Performance Modeling for Correlation-based Neural Decoding of Auditory Attention to Speech

Simon Geirnaert
Dept. of Electrical Engineering
KU Leuven
Leuven, Belgium
simon.geirnaert@kuleuven.be

Jonas Vanthornhout
Dept. of Neurosciences
KU Leuven
Leuven, Belgium
jonas.vanthornhout@kuleuven.be

Tom Francart
Dept. of Neurosciences
KU Leuven
Leuven, Belgium
tom.francart@kuleuven.be

Alexander Bertrand
Dept. of Electrical Engineering
KU Leuven
Leuven, Belgium
alexander.bertrand@kuleuven.be

***Abstract*—Correlation-based auditory attention decoding (AAD) algorithms exploit neural tracking mechanisms to determine listener attention among competing speech sources via, e.g., electroencephalography signals. The correlation coefficients between the decoded neural responses and encoded speech stimuli of the different speakers then serve as AAD decision variables. A critical trade-off exists between the temporal resolution (the decision window length used to compute these correlations) and the AAD accuracy. This trade-off is typically characterized by evaluating AAD accuracy across multiple window lengths, leading to the performance curve. We propose a novel method to model this trade-off curve using labeled correlations from only a single decision window length. Our approach models the (un)attended correlations with a normal distribution after applying the Fisher transformation, enabling accurate AAD accuracy prediction across different window lengths. We validate the method on two distinct AAD implementations: a linear decoder and the non-linear VLAAI deep neural network, evaluated on separate datasets. Results show consistently low modeling errors of approximately 2 percent points, with 94% of true accuracies falling within estimated 95%-confidence intervals. The proposed method enables efficient performance curve modeling without extensive multi-window length evaluation, facilitating practical applications in, e.g., performance tracking in neuro-steered hearing devices to continuously adapt the system parameters over time.**

***Index Terms*—auditory attention decoding, stimulus-response correlations, performance modeling**

I. Introduction

Current hearing devices often struggle in cocktail party scenarios with multiple competing conversations. Selective auditory attention decoding (AAD) addresses this by determining the attended speech signal from neural recordings such as electroencephalography (EEG) [1], [2]. AAD has, therefore, important applications in so-called neuro-steered hearing devices, allowing them to automatically steer beamformers or source separation algorithms towards the attended conversation [3].

This work was supported by a junior postdoctoral fellowship fundamental research of the FWO (for S. Geirnaert, No. 1242524N), FWO project nr. G081722N, the Flemish Government (AI Research program), Internal Funds KU Leuven (project IDN/23/006), and the European Research Council (ERC) under the European Union's research and innovation programme (No 101138304). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or any of the granting authorities. Neither the European Union nor the granting authorities can be held responsible for them. S. Geirnaert, J. Vanthornhout, T. Francart, and A. Bertrand are affiliated to Leuven.AI - KU Leuven institute for AI, Leuven, Belgium.

Various AAD algorithms leverage the neural tracking phenomenon [2], [3], referring to the synchronization of the listener's neural signals with features of the attended speech signal [1]. As illustrated in Fig. 1, these algorithms employ a (linear or non-linear) neural decoder and/or stimulus encoder to generate representations that correlate well. The strength of this relationship is then quantified using the Pearson correlation coefficient, which serves as a decision variable to determine attention to one of multiple speech sources.

A critical parameter in AAD algorithms is the *decision window length*. In neural tracking algorithms, this corresponds to the number of samples used to compute the correlation coefficient. Therefore, this parameter determines the time resolution at which attention decisions are made. However, there exists a fundamental trade-off: shorter windows improve the temporal resolution but typically reduce accuracy. This trade-off is captured in the algorithm's performance curve, showing AAD accuracy against decision window length [4].

Quantification of this performance curve is crucial for optimizing AAD-based gain control systems by determining the optimal decision window length. This is done, for example, based on the minimal expected switch duration [4], which is a metric that uses the performance curve as an input. Currently, generating this performance curve requires evaluating the AAD algorithm across *multiple* decision window lengths. However, this approach becomes impractical for efficiently tracking the performance curve in real time, e.g., to adapt gain control system parameters in a neuro-steered hearing device over time [5].

