Linear stimulus reconstruction *works* on the KU Leuven audiovisual, gaze-controlled auditory attention decoding dataset

Simon Geirnaert[1,2,3]**, Iustina Rotaru**[1,2]**, Tom Francart**[2,3]**, Alexander Bertrand**[1,3]

[1]KU Leuven, Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics
[2]KU Leuven, Department of Neurosciences, ExpORL
[3]Leuven.AI - KU Leuven institute for AI,

E-mail: `simon.geirnaert@kuleuven.be, alexander.bertrand@kuleuven.be`

Abstract. In a recent paper, we presented the KU Leuven audiovisual, gaze-controlled auditory attention decoding (AV-GC-AAD) dataset, in which we recorded electroencephalography (EEG) signals of participants attending to one out of two competing speakers under various audiovisual conditions. The main goal of this dataset was to disentangle the direction of gaze from the direction of auditory attention, in order to reveal gaze-related shortcuts in existing spatial AAD algorithms that aim to decode the (direction of) auditory attention directly from the EEG. Various methods based on spatial AAD do not achieve significant above-chance performances on our AV-GC-AAD dataset, indicating that previously reported results were mainly driven by eye gaze confounds in existing datasets. Still, these adverse outcomes are often discarded for reasons that are attributed to the limitations of the AV-GC-AAD dataset, such as the limited amount of data to train a working model, too much data heterogeneity due to different audiovisual conditions, or participants allegedly being unable to focus their auditory attention under the complex instructions. In this paper, we present the results of the linear stimulus reconstruction AAD algorithm and show that high AAD accuracy can be obtained within each individual condition and that the model generalizes across conditions, across new subjects, and even across datasets. Therefore, we eliminate any doubts that the inadequacy of the AV-GC-AAD dataset is the primary reason for the (spatial) AAD algorithms failing to achieve above-chance performance when compared to other datasets. Furthermore, this report provides a simple baseline evaluation procedure (including source code) that can serve as the minimal benchmark for all future AAD algorithms evaluated on this dataset.

Contents

1. Introduction

Selective auditory attention decoding (AAD) methods aim to identify the sound source a person is attending to amidst a cocktail of sound sources based on neural recordings such as electro- or magneto-encephalography (EEG/MEG) [1, 2]. A typical application of AAD is in cognitively-controlled hearing aids, which allow the user to steer their hearing aid towards the conversation they actually want to listen to, for example, in a cocktail party scenario [3].

AAD methods can generally be classified into two different classes: stimulus decoding and direct classification [3,4]. *Stimulus decoding* leverages so-called neural tracking: the brain tracks (features of) the attended stimulus better than other unattended stimuli [1, 5, 6]. This class of algorithms, therefore, aims to reconstruct such features (e.g., the speech envelope) of the attended stimulus from the neural responses (backward modeling) to compare them with the presented stimuli to identify the correct speaker [2]. Alternatively, the neural responses can be predicted from each stimulus and compared with the actual neural responses recorded through EEG/MEG (forward modeling) [7]. Combinations of backward and forward modeling have also been proposed [8]. The models used to reconstruct can vary from simple linear models to more complex non-linear models. This class of algorithms has been well-established in the literature, yielding robust results that have been countless times reproduced in various datasets [3]. The main disadvantage of this class is that the performance quickly degrades when less data is available to make a decision (i.e., at increasing decision speeds) [9].

The second class of *direct classification* algorithms is becoming increasingly popular along with the rise of deep learning methods in EEG/MEG. While various approaches exist, a particularly popular subbranch consists of determining the *spatial location* of the attended

sound source solely based on the neural responses (spatial auditory attention decoding (Sp-AAD)) [10–14]. These neural responses are believed to reflect certain lateralization patterns based on the listening direction, offering an alternative approach to identifying the attended sound source based on their location relative to the listener. While the models used for Sp-AAD can be simple linear data-driven models (e.g., based on common spatial patterns [11]), there has been a recent surge in various non-linear deep learning models [10, 13, 14]. Their main advantage is that they claim to require much less data to make an accurate decision about the attended speaker.

