DOI: 10.61782/fa.2025.0224

FORUM ACUSTICUM
EURONOISE 2025

COMPARATIVE ANALYSIS ON THE USE OF LINEAR AND NON-LINEAR METHODS FOR AUDITORY ATTENTION DECODING (AAD)

Iñigo García-Ugarte[1*] **Rubén Eguinoa**[1] **Carmen Vidaurre**[2]
Daniel Paternain[2] **Ricardo San Martín**[1]
[1] Acoustics Laboratory, Public University of Navarre, Spain
[2] Department of Computer Science, Public University of Navarre, Spain
[3] Basque Center on Cognition, Brain and Language, San Sebastian, Spain

ABSTRACT

In a cocktail party scenario, the human auditory system can focus on a single stimulus while suppressing others identified as noise [1]. In the context of neuro-steered hearing devices, auditory attention decoding (AAD) aims to replicate this process using different algorithms that decode electroencephalography (EEG) signals to identify the attended stimulus. Traditional approaches often rely on linear models to establish relationships between neural activity and auditory inputs. However, linear algorithms face significant limitations when decoding a complex non-linear system like the brain. The emergence of deep learning has enabled the development of novel non-linear algorithms, which have shown promising results. In this study, different linear and non-linear algorithms are implemented and evaluated using publicly available data. Furthermore, different methods for training deep learning models are considered to enhance the final model accuracy. The results are analyzed to assess the advantages and limitations of linear versus non-linear approaches in real-world scenarios. This work provides a detailed comparison between different AAD methodologies, offering valuable insights for applications in smart hearing aids, auditory prostheses, and hearing-related medical diagnoses.

Keywords: *Auditory attention decoding (AAD), electroencephalography (EEG), speech envelope, deep learning.*

1. INTRODUCTION

In scenarios where multiple speakers, background noise, or music are present, our auditory system can distinguish the attended speaker from other sources by suppressing or ignoring them [1]. This auditory ability diminishes with age, and individuals with congenital hearing impairment are similarly affected. Current hearing devices and cochlear implants often perform poorly in complex listening environments—such as at a cocktail party, in eavesdropping situations, or while driving—because they typically identify the attended source based on loudness or location. Decoding the desired auditory source directly from the user's brain activity could provide crucial information for neuro-steered hearing devices, thereby enhancing the performance of traditional devices in challenging scenarios.

The electroencephalogram (EEG) is the preferred method for measuring brain activity for auditory attention decoding (AAD) because it is non-invasive, cost-effective, and scalable; features that facilitate its use in everyday settings. In [2,3], the authors established a direct relationship between the EEG signal and the low-frequency envelope of the attended stimulus. When linking the attended stimulus with the brain activity, researchers distinguish between forward and backward models based on whether the stimulus or the EEG signal is predicted. Backward

[*] ***Corresponding author***: *inigo.garciaugarte@unavarra.es.*

models have demonstrated superior performance by predicting the attended speech envelope from the EEG signal [4]. Consequently, this study primarily considers the backward paradigm (see Fig. 1 a)).

Several real-time [5] and real-life [6] AAD implementations have been conducted, and new portable, comfortable EEG devices continue to emerge [6]. These experiments only considered classical approaches—specifically linear methods—which have been regarded as state-of-the-art in terms of efficiency and performance. However, in recent years the advent of deep learning techniques significantly impacted data processing tasks. In the field of AAD, recent studies introducing new non-linear methods [7–9] and more extensive datasets [10] have enabled the development of complex non-linear models that outperform classical methods.

Unlike previous works, our study incorporates innovative deep learning models using the same dataset, allowing for a rigorous analysis of their advantages and limitations compared to traditional methods. This provides a comprehensive evaluation of the current state of auditory attention decoding.

