset Nanjing University," Oct. 2022.

[21] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, no. 267, pp. 1–13, 2013.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] W Biesmans, N Das, T Francart, and A Bertrand, "Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, May 2017, Epub 2016 May 24.

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.