However, in a recent paper [15], we uncovered that these Sp-AAD algorithms suffer from various (unwanted) shortcuts or confounds in the datasets on which they are tested. One example of such a shortcut is trial fingerprints. We showed that Sp-AAD models are highly susceptible to these trial fingerprints and are (implicitly) overfitting due to high statistical similarity between segments in the training and test set. However, many papers resort to random cross-validation based on very short (and sometimes overlapping) segments as an evaluation scheme, leading to unrealistically and misleadingly high performance, as pointed out in Puffay et al. [16]. Generalizing across various trials or subjects turns out to be much harder than with stimulus decoding algorithms due to feature shifts across trials. A second shortcut arises from confounding signal components in the neural responses, such as eye movements. Various datasets have no or very limited restrictions on eye gaze (such as the KU Leuven AAD dataset of 2016 [17]), which could lead to congruent eye movements with the direction of auditory attention due to the tendency of listeners to (in)voluntarily direct their gaze toward the direction of the attended speaker. Data-driven models could then leverage eye movements to decode attention with a much higher accuracy, rather than the actual neural processes that are much harder to decode. While decoding eye movements for Sp-AAD might work just fine in many scenarios, it is not a fully robust way of decoding attention (e.g., when eavesdropping). Furthermore, it means that these models might not really decode attention directly from the brain, but instead exploit shortcuts in the EEG recordings that are in fact generated by the eyes, which (in many cases) turn out to be correlated with the direction of auditory attention.

To uncover these potential shortcuts in Sp-AAD algorithms, Rotaru et al. [15] introduced a new two-speaker AAD dataset with forced congruent and incongruent spatial auditory and visual attention (i.e., the audiovisual, gaze-controlled auditory attention decoding dataset KU Leuven (AV-GC-AAD dataset) [18]). This AV-GC-AAD dataset is publicly available as a new benchmarking dataset. However, the original paper [15] did not include an analysis using the first class of stimulus decoding algorithms. In this paper, we want to provide such an analysis for various reasons. Firstly, as stimulus decoding algorithms are very robust and well-established, they allow to verify that the participants in the experiment followed the instructions and were indeed able to attend to the correct speaker, despite the varying visual attention cues that complicate focusing auditory attention. Secondly, it allows to provide a trustworthy baseline benchmark on the new dataset. As previously mentioned, many AAD datasets suffer from shortcuts and confounds regarding eye gaze. We have noticed that several Sp-AAD deep learning algorithms achieve skyrocketing performance on other datasets with minimal control for eye gaze while they fail on our new AV-GC-AAD dataset. However, AAD model developers often discard these failures and attribute them to an insufficient amount of training data, participants being unable to focus their auditory attention due to the distracting incongruent

Condition	Auditory vs. visual attention	Task
No visuals	Incongruent	Fixate on an imaginary point in the middle of the black screen while minimizing eye movements
Static video	Congruent	Fixate on a static video of the attended speaker located on the same side of the corresponding attended speech signal (left or right)
Moving video	Incongruent	Follow a moving video of the attended speaker along a random horizontal trajectory on the screen
Moving target + noise	Incongruent	Follow a moving crosshair along a random horizontal trajectory on the screen. There is additional auditory background babble noise present at $-1\,$dB signal-to-noise ratio (SNR)

Table 1: An overview of the 4 different conditions in the AV-GC-AAD dataset. Per condition, there are 2 trials of 10 min long.

gaze-steering task, or a too-large variety across conditions. With this stimulus decoding analysis, we want to refute those arguments and, therefore, provide a minimal benchmark for future Sp-AAD and other algorithms.

In Section 2, we give a brief overview of the AV-GC-AAD dataset, while the linear stimulus decoding algorithm for AAD is revisited in Section 3. In Section 4, we explain the various experiments and evaluation setups and in Section 5, we show and discuss results.

2. Dataset

The audiovisual, gaze-controlled auditory attention (AV-GC-AAD) dataset from KU Leuven was originally presented in Rotaru et al. [15]. 16 young, normal-hearing participants were instructed to listen to one out of two competing talkers located at $\pm90°$ w.r.t. the participant. This specific perceived spatial separation between speakers, presented through insert earphones, was obtained by convolving the two competing stimuli with the relevant head-related transfer functions. The stimuli consisted of science outreach podcasts in Dutch, for which videos were also available. The EEG was recorded using a 64-channel BioSemi ActiveTwo system, while also 4 electro-oculography (EOG) electrodes were used to record eye movements.

The experiment consisted of 4 different conditions with 2 trials per condition, resulting in 8 trials in total. Each trial was 10 min long. These 4 different conditions, summarized in Table 1, implement different visual conditions such that the effect of eye gaze (movements) on AAD can be investigated. In each trial, the participant always had to listen to the to-be-attended speaker, who, after 5 min, switched sides to emulate a spatial attention switch. Importantly, the randomization was done such that the 2 trials from the same conditions were 40 min apart in time, resulting in a potentially substantial shift in statistics between both trials.