2. AAD ALGORITHMS REVIEW

2.1 Linear models

2.1.1 Linear regression

Linear supervised AAD models facilitate the estimation of the attended stimulus by reconstructing the speech envelope ($\hat{\mathbf{s}}_\mathbf{a}$) applying a decoder matrix on the EEG lagged signal [3]. This reconstruction is achieved by linearly combining time-lagged sequences from the EEG channels, as expressed by:

$$\hat{s}_a(t) = \sum_{c=1}^{C} \sum_{l=0}^{L-1} d_c(l) x_c(t+l) \tag{1}$$

where $x_c(t)$ denotes the value of the c-th EEG channel at time t, $d_c(l)$ represents the decoder coefficient corresponding to the l-th time lag corresponding to the c-th channel, and L and C indicate the total number of time lags and EEG channels, respectively.

Assuming T samples, the decoding coefficients are obtained by minimizing the mean squared error (MSE) between the predicted speech envelope $\hat{\mathbf{s}}_\mathbf{a} = \mathbf{Xd}$ and the actual envelope $\mathbf{s_a}$:

$$\mathbf{argmin}||\mathbf{s_a} - \mathbf{Xd}||_\mathbf{2}^\mathbf{2} \tag{2}$$

with $\mathbf{X} = [x(0)...x(T-1)] \in \mathbb{R}^{T\times LC}$ and $\mathbf{s}(\mathbf{a}) = [s(0)...s(T-1)] \in \mathbb{R}^{T\times LC}$. This leads to the solution $\hat{\mathbf{d}} = (\mathbf{X^TX})^{-\mathbf{1}}\mathbf{X^Ts_a}$ [4], where $\mathbf{R_{xx}} = (\mathbf{X^TX})^{-\mathbf{1}}$ corresponds to the autocorrelation matrix and $\mathbf{R_{xs_a}} = \mathbf{X^Ts_a}$ represents cross-correlation matrix. Thus, the decoder matrix is estimated as: $\hat{\mathbf{d}} \in \mathbb{R}^{LC\times 1}$.

Although these linear models offer simplicity with considerable performance, overfitting remains a common issue in AAD due to the typically limited size of available datasets. To mitigate this, Ridge regression is frequently employed, incorporating to Equation 1 an L2-norm regularization term when computing the decoder matrix: $\lambda\mathbf{z}||\mathbf{d}||^\mathbf{2}$ with $\mathbf{z} = \mathbf{trace}(\mathbf{XX^T})/LC$ [11] Here, λ denotes the regularization hyperparameter that penalizes large decoder weights, and it is determined by selecting an optimal value from a predefined range using a validation procedure.

In this model the reconstructed envelope is correlated with both the attended (ρ_1) and the ignored (ρ_2) real envelopes. The higher correlation coefficient, ρ, is associated with the attended stimulus, providing a measure of the model's accuracy in decoding auditory attention in a multi-speaker scenario (Fig. 1 a)).

2.1.2 Canonical correlation analysis CCA

In section 1, backward models are presented as the top-performing models in Auditory Attention Decoding (AAD). However, Canonical Correlation Analysis (CCA) offers a hybrid approach (see Fig. 1 b)), integrating both forward and backward modeling techniques to enhance decoding performance. This model proposed in [12] to solve the AAD problem, identifies optimal linear transformations for both EEG signals and speech envelopes, thereby minimizing irrelevant variance and maximizing mutual correlation between the transformed domains.

Mathematically, CCA seeks to determine a set of backward spatiotemporal filters, denoted as $\mathbf{w_x} \in \mathbb{R}^{LC\times 1}$, and forward temporal filters, $\mathbf{w_{s_a}} \in \mathbb{R}^{L_a\times 1}$, applied to the EEG and stimulus domains, respectively. These filters maximize the correlation:

$$\max_{\mathbf{w_x},\mathbf{w_{s_a}}} \frac{\mathbf{w_x^T R_{xs_a} w_{s_a}}}{\sqrt{\mathbf{w_x^T R_{xx} w_x}}\sqrt{\mathbf{w_{s_a}^T R_{s_as_a} w_{s_a}}}} \tag{3}$$