To address this, we propose a new method to model the complete performance curve of correlation-based AAD algorithms using (known or measured) correlation statistics for only *one* decision window length. This enables continuous and efficient monitoring of the trade-off between accuracy and decision window length, without having to evaluate the model for all possible window lengths. Furthermore, the proposed technique provides a theoretical model to inform new correlation-based AAD algorithm design by translating target accuracy requirements into corresponding correlation coefficient targets, and vice versa. As this is an accuracy evaluation framework, we assume ground-truth attention labels are known, and leave the integration with unsupervised accuracy

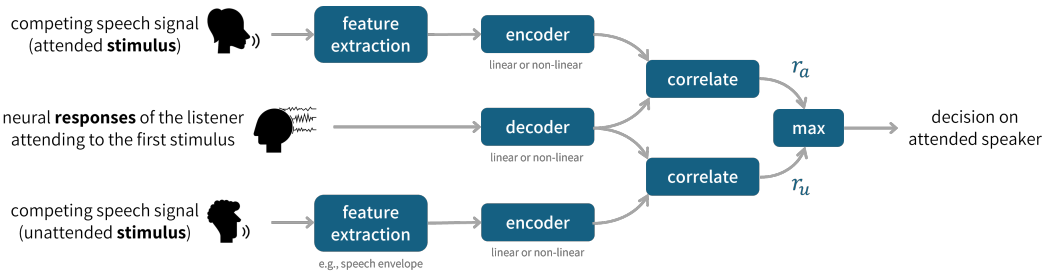

Fig. 1. In this paper, we start from correlation-based AAD algorithms that use a stimulus encoder and/or neural decoder of which the outputs are correlated to determine the attended speaker.

estimation as in López-Gordo et al. [6] for future work.

The proposed method is explained in Section II. Section III describes the experiment setup, including the datasets and AAD algorithms used for validating the proposed method. Results and discussion are provided in Section IV.

II. Performance Curve Modeling Based on One Decision Window Length

A. Correlation-based AAD algorithms

In this paper, we start from correlation-based AAD algorithms that use neural tracking, as shown in Fig. 1. These algorithms employ both stimulus encoders and neural response (EEG) decoders, which can be implemented using either linear or non-linear approaches (see Section III for specific examples).

For simplicity, we consider the case of two competing speakers, for which the AAD pipeline compares correlation coefficients between the decoded response and each encoded competing speech signal, identifying the highest correlation as the attended speaker. This approach is based on the fundamental assumption that the expected correlation with the attended speaker is higher than the expected correlation of the unattended speaker(s).

B. Modeling of correlations in the AAD decision system

In this paper, we evaluate the AAD accuracy assuming the ground-truth labels are known, as opposed to the unsupervised accuracy estimation technique in López-Gordo et al. [6]. For a given window length w and sampling frequency f_s, the measured correlation coefficient r_a over $N = wf_s$ samples between the (zero-mean) decoded response $\mathbf{x} \in \mathbb{R}^N$ and encoded attended speech signal $\mathbf{y}_a \in \mathbb{R}^N$ corresponds to:

$$r_a = \frac{\sum_{n=1}^{N} x_n y_{a,n}}{\sqrt{\sum_{n=1}^{N} x_n^2}\sqrt{\sum_{n=1}^{N} y_{a,n}^2}}, \tag{1}$$

with $\rho_a = \mathbb{E}\{r_a\}$ the expected attended correlation (and similarly for the unattended speech $\mathbf{y}_u$, resulting in the (expected) unattended correlations r_u and ρ_u).

Previous work [7] has shown that stimulus-response correlations in neural tracking such as $r_{a/u}$ can then be effectively modeled using a normal distribution after applying the Fisher transformation (i.e., using an inverse hyperbolic tangent) [8]:

$$z_{a/u} = \operatorname{artanh}\left(r_{a/u}\right) \sim \mathcal{N}\left(\mu_{a/u}, \sigma^2_{a/u}\right).$$

This modeling approach has been validated in the context of neural tracking to single speakers [7] and is further supported by our experimental results in Section III.