A full description of the dataset can be found in Rotaru et al. [15]. This dataset is publicly available online [18], minus the EEG data from three participants (subject 2, 5, and 6) who

did not consent to making their data publicly available. For a few subjects, a particular trial or condition is not present (see [18] for details), which is taken into account during evaluation. In this report, we use the version of the dataset that can be found online, so that all results are fully reproducible.

3. Methods

In this section, we revisit the basics of the linear stimulus decoding or reconstruction (backward modeling) approach for AAD. As shown in Figure 1, the stimulus decoding approach aims to reconstruct certain temporal features of the attended speech signal from the neural responses of the listener. A typical example of such a feature is the speech envelope, which we will also use in the remainder of this report. Other acoustic and linguistic features can also be used [19–21]. This reconstruction can then be correlated with (features of) the presented competing speech signals to determine the attended speech signal [2, 3].

3.1. Training the stimulus decoder

To decode or reconstruct the attended speech envelope $\hat{s}_a(t)$, with t the time sample index, we linearly combine time-lagged copies of the EEG channels [2, 3]:

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

where $x_c(t)$ is the value of the c-th EEG channel at time t, $d_c(l)$ is the l-th decoder coefficient for channel c, and C and L are the number of EEG channels and decoder time lags, respectively. As shown in (1), the stimulus decoder $d_c(l)$ is an *anti-causal* filter, where only time lags l ranging from 0 to $L - 1$ *after* the current stimulus sample at time t are used, given we are trying to reconstruct the stimulus from the response. Using vector notations, (1) can be rewritten as:

$$\hat{s}_a(t) = \mathbf{x}(t)^{\mathrm{T}} \mathbf{d},$$

where $\mathbf{x}(t)$ contains all time lags per EEG channel:

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_1(t + L - 1) \\ x_2(t) \\ \vdots \\ x_C(t + L - 1) \end{bmatrix} \in \mathbb{R}^{CL},$$

and $\mathbf{d} \in \mathbb{R}^{CL}$ similarly stacks all spatio-temporal decoder coefficients $d_c(l)$.

To train the decoder coefficients $d_c(l)$, assume the availability of a training set of T time samples, i.e., $\{\mathbf{X}, (\mathbf{s}_1, \mathbf{s}_2), \mathbf{y}\}$, with $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_C \end{bmatrix} \in \mathbb{R}^{T \times CL}$, where $\mathbf{X}_c \in \mathbb{R}^{T \times L}$ is a Hankel matrix containing the time-lagged EEG data of the cth channel:

$$\mathbf{X}_c = \begin{bmatrix} x_c(0) & x_c(1) & \cdots & x_c(L-1) \\ x_c(1) & x_c(2) & \cdots & x_c(L) \\ \vdots & \vdots & & \vdots \\ x_c(T-1) & 0 & \cdots & 0 \end{bmatrix}.$$
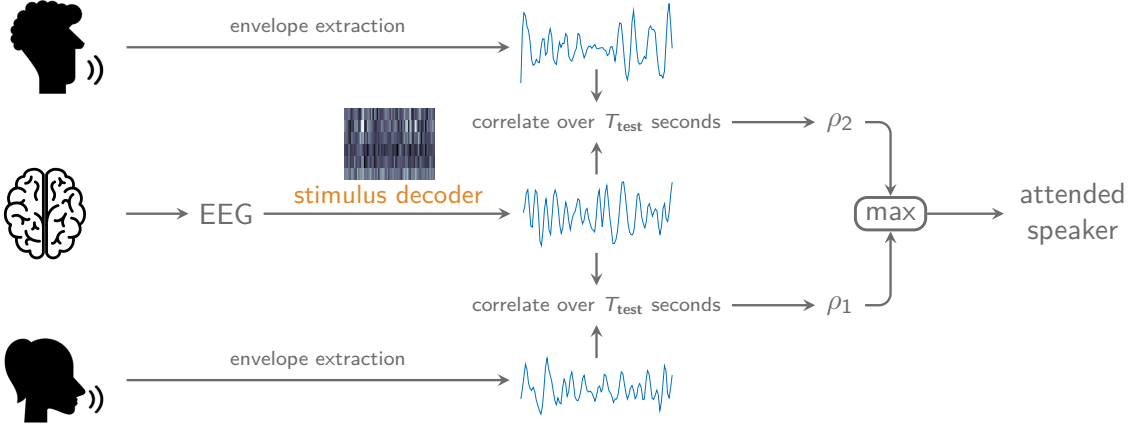
Figure 1: An overview of the linear stimulus decoding algorithm for AAD, in which the attended speech envelope is reconstructed from the neural responses and correlated with the presented speech envelopes to identify the attended one through the Pearson correlation coefficient. Based on Figure 3a in [3] and Figure 2 in [22].