This optimization is typically solved via generalized eigenvalue decomposition, where the optimal filters correspond to the eigenvectors with the largest eigenvalues. By selecting the top J eigenvectors, we obtain filter matrices $\mathbf{W_x} \in \mathbb{R}^{LC\times J}$ and $\mathbf{W_{s_a}} \in \mathbb{R}^{L_a\times J}$. As depicted

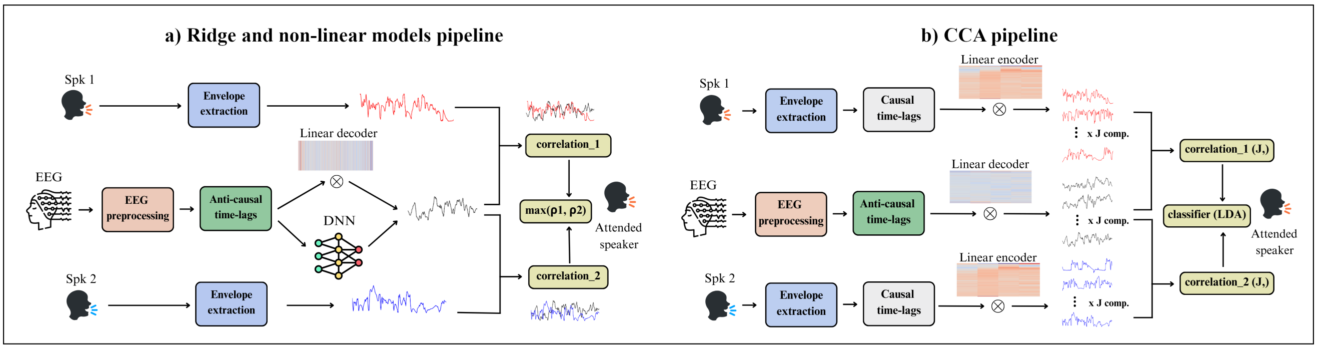

Figure 1. Different analytical pipelines followed for Auditory Attention Detection (AAD). (a) Implements only a backward decoder for Ridge regression and non-linear models. (b) Incorporates both a backward decoder and a forward encoder for Canonical Correlation Analysis (CCA), extracting multiple components.

in Fig. 1 b) for the classification task, these J correlation coefficients, both the ignored and the attended, serve as features for linear discriminant analysis (LDA) classifiers. Further details regarding specific model parameters and implementation can be found in Section 4.

2.2 Non linear models

2.2.1 CNN

In [8], two deep-learning-based backward models (map EEG signal to stimulus) are proposed. These algorithms predict the ongoing sample of the input sequence, operating in an anti-causal fashion. Among the two models presented in the study, we selected the Convolutional Neural Network (CNN) model, based on the original network EEGNet [13], as it proved to be a more efficient network when compared with the alternative Fully-Connected Neural Network (FCNN) solution. This model applies convolution operations along both temporal and channel dimensions to extract features from the EEG. After that, a depth-wise separable convolution is employed to capture global features, which are fed into a linear classifier to predict the initial sample of attended stimulus.

2.2.2 EEG Conformer

Recognizing that the CNN model is relatively simple, we explored more complex architectures in this study. The network proposed in [9], known as EEG Conformer, is considered state-of-the-art for EEG classification, having achieved superior results on motor-imagery (MI) and emotion detection datasets. This network integrates some Convolutional modules from EEGNet with a transformer self-attention block, thereby combining local feature extraction with global context modeling. For adaptation to the auditory attention decoding (AAD) paradigm, the network was modified to predict the ongoing segment of the envelope, as for the CNN approach, and an extensive hyperparameter search was conducted.

