forum acusticum 2026
8-12 September · Graz, Austria

SPATIAL AUDITORY ATTENTION DECODING USING GAMMATONE-BASED REPRESENTATIONS UNDER CLEAN AND BINAURAL CONDITIONS

Maryam Bajool, Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, Germany

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Auditory attention decoding (AAD) aims to determine from neural recordings, which sound source a listener is attending to. Here, we formulate spatial auditory attention decoding (SpAAD) as a subject-independent end-to-end task, jointly modeling EEG signals and Gammatone-based audio representations to decode the attended spatial direction. While conventional approaches assume access to clean speech signals, real-world conditions provide only binaural mixtures. Therefore, we investigate the feasibility of decoding attention directly from mixture signals, in addition to ear-separated (clean) representations on the DTU and KUL datasets. Results show that decoding performance increases consistently with decision window length for all conditions. On the DTU dataset, binaural mixtures achieve performance comparable to ear-separated signals, with no significant effect of condition. On the KUL dataset, the HRTF-based binaural condition outperforms the ear-separated (dichotic) condition as the decision window increases, with the difference between conditions increasing with window length. These findings demonstrate that spatial auditory attention can be decoded directly from binaural mixtures in a subject-independent setting and highlight the potential of Gammatone-based representations for realistic AAD systems.

Keywords: *Auditory attention decoding (AAD), electroencephalography (EEG), BCI.*

1. Introduction

In everyday listening environments, the auditory system is exposed to complex mixtures of sounds originating from multiple sources. The process by which the brain organizes these acoustic inputs into perceptually meaningful auditory objects is known as auditory scene analysis (ASA) [1]. A key component of ASA is selective auditory attention, which enables listeners to focus on a target sound source while suppressing competing sources, a phenomenon known as the cocktail party effect [2]. Understanding the neural mechanisms underlying selective auditory attention has motivated the development of auditory attention decoding (AAD), which aims to infer the attended speaker from neural recordings such as electroencephalography (EEG).

Traditional AAD approaches typically reconstruct the attended speech envelope from neural signals and identify the attended speaker by correlating the reconstructed envelope with the clean speech envelopes of candidate speakers [3]–[6]. However, this assumption is rarely satisfied in real-world scenarios where clean speech signals are unavailable.

To address this limitation, several studies have proposed frameworks that operate on speech mixtures by incorporating a speech separation or enhancement stage prior to attention decoding [7]–[11]. In these approaches, the acoustic mixture is first separated into speaker estimates, which are subsequently compared with neural responses to determine the attended source. For example, Han et al. [10] and Ceolini et al. [11] demonstrated that neural signals can be combined with deep learning–based speech separation to identify or extract the attended speaker without access to clean speech sources. Although these approaches reduce reliance on clean speech references, they still require an explicit separation stage, which increases computational complexity and is prone to error. Alternatively, decoding directly from binaural mixtures avoids explicit separation but typically leads to reduced performance, particularly when using linear models [12].

In parallel, deep learning approaches have improved AAD performance by modeling nonlinear relationships between EEG and acoustic features. Architectures such as AADNet [13] enable subject-independent de-

***Corresponding author**: maryam.bajool@tum.de.*

coding by jointly processing EEG signals and acoustic features. Among existing approaches, AADNet is particularly suitable due to its end-to-end architecture, which directly models the relationship between EEG signals and acoustic features without requiring explicit envelope reconstruction. However, the role of acoustic feature representations in such frameworks remains insufficiently explored, particularly under realistic binaural listening conditions.

In this work, we focus on the impact of acoustic representations for spatial auditory attention decoding. Specifically, we investigate the use of Gammatone filterbank features, which approximate auditory peripheral processing and preserve spectral information [14]. Unlike conventional approaches that collect filterbank outputs into broadband envelopes, we retain subband-level representations to capture richer spectro-temporal structure.

Furthermore, we compare two acoustic conditions: (i) clean, ear-separated speech signals, representing an idealized scenario, and (ii) spatialized binaural mixtures generated using head-related transfer functions (HRTFs), reflecting realistic listening environments. This comparison allows us to quantify how spatial filtering and signal mixing affect attention decoding performance. The goal is to decode the direction of auditory attention (left vs. right speaker) in a subject-independent setting. We build on the AADNet architecture to jointly process EEG signals and acoustic features for spatial attention decoding. By operating on spatialized binaural mixtures rather than clean speech signals, the proposed approach reduces the reliance on speech preprocessing and explicit source separation stages commonly required in conventional AAD pipelines.

2. Materials and Data

In this work, we use two publicly available datasets recorded in a competing-talker setup.