Since the AAD decision rule compares per-window attended and unattended correlations (r_a and r_u), and given that the inverse hyperbolic tangent is *monotonically increasing*, we can use $\max(z_a, z_u)$ instead of $\max(r_a, r_u)$ in Fig. 1 to make the binary decision on attention. Under the assumption that r_a and r_u are uncorrelated (as empirically validated in López-Gordo et al. [6]), we can use $z_a - z_u \sim \mathcal{N}\left(\mu_a - \mu_u, \sigma_a^2 + \sigma_u^2\right) \lessgtr 0$ as our decision variable. The AAD accuracy can then be computed using the cumulative distribution function (CDF) $F(\cdot)$ of the normal distribution, evaluated in zero [6]:

$$\text{AAD accuracy} = 100\left(1 - F(0; \mu_a - \mu_u, \sigma_a^2 + \sigma_u^2)\right). \tag{2}$$

To determine the parameters required for computing the AAD accuracy in (2), we employ the first-order Hotelling series expansion to approximate the mean and variance of the Fisher-transformed variables $z_{a/u}$ [9], [10]:

$$\mu_{a/u} \approx \operatorname{artanh}\left(\rho_{a/u}\right) + \frac{\rho_{a/u}}{2(N-1)}, \sigma^2_{a/u} \approx \frac{1}{N-1},$$

where $N = wf_s$ represents the number of samples used in the correlation computation in (1), derived from the decision window length w and sampling frequency f_s. With this, the parameters of the decision variable distribution $\mathcal{N}\left(\mu_a - \mu_u, \sigma_a^2 + \sigma_u^2\right)$ can be approximated as:

$$\begin{cases} \mu_a - \mu_u \approx \operatorname{artanh}(\rho_a) - \operatorname{artanh}(\rho_u) + \dfrac{\rho_a - \rho_u}{2(N-1)}, & (3) \\ \sigma_a^2 + \sigma_u^2 \approx \dfrac{2}{N-1}. & (4) \end{cases}$$

C. Modeling the entire performance curve

A fundamental insight underpinning our method is that the mean of the measured correlation values $r_{a/u}$ (represented by $\rho_{a/u}$) remains constant across different decision window lengths, as demonstrated in [6]. Only the variance of $r_{a/u}$ changes with window length w. Leveraging this property, we

can derive the relationship between parameters at different window lengths. Let w_1 be the baseline window length for which we effectively compute the $r_{a/u}$ values and w_2 the target window length. By subtracting $\mu_a^{(1)} - \mu_u^{(1)}$ at baseline window length w_1 from $\mu_a^{(2)} - \mu_u^{(2)}$ at w_2 using (3), we obtain:

$$\mu_a^{(2)} - \mu_u^{(2)} \approx \mu_a^{(1)} - \mu_u^{(1)} + \frac{(N_2 - N_1)(\rho_a - \rho_u)}{2(N_2 - 1)(N_1 - 1)}, \quad (5)$$

where $N_1 = w_1 f_s$ and $N_2 = w_2 f_s$. Both $\mu_a^{(1)} - \mu_u^{(1)}$ and $\rho_{a/u}$ can be estimated from the available correlations $r_{a/u}$ at the initial window length w_1.

Similarly, the variance at the target window length w_2 can be derived by dividing (4) between the two window lengths:

$$\sigma_a^{(2)^2} + \sigma_u^{(2)^2} \approx \left(\sigma_a^{(1)^2} + \sigma_u^{(1)^2}\right) \frac{N_1 - 1}{N_2 - 1}. \quad (6)$$

This relationship confirms the intuition that larger window lengths lead to reduced variance in the decision variable, as more samples are available for correlation computation.

Eq. (5) and (6) provide the necessary correction factors to extrapolate the parameters of the Fisher-transformed normal distribution from one decision window length to another. This enables modeling of the entire performance curve from correlations measured at a single window length. To quantify the uncertainty in the predictions of the accuracies, we employ the bias-corrected and accelerated bootstrapping method [11] to compute 95%-confidence intervals for the estimated AAD accuracy. 1000 bootstrap samples are created. Algorithm 1 presents a complete summary of our proposed method, and all MATLAB code can be found online [12].

Algorithm 1 Accuracy prediction for performance curve modeling in correlation-based AAD algorithms

Input: M labeled correlation coefficients $\{r_{a/u}^{(m)}\}_{m=1}^{M}$ at decision window length w_1 resulting from a correlation-based AAD algorithm as in Fig. 1, sampling frequency f_s, target decision window length w_2

Output: AAD accuracy at w_2

1: Compute $\rho_{a/u} = \underset{\text{over } m}{\text{mean}}(r_{a/u}^{(m)})$

2: Apply the Fisher transformation $z_{a/u}^{(m)} = \operatorname{artanh}\left(r_{a/u}^{(m)}\right)$ and compute from the given correlations

$$\begin{cases} \mu_a^{(1)} - \mu_u^{(1)} = \underset{\text{over } m}{\text{mean}}(z_a^{(m)} - z_u^{(m)}), \\ \sigma_a^{(1)^2} + \sigma_u^{(1)^2} = \underset{\text{over } m}{\text{variance}}(z_a^{(m)} - z_u^{(m)}) \end{cases}$$