2.2.3 VLAAI

The study in [7] describes a convolutional network architecture, VLAAI, which was evaluated on the SparrKULee dataset [10] and outperformed both CNN (see section 2.2.1) and linear models. The architecture comprises N distinct blocks, each containing a CNN module with M consecutively stacked convolutional layers. The model also includes an output context module that applies a left zero padding to the convolutional operation. Unlike the previous non-linear models, which predict only the ongoing segment of the sequence, VLAAI predicts the entire window at each forward pass. Moreover, since the dataset used differed from that used in the original study, we performed an independent parameter search.

3. DATA: DTU DATASET

This work relies on a well-established dataset, known as the DTU dataset [14], to train and evaluate the models described in previous sections. The dataset comprises recordings from 18 subjects who listened to one of two competing audio streams, each featuring a male and a female narrator reading a book in Dutch. All subjects were young, normal-hearing individuals, and the recordings were obtained under various simulated reverberation scenarios. A 64-channel BioSemi ActiveTwo system sampled at a frequency of 512 Hz is employed to obtain the EEG signals.

The extraction of the speech envelope was conducted using the `COCOHA MATLAB toolbox` [15]. The function `co_auditoryfilterbank.m` is used to implement a Gamma-tone filter bank on the unprocessed audio signal, sampled at $f_s = 44100Hz$. The signal is decomposed into 31 frequency bands, with center frequencies equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale, ranging from 80 Hz to 8000 Hz. The envelope of each sub-band is then computed by taking the absolute value of the filtered signal and applying a nonlinear compression by raising it to the power of 0.3. Subsequently, the audio is down-sampled to align with the EEG sampling rate ($f_s = 64Hz$), and all frequency components are aggregated to compute the average and establish the final envelope.

The `COCOHA MATLAB toolbox` was also utilized to preprocess the EEG signal for the Auditory Attention Decoding (AAD) task. The signal is down-sampled to $f_s = 64Hz$ and subjected to a second-order Butterworth filter operating within the frequency range of $f_l = 0.5Hz$ to $f_h = 32Hz$. For further details on EEG preprocessing and artifact removal, refer to the preprocessing pipeline implemented in `preproc_data.m`, as described in [14].

4. METHODS

4.1 Validation procedure

AAD datasets contain subject-specific information recorded across different trials. Including data from the same trial in both training and validation could lead to misleading results, as models might learn trial-specific patterns, causing overfitting [16]. To prevent this, we implement a cross-trial validation procedure, distinguishing between three different approaches: subject-specific (SS), subject-independent (SI) and population procedures.

For subject-specific decoders (SS), models are trained and evaluated using data from the same individual, resulting in one model per subject. Given the limited availability of subject-specific data, we applied a 5-fold cross-validation strategy. The total number of trials was divided into five equal sets (e.g.: 12 trials per set in the DTU dataset). One set was used for validation, another for testing, and the remaining three for training.

For subject-independent decoders (SI), we used a leave-one-subject-out (LOSO) validation approach. The test set consisted of a single subject, while the remaining subjects were included in the training and validation sets. Five randomly selected subjects were assigned to the validation set and the rest of them formed the training set. A single model per subject is obtained when using this validation paradigm.

In addition to these validation strategies, we implemented a population-level baseline model, trained on data from all subjects. In this case, we applied the same 5-fold cross-validation strategy as in the subject-specific approach to enhance robustness. This resulted in a single model per dataset per fold. The population model served as a baseline for the subject finetuning (SF) training strategy (see Section 4.3).

4.2 Evaluation metrics

4.2.1 *Decoding accuracy*

The accuracy obtained by the model is calculated by comparing the Pearson r coefficients of the attended and unattended stimuli. This Pearson correlation coefficient r quantifies the similarity between two temporal sequences ranging from -1 to 1, where 1 indicates maximum correlation and 0 denotes no correlation. The envelope corresponding to the higher coefficient is identified as the attended one. Classification accuracy is determined as the ratio of correctly predicted windows to the total number of windows. Accuracy depends on window size, as larger windows incorporate more samples, increasing the likelihood of correct classification. Six window sizes were evaluated, ranging from 1s to 50s, the last one corresponding to the trial length of the DTU dataset.