Furthermore, $\mathbf{s}_1$ and $\mathbf{s}_2 \in \mathbb{R}^T$ contain the T samples of both competing speech envelopes. Additionally, during training, we have knowledge of the attention labels $\mathbf{y} \in \mathbb{R}^T$, with $y(t) \in \{1, 2\}$ indicating at each time sample which speech envelope corresponds to the attended envelope $\mathbf{s}_a \in \mathbb{R}^T$, i.e.,

$$s_a(t) = \begin{cases} s_1(t) & \text{if } y(t) = 1, \\ s_2(t) & \text{if } y(t) = 2. \end{cases} \tag{2}$$

The stimulus decoder can then be trained by minimizing the squared error between the reconstructed envelope $\hat{\mathbf{s}}_a = \mathbf{X}\mathbf{d}$ and attended one $\mathbf{s}_a$:

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\arg\min} \, ||\mathbf{s}_a - \mathbf{X}\mathbf{d}||_2^2 \,. \tag{3}$$

The solution of (3) can be found by solving the normal equations, leading to:

$$\hat{\mathbf{d}} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xs},$$

where

$$\mathbf{R}_{xx} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^{CL \times CL} \tag{4}$$

corresponds to the (unnormalized) EEG autocorrelation matrix and

$$\mathbf{r}_{xs} = \mathbf{X}^{\mathrm{T}}\mathbf{s}_a \in \mathbb{R}^{CL} \tag{5}$$

to the (unnormalized) cross-correlation vector between the EEG and attended speech envelope.

3.2. Using the stimulus decoder for AAD

Given a new, unseen segment of EEG data $\mathbf{X}^{(\text{test})} \in \mathbb{R}^{T_{\text{test}} \times CL}$ and corresponding competing speech envelopes $\mathbf{s}_1^{(\text{test})}$ and $\mathbf{s}_2^{(\text{test})} \in \mathbb{R}^{T_{\text{test}}}$, the goal is to identify the attended speaker in this segment, i.e., to determine the correct label $y^{(\text{test})}$. The trained decoder $\hat{\mathbf{d}}$ can now be used to reconstruct the speech envelope of the attended speaker $\hat{\mathbf{s}}_a^{(\text{test})} = \mathbf{X}^{(\text{test})}\hat{\mathbf{d}}$ from

the EEG of the listener, which can then be compared with the competing speech envelopes through the Pearson correlation coefficient. The speaker that exhibits the highest correlation $(\rho(\hat{\mathbf{s}}_a^{(\text{test})}, \mathbf{s}_1^{(\text{test})})$ or $\rho(\hat{\mathbf{s}}_a^{(\text{test})}, \mathbf{s}_2^{(\text{test})}))$ is identified as the attended speaker. This is summarized in Figure 1.

A crucial parameter in determining the AAD accuracy (i.e., number of correct decisions) is the decision window length T_{test}, i.e., the number of samples used to compute the correlation. Given that the variability of these correlations increases with lower number of samples, the accuracy typically quickly decreases when using shorter decision window lengths [9].

4. Experiments

In Section 4.1, we explain the preprocessing steps on the speech stimuli and EEG data. Section 4.2 gives the details of the stimulus decoder settings, while Section 4.3 explains all procedures to evaluate the AAD stimulus decoding algorithm on the AV-GC-dataset.

All MATLAB code to reproduce all experiments and results is available at `https://github.com/AlexanderBertrandLab/linear-stimulus-reconstruction-AAD-AV-GC-AAD-dataset`.

4.1. Preprocessing

To ensure full reproducibility, we start from the publicly online version of the dataset [18]. In the online version of the dataset, the EEG signals are already bandpass-filtered between 128 Hz using a type II zero-phase Chebyshev filter and downsampled to 128 Hz. The speech envelopes are precomputed by applying a gammatone filterbank and computing the envelope per subband signal using a powerlaw operation with exponent 0.6. All subband envelopes are then summed to one envelope, bandpass-filtered between 1–40 Hz and downsampled to 128 Hz similarly to the EEG.

We have, additionally, bandpass-filtered both EEG data and speech envelopes between 1–9 Hz [23] using a 4th-order zero-phase Butterworth filter and downsampled them to 20 Hz. All EEG channels and speech envelopes were subsequently z-scored (mean and standard deviation put to 0 and 1, respectively) per 10 min-trial.