2.1. DTU Dataset

This dataset [15], [16] was collected from 18 normal-hearing subjects in a double-walled soundproof booth using a 64-channel BioSemi ActiveTwo EEG system, sampling EEG data at 512 Hz. Each participant completed 60 trials of 50 seconds. In each trial, two competing speech streams, narrated by a male and a female speaker, were presented binaurally via insert earphones. The stimuli were spatialized using head-related impulse responses (HRIRs) and lateralized at $\pm 60°$ azimuth. Subjects were asked to attend to one speaker and ignore the other during each trial. After each trial, they answered a question related to the content of the attended speech stream. The raw EEG recordings and the clean speech signals are publicly available in the original repository on Zenodo. In this work, clean speech signals are used for "Ear-separated" representations, while binaural mixtures, referred to as "Mixture", for the left and right ears were created using non-individual KEMAR HRIR filters [17].

2.2. KUL Dataset

This dataset [18], [19] was also collected using a competing-speaker paradigm. Recordings were conducted with a 64-channel BioSemi ActiveTwo EEG system at a sampling rate of 8192 Hz. The dataset includes 16 normal-hearing participants, each of whom completed 20 trials, drawn from three experiments. In the first two experiments, four Dutch stories (narrated by male speakers), two per experiment, were each divided into two 6-minute parts and presented simultaneously to both ears; each experiment yielded four six-minute trials (8 trials total). The third experiment reused material from the four-story stimuli of experiment one. For each stimulus, the first 2 minutes were extracted and repeated three times, resulting in 12 additional 2-minute trials. Participants were instructed to attend to the speech stream presented to one ear while ignoring the competing stream in the other ear. After each trial, subjects answered multiple-choice questions to verify attentional engagement. This dataset included two presentation conditions: a dichotic condition, in which clean speech signals were delivered separately per ear, and an HRTF condition, in which stimuli were spatialized at $\pm 90°$ azimuth using head-related transfer functions. In this study, the HRTF condition was used for the binaural processing framework and referred to as "Mixture", while the dichotic condition was used as "Ear-separated" clean-signal analysis.

3. Data Pre-Processing

In this section, we present the processing steps for both datasets before modeling.

3.1. EEG

The EEG data were first band-pass filtered with an FIR filter with cut-off frequencies at 1 Hz and 32 Hz. The data were downsampled to 64 Hz and re-referenced to the average of all channels.

3.2. Audio

An audio processing pipeline was designed to generate representations for both clean ear-separated and binaural speech conditions. To obtain realistic spatialized mixture representations for the DTU dataset, the clean signals were first spatialized using non-individual KEMAR head-related impulse responses (HRIRs) [17]. The attended and unattended speech streams, x_a and x_u, respectively, were convolved with HRTFs corresponding to $\pm 60°$ azimuth to generate spatialized signals at the left and right ears. Depending on the attended direction label in each trial, the appropriate HRIR pair was assigned to the attended and competing sources. The ear signals corresponding to both sources were then summed at each ear to form the final binaural mixture signals. For the KUL dataset, binaural mixtures were directly obtained from the provided spatialized signals (HRTF condition), while the dichotic condition was used for ear-separated (clean) representations. To obtain auditory representations, the resulting left- and right-ear signals, y_L and y_R,

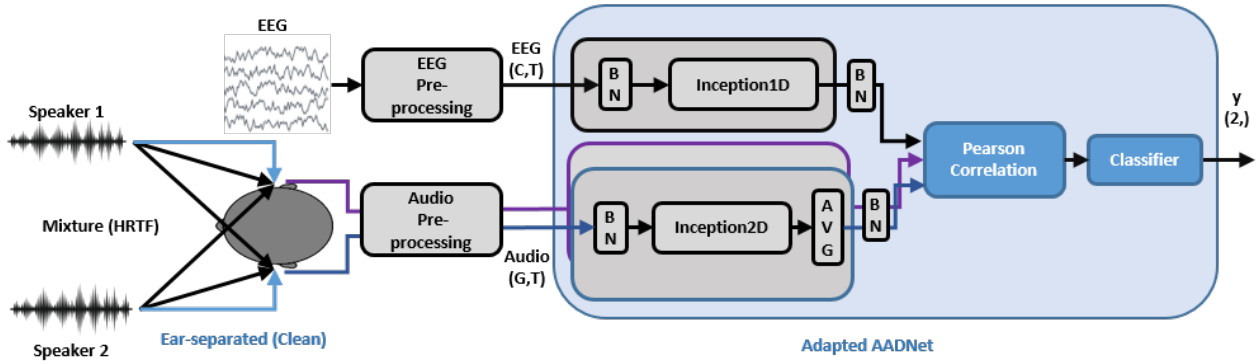

Figure 1: Overview of the proposed framework. Binaural signals (clean or spatialized mixtures) are transformed using Gammatone-based features. BN: Batch Normalization; AVG: Global Average Pooling.