4.2.2 *MESD*

Decoding accuracy, as previously mentioned, is a length-dependent metric that yields varying results based on the evaluation window. In [17], a unique metric is introduced to assess overall model performance. This metric offers insight into the model's potential performance in a real hearing-aid device by quantifying the minimal-expected switch duration (MESD) using a Markov chain approach. It is measured in seconds, and a lower value is preferable for the model, as it indicates the minimum expected duration required for an attention shift. Its computation relies on an adaptive gain system that depends on a hypothetical number of gain levels. All calculations were performed using the `MESD-toolbox` implemented in Python [18], which processes classification window results and returns an estimate for the duration of a hypothetical attention switch.

4.3 Training process

In linear decoders, time-lag matrices were computed by fixing the time-lag value to $L = 26$ in LSR, which left-shifts the EEG signal by up to 400 ms. For CCA, we set $L_a = 80$ for the encoder and $L = 16$ for the decoder, corresponding to a right-shift of the stimulus by up to 1.25 s and a left-shift of the EEG by up to 250 ms. These time-lag values were chosen based on previous studies [3,4]. To determine the optimal number of components J for CCA, we first trained a CCA model with $\min(L_a, L)$ components and then optimized an LDA classifier via grid-search over J (ranging from 1 to $\min(L_a, L)$); the J yielding the best validation performance was selected. Similarly, for Ridge regression, a grid-search was performed for the regularization parameter λ over the range 10^{-7} to 10^7. Both linear models were implemented using the scikit_learn library, specifically `cross_decomposition.CCA` and `linear_model.Ridge` functions.

For non-linear models, the optimization objective was to minimize the negative correlation coefficient. The Adam optimizer was employed and models were trained for up to 200 epochs with early stopping after 5 epochs without improvement in validation loss. A hyperparameter search was conducted for the VLAAI and Conformer models, as they were originally designed for different tasks and datasets. All deep learning models were implemented in PyTorch using an NVIDIA A600 GPU.

Two training strategies were adopted to obtain the best subject-specific model. For non-linear models, we compared performance when training the model from scratch (see SS in 4.1) versus fine-tuning a pre-trained baseline model (see population in 4.1), where the final classification layers were adapted to subject-specific data while the rest of the model remained with the same parameters. A parametric paired test with Bonferroni correction was used for the statistical analysis.

5. RESULTS

5.1 Hyper-parameter search

To ensure a fair comparison among the reviewed models, we performed a hyperparameter search for all models except the CNN, which was already optimized in [8] for the dataset used.

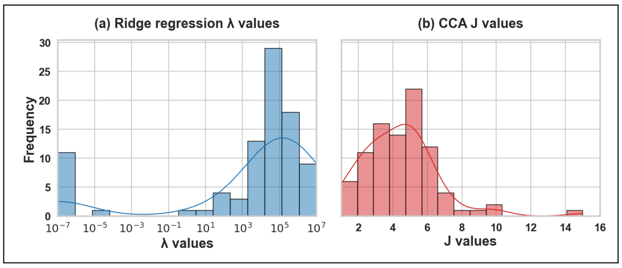

Figure 2. Linear models parameter distribution: a) Ridge regression λ values b) CCA J values

Fig. 2 a) illustrates the λ values for the Ridge regression model during subject-specific validation, selected based on validation loss. Most models yielded high regularization values that penalize large weights.

Fig. 2 b) displays the optimal J values for the CCA models under subject-specific validation. Typically, the best J values involve seven or fewer components, indicating that additional components are unnecessary.

For the non-linear models originally trained on different datasets, we conducted a hyperparameter search on the VLAAI and Conformer models. Tab. 1 summarizes the differences in model sizes and validation performance, with results averaged across folds. Note that the window sizes differ: the VLAAI model was trained on 5-second windows, while the Conformer and CNN models predicted 2-second windows on the training stage.