4.2. Decoder settings

For the decoder, a filter range of 0–400 ms post-stimulus lags is chosen, such that $L = 9$ (at 20 Hz), while $C = 64$ given the 64-channel EEG system.

To avoid overfitting when only limited amounts of training data are available (e.g., in the subject-specific, per-condition evaluation (see below)), we use shrinkage to regularize the estimation of the EEG covariance matrix in (4) [24, 25]. The regularization parameter is heuristically determined using the Ledoit-Wolf shrinkage estimator [26], which is recommended as the state of the art [27].

4.3. Evaluation procedure

Several deep learning approaches in AAD and neural tracking suffer from shortcuts and implicit trial-overfitting due to the uncareful validation procedure they employ [16]. Although the risk

for (linear) stimulus reconstruction approaches that rely on correlation-based decisions is much lower, we will employ careful cross-validation to exclude all risks of overfitting to such potential shortcuts. We will use both subject-specific and subject-independent decoders, where the latter allows to assess the generalization abilities of the AAD algorithm in the hardest sense, i.e., across different subjects.

In *subject-specific* decoders, the stimulus decoder is trained and tested on data from the same subject. Such decoders generally perform better as they are tailored to the test subject. To avoid implicit overfitting, we will use a *leave-one-trial-out* cross-validation (LOTO-CV) procedure, i.e., we always leave out a complete trial of 10 min from the training set. We will do this both within conditions and across conditions. Within-condition evaluation boils down to only taking the two trials corresponding to one of the four conditions in Table 1 and performing LOTO-CV on those two trials. In this case, there is only 10 min of training data. Across-condition evaluation refers to LOTO-CV on all trials of all four conditions combined. Alternatively, we will also perform *leave-one-condition-out* CV (LOCO-CV) across all conditions, i.e., we will leave out both trials from one condition in the training set. This allows to check generalization to specific conditions.

In *subject-independent* decoders, the stimulus decoders are trained on data from different subjects than the test subject, i.e., a *leave-one-subject-out* CV (LOSO-CV) procedure is used. We will perform LOSO-CV across all conditions combined. Secondly, to refute arguments that there is insufficient training data to train more complex models, we will also train a stimulus decoder based on all subjects of the KU Leuven AAD dataset 2016 [17] (16 subjects $\times 72$ min ≈ 19 h of training data). This subject-independent decoder is then again tested on all subjects of the AV-GC-AAD dataset. The same preprocessing steps as in Section 4.1 are applied on the KU Leuven AAD dataset 2016, which was recorded using the same EEG system.

The significance level for the decoding accuracy is computed as the 95-percentile of the inverse binomial cumulative distribution function with 50% chance of success and the corresponding number of decisions.

5. Results and discussion

In Section 5.1, we show and discuss the results of the experiments using a subject-specific decoder, while in Section 5.2, we show the results of subject-independent decoding.

5.1. Subject-specific decoding

Figure 2 shows the results of the LOTO-CV on a per condition basis, while Figure 3 shows the results across conditions.

5.1.1. Per condition evaluation
For the per-condition LOTO-CV, one 10 min-trial of each condition is used as training data, while the other one is used as the testing trial, and vice versa. The first crucial observation from Figure 2 is that the stimulus decoding algorithm works for *each* condition, as significant AAD accuracies are achieved for every single condition. This implies that the participants were generally able to focus their attention on the instructed speaker despite the various visual instructions. Even in the case of a visually moving target