3: Extrapolate parameters to window length w_2, with $N_1 = w_1 f_s, N_2 = w_2 f_s$:

$$\begin{cases} \mu_a^{(2)} - \mu_u^{(2)} = \mu_a^{(1)} - \mu_u^{(1)} + \frac{(N_2 - N_1)(\rho_a - \rho_u)}{2(N_2 - 1)(N_1 - 1)}, \\ \sigma_a^{(2)^2} + \sigma_u^{(2)^2} = \left(\sigma_a^{(1)^2} + \sigma_u^{(1)^2}\right) \frac{N_1 - 1}{N_2 - 1}. \end{cases}$$

4: Predict AAD accuracy at w_2 using the CDF $F(\cdot)$ of the normal distribution at 0 with the extrapolated parameters:

$$\text{accuracy} = 100\left(1 - F(0; \mu_a^{(2)} - \mu_u^{(2)}, \sigma_a^{(2)^2} + \sigma_u^{(2)^2})\right)$$

III. Experiment setup

To demonstrate the broad applicability of our performance curve modeling method, we validate it using two variants of the correlation-based AAD pipeline shown in Fig. 1: a linear decoder [2], [3] and the non-linear VLAAI deep neural network [13]. Each variant is evaluated on a different dataset for generalization purposes.

A. Linear decoding on Das-2016 dataset

The first setup uses the commonly used linear least-squares decoder (without encoder) to reconstruct the attended speech envelope in the $1 - 9\,\text{Hz}$ frequency range [2], [3]. This decoder is a spatio-temporal filter that integrates EEG channels across time lags to minimize the squared error between the reconstructed and attended speech envelopes. We follow all preprocessing and implementation details from [5], with the sole modification of using a $[100, 400]$ms integration window.

This AAD algorithm is evaluated on the publicly available Das-2016 dataset, which contains EEG recordings from 16 participants attending to one of two competing speakers for $72\,\text{min}$ [14]. We compute the attended and unattended correlations using user-specific 5-fold cross-validation on $60\,\text{s}$ trials. Ground-truth AAD accuracies and (un)attended correlations are computed across decision window lengths of $\{60, 30, 20, 10, 5, 1\}$s. To analyze the impact of the amount of estimation data, we vary the amount of data used ($\{2, 5, 10, 30, 60\}$ min) by randomly sampling from the available test correlations, which is repeated ten times per condition. This AAD algorithm and evaluation serves as the default experimental setup in Section IV unless specified otherwise.

B. VLAAI deep neural network on Fuglsang-2018 dataset

The second independent validation employs the non-linear VLAAI deep neural network, which is pre-trained on a large single-speaker dataset to reconstruct the $1 - 32\,\text{Hz}$ speech envelope [13]. We apply this network as a user-independent AAD decoder to the publicly available Fuglsang-2018 dataset, containing EEG recordings from 18 participants attending to one of two competing talkers, with 60 trials of $50\,\text{s}$ per participant [15]. Data preprocessing follows [13], except for the multi-channel Wiener filter which is replaced by regressing out the EOG from EEG channels for artifact removal. Ground-truth AAD accuracies and correlations are computed for decision window lengths of $\{50, 25, 10, 5, 1\}$s.

IV. Results and discussion

We evaluate our performance modeling method to estimate the average performance curve (Section IV-A), investigate the impact of the amount of estimation data (Section IV-B), and show per-participant performance curve estimation (Section IV-C). All experiments use the linear decoder on the Das-2016 dataset (Section III-A), while evaluation on the VLAAI setup is provided in Section IV-D.