Table 1. Model sizes and validation results for non-linear models

Model	Size	Loss (ρ)	Accuracy (%)
CNN [8]	**9.65K**	**0.150**	**61.32**
VLAAI [7]	1.71M	0.058	56.40
VLAAI (enhanced)	**1.71M**	**0.069**	**57.08**
Conformer [9] (adapted)	241K	0.139	60.34
Conformer (enhanced)	**508K**	**0.144**	**60.62**

5.2 Subject-finetuning (SF)

Each population model was fine-tuned with subject-specific data to compare two training methodologies: subject-specific versus subject-fine-tuned (see Section 4.3). As shown in Fig. 3, except for VLAAI, fine-tuning did not improve validation accuracy, with Conformer and CNN models showing no significant differences between the two strategies. Therefore, we will use subject-specific validation for these models. In contrast, the VLAAI model exhibited a significant improvement with fine-tuning ($p < 0.01$), consistent with [7]. Hence, this strategy will be applied to VLAAI in the subsequent evaluation.

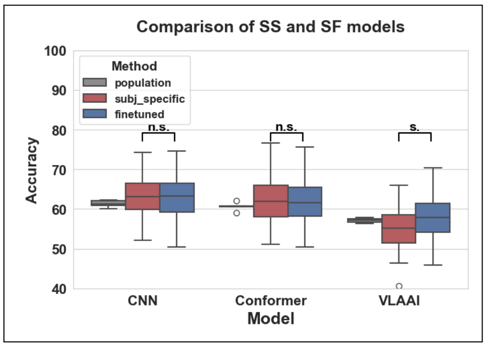

Figure 3. Subject specific and subject finetuned comparison based on the validation accuracy. Statistical analysis: **n.s.** : $p > 0.01$; **s.** : $p < 0.01$

5.3 Subject-specific (SS) and Subject-independent (SI)

Fig. 4 a) presents the subject-specific evaluation results, comparing accuracies across different window lengths and corresponding inter-model MESD values. The CNN and Conformer models achieve the best median MESD values, at 16.7 and 17.4 respectively. Although the CCA model appears to perform better for long windows, for short decision windows (1–2 seconds) the CNN and Conformer models outperform the others, thereby significantly influencing the overall MESD. Conversely, for long windows the CCA model performs better, reaching 92.2% at 50s, also producing fewer MESD outliers. The Ridge regression model performed substantially worse in both MESD and accuracy, while the VLAAI model performed close to the singnificance level, indicating its unsuitability for this dataset.

The results for subject-independent methodology, which renders the system ready for use (see Section 4.1), are illustrated in Fig. 4 b). As expected, subject-independent models performed significantly worse than subject-specific ones; the VLAAI model was excluded since it did not exceed the significance level. Nevertheless, the remaining models achieved above-chance, and even acceptable accuracies with long decision windows. The CCA model attained the best performance for long windows, reaching an accuracy of 83.6% on a 50-second window, whereas the CNN model—excelling in short decision windows—delivered the best MESD at 33.3 seconds. These findings indicate that the models can generalize across subjects when subject-specific data is unavailable.

6. DISCUSSION

6.1 Linear vs. non-linear models

Results in Section 5.3 show that the best linear model (CCA) and the CNN performed similarly in terms of MESD. However, these models differ in accuracy across different window lengths, highlighting the MESD metrics sensitivity to short-time windows. We attribute the superior performance of these models to two factors. First, CCA leverages both stimulus and EEG information to capture relevant variance from each domain [12]. Second, the performance of the CNN and Conformer nonlinear models can be attributed to their ability to adapt to a relatively small dataset, such as the DTU dataset.