Subject-specific per-condition AAD performance

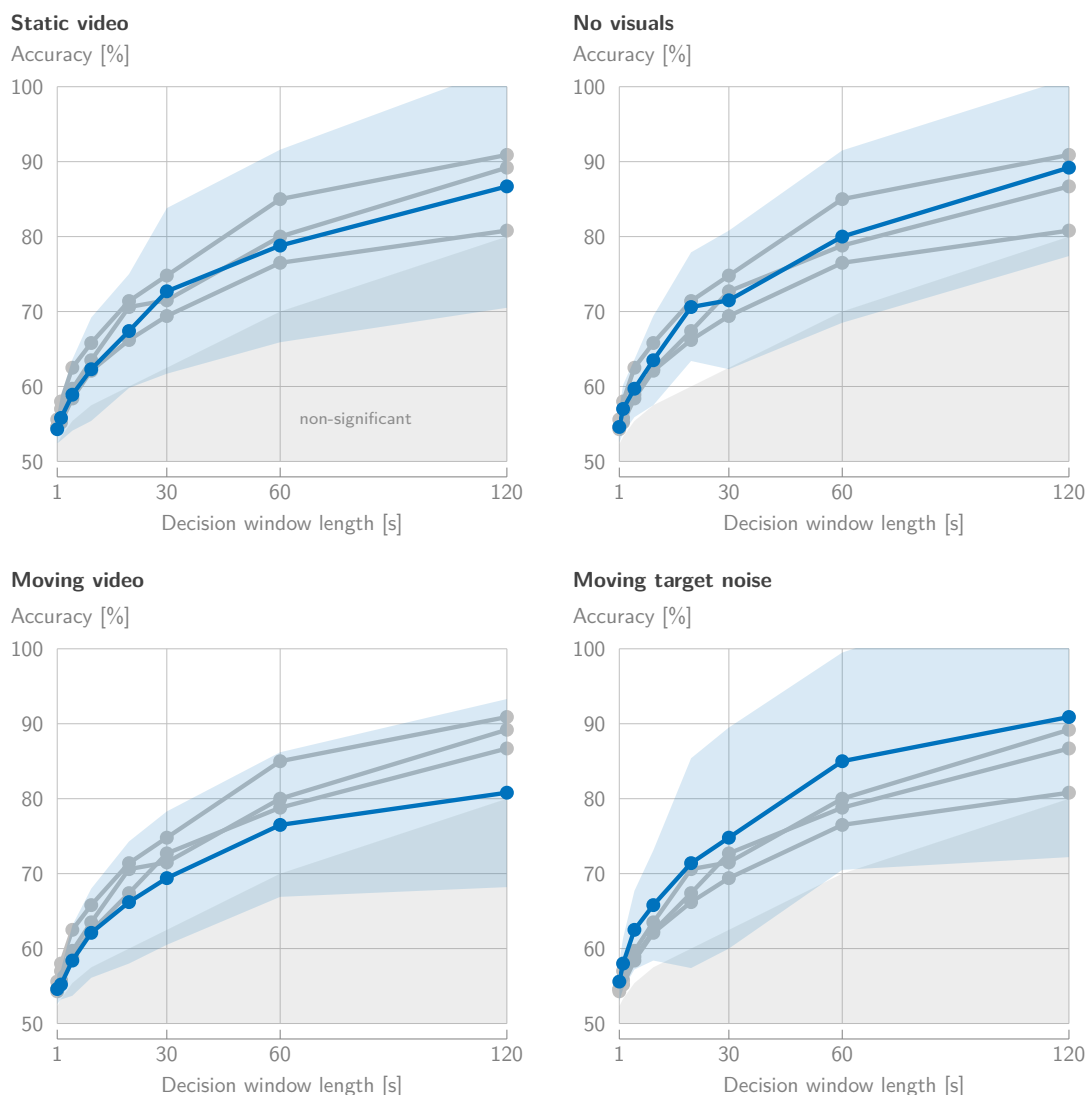Leave-one-trial-out CV | Mean ± standard deviation (shaded areas)

Figure 2: Using leave-one-trial-out CV for subject-specific decoding per individual condition leads to significant AAD performances for every single condition, even when the visual instruction is incongruent with the direction of auditory attention (moving video and moving target + noise). Gray lines are replicas per condition, provided as a reference.
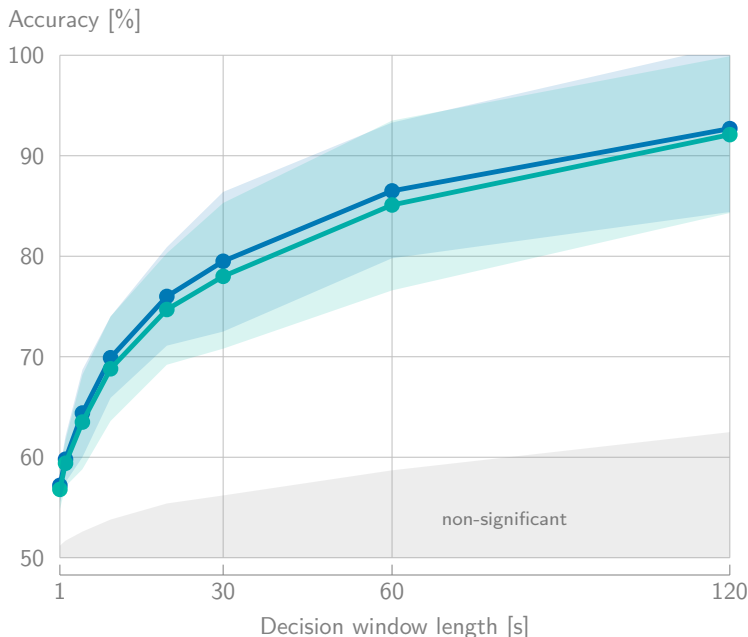
(as in the moving video or moving target + noise condition), significant performances are obtained.