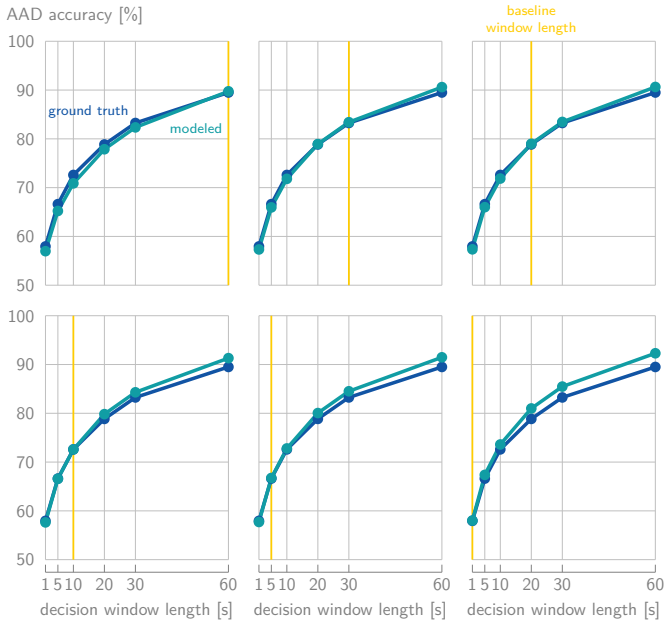

Fig. 2. The modeled (predicted) vs. ground-truth average performance curves of the linear decoder on the Das-2016 dataset, starting from various baseline window lengths indicated in yellow and using 60 min of estimation data.

A. Average performance curve estimation

We first assess the method's ability to model performance curves across all participants. Fig. 2 compares predicted and ground-truth average performance curves (by averaging the per-participant modeled performance curves) for different baseline decision window lengths, using 60 min of randomly sampled estimation data (ten repetitions).

The first overall observation is that the modeled performance curves generally closely approximate the ground truth. As expected, optimal prediction accuracy is observed at the corresponding baseline window length, confirming the modeling of the AAD decision variable using a normal distribution after the Fisher transformation. Prediction error increases with the extrapolation distance between baseline and target window lengths. The baseline window length can, therefore, be tailored towards the range of window lengths of interest (for example, the shorter ones in the context of AAD gain control systems [4]). In general, a good choice is in the middle of the considered range of window lengths. Here, based on Fig. 2, we select 20 s as the baseline for subsequent analyses.

B. Amount of estimation data

Using the 20 s baseline window length, we analyze prediction performance across varying amounts ($\{2, 5, 10, 30, 60\}$ min) of estimation data. Per amount of estimation data, a corresponding number of random 20 s-window correlations are sampled, which is repeated ten times. Fig. 3 shows the mean absolute percent point (pp) error between modeled and true AAD accuracy across the average dataset-wide performance

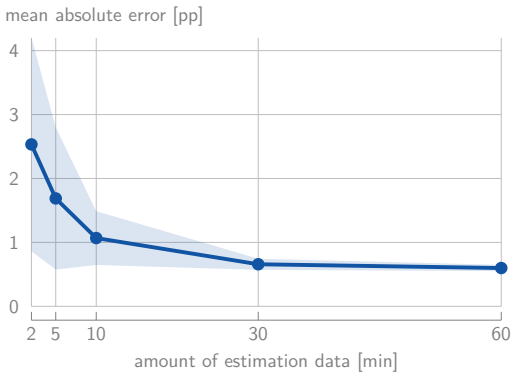

Fig. 3. The mean ($\pm$ standard deviation) absolute error in percent point (pp) between the modeled and true AAD accuracy across the average performance curve of the linear decoder on the Das-2016 dataset in fucntion of the amount of estimation data.

curve, with uncertainty bands indicating standard deviation across the ten repetitions.

As expected, the prediction error and its variability increase when less estimation data are available, yet remains reasonably low (≈ 2.5 pp) even with only 2 min of estimation data. The prediction error stabilizes beyond 30 min, which we adopt in subsequent analyses. In practical applications, however, the AAD accuracy and correlation distributions can change over time due to variations in the EEG signal quality, user performance, or when time-adaptive decoders are used [5]. Such cases warrant a smaller amount of estimation data to better track the performance curve over time, at the cost of larger performance estimation errors as shown in Fig. 3.

C. Individual performance curve estimation

Table I presents per-participant ground-truth and predicted accuracies, mean and standard deviation of the absolute error in pp, and the proportion of true accuracies within the estimated 95%-confidence interval, using 30 min of estimation data and 20 s baseline window length, averaged across ten repetitions.

The results clearly show that the accuracy prediction method closely approximates the ground-truth accuracy, even on a per-participant level, with a mean absolute error of 2.1 pp (standard deviation 1.7 pp). The estimated 95%-confidence intervals contain the true accuracy in 94.0% of cases, validating our uncertainty quantification.