In many machine learning applications, the trend is to develop larger, more complex models trained on extensive datasets. Following this approach, a non-linear model that integrates both stimulus and EEG signals could potentially yield even better performance. Nonetheless, when selecting an algorithm for AAD, linear models offer clear advantages: they provide lightweight solutions and simpler, more interpretable pipelines. In many cases, these characteristics may be prioritized over marginal gains in final performance.

6.2 Non-linear models prediction

In Section 5.2, the only model that benefited from subject fine-tuning was VLAAI. However, this model exhibited the poorest overall performance, in contrast to [7]. We attribute this discrepancy primarily to the window prediction paradigm, as other linear models predicting the ongoing sample performed much better. Furthermore, the DTU dataset [14] differed substantially in both experimental design and size from that in [7], which further explains the under-performance of VLAAI in Section 5.3. In contrast, the CNN and Conformer nonlinear models adequately fit the dataset, achieving the best results in terms of MESD and accuracy for small window sizes. When comparing these two models, despite expectations regarding model

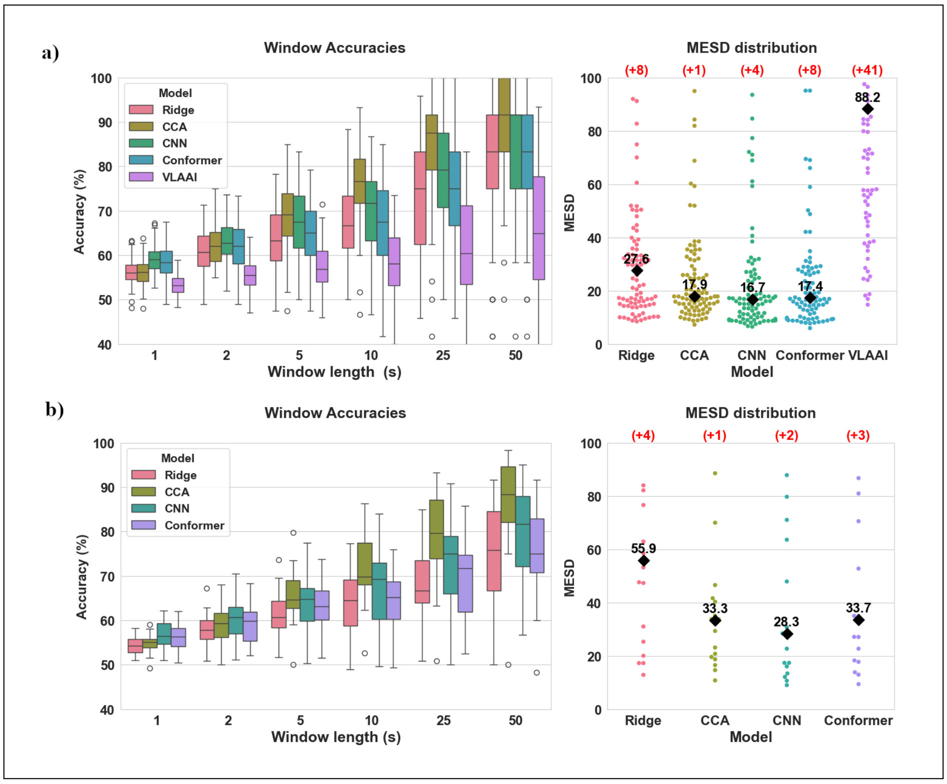

Figure 4. a) Subject-specific (SS) model comparison, b) Subject-independent (SI) model comparison.

size and complexity, empirical results indicate that the CNN outperformed the Conformer model. This results suggest that a depth-wise block is sufficient to learn from EEG signals, rather than relying on a self-attention layer.

6.3 AAD feasibility and actual constraints

The reviewed models demonstrated robust performance for the AAD paradigm. However, when implementing an AAD system in real-life scenarios, several constraints must be considered. First, all algorithms rely on access to clean envelopes, which requires an effective speech separation algorithm. Second, the data in this study were recorded using a wet 64-electrode EEG system in a controlled environment, a setting that does not reflect real-world conditions. Another critical factor is the computational cost and real-time feasibility, both aspects not addressed by the pipelines developed in this work.