When comparing the different conditions, the moving video condition is the worst one. This is no surprise, as it is one of the hardest conditions, given the continuously and randomly moving video, independent of the direction of the attended speaker. On the other hand, the moving target + noise also has a randomly moving target (albeit a crosshair) and results in the highest accuracies. In this moving target + noise condition, the difficulty of the moving crosshair could be offset by the added auditory background noise. Das et al. [28] observed an increasing performance with mild background noise at $-1.1\,\mathrm{dB}$ SNR and $180°$ speaker separation w.r.t. no background noise, potentially due to the higher listening effort of the participant resulting in stronger neural tracking. The same effect could explain the

Figure 3: (a) Using leave-one-trial-out or leave-one-condition-out CV across all conditions at once both leads to significant and expected AAD performances. **(b)** A breakdown of the leave-one-condition-out CV accuracies per condition shows that generalization to every other condition is possible and similar. One dot represent the average 60 s AAD accuracy for one subject.

higher accuracy in the moving target + *noise* condition w.r.t. the other conditions without background noise, as it is recorded using the same settings as in Das et al. [28].

However, per-condition accuracies are generally lower than expected in AAD (e.g., around 90% on 60 s decision windows in O'Sullivan et al. [2]). However, this lower accuracy can be attributed to the low amount of training data per condition (10 min), as the performance increases with LOTO-CV across all conditions (see later) and, for example, when using random 10-fold CV (= 18 min of training data) based on shorter segments of 60 s (86.9% on 60 s decision windows for the no visuals condition).

Subject-independent across-conditions AAD performance

Leave-one-subject-out CV and trained on KU Leuven AAD dataset 2016 [17]
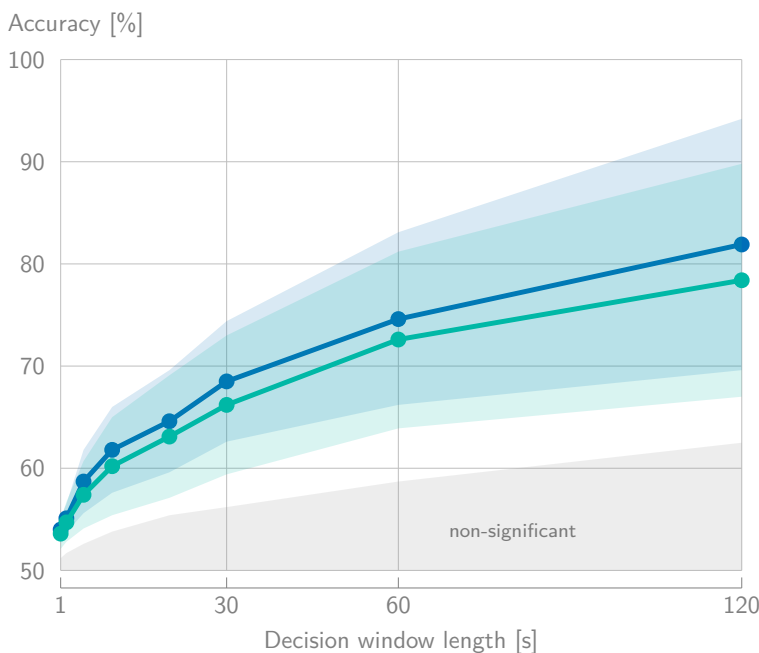Mean ± standard deviation (shaded areas)

Figure 4: Using leave-one-subject-out CV and generalizing from the KU Leuven AAD dataset 2016 (i.e., subject-independent decoding) leads to significant AAD performances, showing that generalization across subjects and datasets is possible.

5.1.2. Across conditions evaluation Across conditions, both leave-one-trial-out and leave-one-condition-out CV are used. Figure 3a shows that both CV evaluation schemes achieve very comparable and significant accuracies. This shows that generalization across all the different conditions using linear stimulus decoding is possible. Moreover, the accuracies are now comparable with other datasets (such as in O'Sullivan et al. [2] and Geirnaert et al. [24]) as it achieves performances of 86.5% (LOTO-CV) and 85.1% (LOCO-CV) on 60 s decision windows.

To show that this generalization holds to every other condition, Figure 3b shows a breakdown of the per-condition and per-subject AAD accuracy using 60 s decision windows. All conditions have a similar mean accuracy around 85% with a similar spread.