D. Validation on VLAAI network on Fuglsang-2018 dataset

We validate the performance modeling method on the second experimental setup with the VLAAI deep neural network on the Fuglsang-2018 dataset (Section III-B). Informed by the previous results, we use a 25 s baseline decision window length (closest to 20 s in the range of window lengths in Section III-B) and 30 min of estimation data, again with ten repetitions.

Similarly to Fig. 2, Fig. 4 shows the modeled average performance curve across participants vs. the ground-truth one. Table II shows the per-participant mean and standard deviation of the estimation errors across target decision window lengths and repetitions. Both demonstrate comparable performance to

TABLE I

THE PER-PARTICIPANT PREDICTION RESULTS FOR THE LINEAR DECODER ON THE DAS-2016 DATASET, STARTING FROM 20 s BASELINE WINDOW LENGTH AND USING 30 min OF ESTIMATION DATA. CI = CONFIDENCE INTERVAL

Part.	Accuracy [%] per decision window length [s]												Abs. error		% in CI
	60		30		20		10		5		1				
	true	pred	true	pred	true	pred	true	pred	true	pred	true	pred	mean	std	
1	81.9	80.4	74.3	72.9	66.2	69.1	64.1	63.8	60.9	59.9	54.7	54.4	2.2	1.8	100.0
2	94.4	95.3	93.8	88.4	85.6	83.6	77.8	75.6	71.5	68.8	59.7	58.7	2.6	2.1	80.0
3	94.4	96.0	91.7	89.5	86.6	84.8	78.0	76.7	72.8	69.7	59.8	59.1	2.4	1.6	90.0
4	93.1	95.4	86.1	88.4	81.9	83.5	74.8	75.5	68.8	68.7	59.4	58.6	1.7	1.2	96.7
5	94.4	97.4	92.4	91.7	88.9	87.1	80.6	78.9	70.6	71.6	59.8	60.1	2.0	1.6	83.3
6	93.1	91.2	82.6	83.2	80.1	78.4	70.6	71.1	66.7	65.3	57.4	57.0	1.6	1.3	100.0
7	93.1	93.5	83.3	85.9	79.2	81.0	75.5	73.3	68.2	67.0	58.1	57.8	1.7	1.4	100.0
8	79.2	75.7	71.5	69.0	67.1	65.7	63.0	61.3	58.6	58.0	55.2	53.6	2.4	2.6	91.7
9	76.4	79.3	68.1	72.0	64.4	68.4	62.3	63.3	58.2	59.5	54.1	54.3	3.5	2.9	88.3
10	91.7	94.3	84.7	87.0	81.5	82.1	74.5	74.3	66.4	67.8	58.8	58.2	2.0	1.4	100.0
11	86.1	86.5	80.6	78.3	72.7	73.9	69.0	67.5	63.2	62.6	56.6	55.7	2.0	1.6	96.7
12	79.2	81.8	69.4	74.1	70.4	70.1	65.0	64.6	63.0	60.4	56.3	54.7	2.7	2.2	91.7
13	90.3	90.9	85.4	82.8	79.6	78.1	71.5	70.9	65.2	65.1	57.2	56.9	1.8	1.6	98.3
14	97.2	98.7	95.1	94.3	91.2	90.1	82.9	82.0	76.5	74.1	62.4	61.4	1.5	1.1	91.7
15	91.7	93.1	85.4	85.4	81.9	80.5	74.3	72.9	67.4	66.7	58.7	57.6	1.7	1.1	100.0
16	95.8	95.4	87.5	88.4	84.3	83.6	77.8	75.6	68.1	68.9	58.8	58.7	1.6	1.5	95.0
Mean	95.8	95.4	87.5	88.4	84.3	83.6	77.8	75.6	68.1	68.9	58.8	58.7	2.1	1.7	94.0

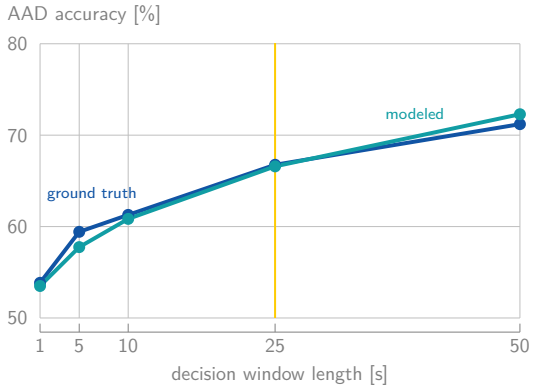

Fig. 4. The modeled vs. ground-truth average performance curves of the VLAAI deep neural network on the Fuglsang-2018 dataset, starting from 25 s baseline window length and using 30 min of estimation data.

the first setup on the Das-2016 dataset, with a mean absolute error across participants of 2.5 pp (standard deviation 2.1 pp), and 92.2% coverage by the estimated 95%-confidence interval.