Ultimately, accurately measuring attention remains challenging. In experimental settings, attention is typically inferred from participant's responses to a series of questions after completing the trials. However, in real-world situations, attention fluctuates dynamically and cannot be directly measured, making it fundamentally different from controlled experimental assessments. Thus, future studies must search for an adequate trade-off between window length and accuracy and incorporate an adaptive gain system that allows users to switch attention.

Although significant progress has been made to overcome these issues [5, 6], further research is essential to transform neuro-steered hearing aids into a practical solution.

7. CONCLUSION

In this study, different linear and non-linear auditory-attention decoding (AAD) models were presented and evaluated using the same dataset. All models were adjusted to the dataset through a hyperparameter search to obtain the optimal configuration. For non-linear models, a subject fine-tuning training strategy was considered. The CNN model of [8], performed the best on short decision windows achieving the lowest MESD on both SS and SI models by predicting the ongoing sample of the introduced context. Nevertheless, the CCA linear model achieved higher accuracies on long decision windows and similar MESD results, offering a simpler and more interpretable alternative to non-linear models. In conclusion, this work provides a fair comparison between different AAD algorithms using the same dataset and may serve as a basis for future studies evaluating alternative AAD approaches. It also contributes to further research by linking brain activity with auditory stimuli, thereby providing useful insights for neuro-steered hearing aids and hearing-related medical diagnoses.

8. REFERENCES

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, June 2005.

[2] S. J. Aiken and T. W. Picton, "Human Cortical Responses to the Speech Envelope," *Ear & Hearing*, vol. 29, pp. 139–157, Apr. 2008.

[3] J. A. O'Sullivan, A. J. Power, and Mesgarani, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, pp. 1697–1706, July 2015.

[4] S. Geirnaert and Vandecappelle, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, pp. 89–102, July 2021. Conference Name: IEEE Signal Processing Magazine.

[5] J. Hjortkjær and Wong, "Real-time control of a hearing instrument with EEG-based attention decoding," *Journal of Neural Engineering*, vol. 22, p. 016027, Feb. 2025. Publisher: IOP Publishing.

[6] L. Straetmans, K. Adiloglu, and S. Debener, "Neural speech tracking and auditory attention decoding in everyday life," *Frontiers in Human Neuroscience*, vol. 18, Nov. 2024. Publisher: Frontiers.

[7] B. Accou and Vanthornhout, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Scientific Reports*, vol. 13, p. 812, Jan. 2023. Publisher: Nature Publishing Group.

[8] M. Thornton and Mandic, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of Neural Engineering*, vol. 19, p. 046007, Aug. 2022.

[9] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.

[10] L. Bollens, B. Accou, H. Van hamme, and T. Francart, "SparrKULee: A Speech-evoked Auditory Response Repository of the KU Leuven, containing EEG of 85 participants," 2023.

[11] D. D. E. Wong and Fuglsang, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, p. 531, Aug. 2018.

[12] A. de Cheveigné and Wong, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.

[13] V. J. Lawhern and Solon, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[14] S. A. Fuglsang and Wong, "EEG and audio dataset for auditory attention decoding," Mar. 2018.

[15] D. D. Wong and Hjortkjær, "COCOHA Matlab Toolbox," Mar. 2018.

[16] I. Rotaru and Geirnaert, "What are wereallydecoding? Unveiling biases in EEG-based decoding of the spatial focus of auditory attention," *Journal of Neural Engineering*, vol. 21, Feb. 2024.

[17] S. Geirnaert and Francart, "An Interpretable Performance Metric for Auditory Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, pp. 307–317, Jan. 2020. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[18] EXPORL, "Mesd-toolbox: A toolbox for model-based evaluation of speech detectors." `https://github.com/exporl/mesd-toolbox`, 2024. GitHub repository.