5.2. Subject-independent decoding

Figure 4 shows the results of the leave-one-subject-out CV evaluation, where all data from one subject is entirely removed from the training set, and the generalization from the KU Leuven AAD dataset 2016 (i.e., subject-independent decoding). The significant AAD classification performance in both cases shows that even generalization across subjects and datasets with the linear stimulus decoder is possible.

For LOSO-CV, this good generalization is achieved despite the large variety of audiovisual conditions. A similar, significant performance is achieved when training the decoder on the KU Leuven AAD dataset 2016 on all conditions (on 60 s: 70.4% on no visuals, 72.7% on

static video, 70.8% on moving video, and 76.8% on moving target + noise, respectively). This shows that even generalization across datasets to the AV-GC-AAD dataset is possible. The implication is that all the data of the KU Leuven AAD dataset 2016 could be additionally used to generate more training data for more complex models.

6. Conclusion

In this short report, we have shown that linear stimulus reconstruction works on the KU Leuven audiovisual, gaze-controlled auditory attention decoding (AV-GC-AAD) dataset. Significant and expected performances are obtained for subject-specific decoding, both within all different audiovisual conditions and across all conditions. Generalizing across different subjects with a subject-independent decoder also yields significant AAD performance. Moreover, even generalizing from a different dataset to the AV-GC-AAD dataset is possible, opening up possibilities to generate more training data to train more complex models. The results show that the participants were able to follow the instructions and attend to the correct speaker and that their auditory attention is decodable. Moreover, we have provided a baseline analysis and performance for all future AAD algorithms on the AV-GC-AAD dataset, showing that above-chance accuracy is easily achievable on these data with simple linear models. Future algorithms should be able at least to beat the linear stimulus reconstruction algorithm on this dataset.

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, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. 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, 2014.

[3] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[4] S. Geirnaert, "Signal processing algorithms for eeg-based auditory attention decoding," Ph.D. dissertation, KU Leuven, May 2022.

[5] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[6] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. G. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder, "Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[7] D. D. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, no. 531, 2018.

[8] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. C. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[9] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 307–317, 2020.

[10] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *eLife*, vol. 10, no. e56481, 2021.

[11] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1557–1568, 2021.

[12] ——, "Riemannian Geometry-Based Decoding of the Directional Focus of Auditory Attention Using EEG," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1115–1119.

[13] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention from EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, 2022.

[14] S. Pahuja, S. Cai, T. Schultz, and H. Li, "Xanet: Cross-attention between eeg of left and right brain for auditory attention decoding," in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2023, pp. 1–4.

[15] I. Rotaru, S. Geirnaert, N. Heintz, I. Van de Ryck, A. Bertrand, and T. Francart, "What are we really decoding? Unveiling biases in EEG-based decoding of the spatial focus of auditory attention," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016017, 2024.

[16] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, T. Francart *et al.*, "Relating EEG to continuous speech using deep neural networks: a review," *Journal of Neural Engineering*, vol. 20, no. 4, p. 041003, 2023.

[17] N. Das *et al.*, "Auditory Attention Detection Dataset KULeuven," Zenodo, 2019. [Online]. Available: https://zenodo.org/record/3997352

[18] I. Rotaru, S. Geirnaert, A. Bertrand, and T. Francart, "Audiovisual, Gaze-controlled Auditory Attention Decoding Dataset KU Leuven (AV-GC-AAD)," Zenodo, Apr. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.11058711

[19] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[20] M. Gillis, J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck, "Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics," *Journal of Neuroscience*, vol. 41, no. 50, pp. 10 316–10 329, 2021.

[21] J. Van Canneyt, J. Wouters, and T. Francart, "Enhanced Neural Tracking of the Fundamental Frequency of the Voice," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 12, pp. 3612–3619, 2021.

[22] S. Geirnaert, R. Zink, T. Francart, and A. Bertrand, "Fast, Accurate, Unsupervised, and Time-Adaptive EEG-Based Auditory Attention Decoding for Neuro-steered Hearing Devices," in *Brain-Computer Interface Research: A State-of-the-Art Summary 11*, C. Guger, B. Allison, T. M. Rutkowski, and M. Korostenskaja, Eds. Springer Nature Switzerland, 2024, pp. 29–40.

[23] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of Neural Engineering*, vol. 13, no. 056014, 2016.

[24] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE Journal on Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3955–3966, 2021.

[25] ——, "Time-Adaptive Unsupervised Auditory Attention Decoding Using EEG-Based Stimulus Reconstruction," *IEEE Journal on Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3767–3778, 2022.

[26] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[27] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[28] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 066017, 2018.