V. CONCLUSION

We proposed a new AAD accuracy prediction method for correlation-based AAD algorithms that models the entire accuracy-decision window length performance curve starting from the (un)attended correlations at a single window length. Validation across two distinct AAD implementations shows consistent low prediction errors around 2 percent points. Current limitations include the assumption of two competing speech signals and the requirement for *labeled* baseline correlations. Future work will address multi-speaker scenarios and integration with unsupervised accuracy estimation [6].

TABLE II

THE PER-PARTICIPANT PREDICTION RESULTS FOR THE VLAAI NEURAL NETWORK ON THE FUGLSANG-2018 DATASET, STARTING FROM 25 s BASELINE WINDOW LENGTH AND USING 30 min OF ESTIMATION DATA.

	Participant																		Mean
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
mean abs err.	2.3	2.9	2.5	2.4	3.1	2.4	2.2	2.3	2.4	1.9	2.7	2.7	2.6	2.4	2.1	2.9	2.6	2.5	2.5
std abs err.	1.8	2.7	2.8	1.9	2.2	2.4	1.7	1.4	2.3	1.5	2.6	3.0	2.1	1.8	1.6	2.1	1.7	2.4	2.1
% in 95%-CI	88	94	92	98	86	92	94.	90	94	96	98	98	90	88	96	88	82	96	92.2

The proposed AAD accuracy prediction method enables efficient performance curve modeling without continuously re-evaluating AAD performance across multiple decision window lengths, facilitating practical neuro-steered hearing device applications by, e.g., allowing to steer AAD-based gain control systems and providing a theoretical model to translate target AAD accuracies to correlation targets and vice versa.

REFERENCES

[1] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, pp. 233–236, 2012.

[2] J. A. O'Sullivan *et al.*, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cereb. Cortex*, vol. 25, no. 7, pp. 1697–1706, 2014.

[3] S. Geirnaert *et al.*, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 89–102, 2021.

[4] S. Geirnaert, T. Francart, and A. Bertrand, "An Interpretable Performance Metric for Auditory Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 307–317, 2020.

[5] ——, "Time-adaptive Unsupervised Auditory Attention Decoding Using EEG-based Stimulus Reconstruction," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 3767–3778, 2022.

[6] M. A. Lopez-Gordo, S. Geirnaert, and A. Bertrand, "Unsupervised Accuracy Estimation for Brain-Computer Interfaces based on Selective Auditory Attention Decoding ," *IEEE Trans. Biomed. Eng.*, 2025.

[7] S. Geirnaert, J. Vanthornhout, T. Francart, and A. Bertrand, "Modeling Correlations, Null Distributions and Significance Levels in Neural Tracking of Natural Stimuli," KU Leuven, Tech. Rep., 2025, https://simongeirnaert.com/publication/modeling2025geirnaert/.

[8] R. Fisher, "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample," *Metron*, vol. 1, pp. 1–32, 1921.

[9] H. Hotelling, "New light on the Correlation Coefficient and its Transforms," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 15, no. 2, pp. 193–232, 1953.

[10] R. T. Fouladi and J. H. Steiger, "The Fisher transform of the Pearson Product Moment Correlation Coefficient and Its Square: Cumulants, Moments, and Applications," *Commun. Stat. Simul. Comput.*, vol. 37, no. 5, pp. 928–944, 2008.

[11] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Stat. Sci.*, vol. 11, no. 3, pp. 189–228, 1996.

[12] S. Geirnaert, T. Francart, and A. Bertrand, "AAD performance curve modeling code." [Online]. Available: https://github.com/AlexanderBertrandLab/AAD-performance-curve-modeling

[13] B. Accou, J. Vanthornhout, H. Van hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Sci. Rep.*, vol. 13, no. 812, 2023.

[14] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, 2017.

[15] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *Neuroimage*, vol. 156, pp. 435–444, 2017.