were passed through a Gammatone filterbank consisting of 12 filters with 2-ERB spacing, covering the frequency range of 80−4000 Hz, thereby approximating the frequency selectivity of the human cochlea. The resulting subband signals were then converted to envelope representations by taking the absolute value and raising each subband output to a power-law compression of 0.3. This nonlinear compression mimics the compressive behavior of the auditory periphery, reducing the signal dynamic range while preserving envelope fluctuations relevant to speech perception. Finally, the envelope features were downsampled to 64 Hz to match the EEG signal rate. The overall processing pipeline of the proposed method is illustrated in Fig. 1.

4. Decoding Method

The proposed method builds upon the AADNet architecture [13], which was originally developed for attended speaker identification in auditory attention decoding tasks and utilizes an inception-based convolutional backbone for feature extraction. In this work, AADNet is used as a baseline framework due to its ability to integrate EEG and audio representations in an end-to-end manner without requiring intermediate reconstruction stages. Its inception-based design enables multi-scale temporal feature extraction which captures multi-scale patterns through parallel convolutional branches with different kernel sizes.
$X \in \mathbb{R}^{C\times T}$ denotes an EEG segment with C channels and T temporal samples. In the EEG branch, the preprocessed EEG signals are first normalized using a batch normalization layer and then passed through a modified one-dimensional inception block. This module consists of six parallel one-dimensional convolutional branches with kernel sizes of 1, 19, 25, 33, 39, and a max-pooling branch (kernel size of 3). These kernel sizes permit the model to capture both short- and long-range temporal patterns, ranging from local context to approximately 0.6 seconds at a sampling rate of 64 Hz. The outputs of the parallel branches are concatenated to form the EEG feature representation. For each ear, $e \in \{L, R\}$, the audio signal is represented using Gammatone-filtered time-frequency features $A_e \in \mathbb{R}^{G\times T}$, where G denotes the number of Gammatone filterbank channels and T the number of time frames. The audio branch follows the same architectural structure as the original AADNet, but modifies the audio feature extractor to Inception2D, utilizing two-dimensional convolutional layers to process time-frequency audio representations instead of one-dimensional convolutions. Each ear-specific audio stream is processed independently using the same encoder to preserve better-ear benefits.

The audio encoder first applies batch normalization to the input time-frequency representations, followed by the Inception2D block, which consists of multiple parallel spectro-temporal convolutional branches. The convolutional kernels are derived from the temporal kernels in the original AADNet architecture and extended to two dimensions, yielding kernel sizes of 1×1, 3×65, 3×81. This configuration allows the network to capture both long-range temporal dependencies and local spectral patterns within the audio signal. After the Inception2D module, the resulting feature maps are averaged along the frequency dimension, producing a compact temporal audio representation.
The extracted EEG and audio features are combined via a Pearson correlation computed over time, which quantifies the similarity between neural responses and their corresponding audio representations. The resulting correlation values form a feature vector, which is flattened and passed through a dropout layer and mapped to class logits via a single fully connected layer. Softmax normalizes this decision variable into class scores, from which the attended direction (left or right) is predicted, enabling spatial auditory attention decoding.

5. Evaluation Procedure

Deep learning approaches for auditory attention decoding may inadvertently exploit subtle dataset biases, such as within-trial neural fingerprints or stimulus-specific patterns, thereby artificially inflating decoding performance. It has been highlighted [13], [20] that appropriate data partitioning is essential to mitigate

such biases. To mitigate this issue, we employed a trial-based data splitting strategy and evaluated generalization using subject-independent modeling. Specifically, a leave-one-subject-out (LOSO) cross-validation scheme was adopted. In each iteration, data from one subject were held out as the test set, while data from the remaining subjects were used for training and validation. The trials from the training subjects were pooled together and further split into training and validation sets using five-fold cross-validation on a trial basis. Models were trained on the training set, tuned using the validation set, and finally evaluated on the held-out subject. This procedure was repeated for all subjects to obtain the overall subject-independent decoding performance. The training used the cross-entropy loss function for binary classification of auditory attention direction (left vs. right). The model was trained on 10-second EEG–audio segments, extracted via a sliding window across the full duration of each trial. Optimization was performed using the AdamW [21] optimizer with a learning rate of 5×10^{-5} and weight decay of 0.05, and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Early stopping was applied based on the validation loss.

Classification accuracy was evaluated as a function of the decision window length (1, 2, 5, 10, 20, and 40 seconds), with predictions aggregated over consecutive segments corresponding to each window duration. During training, a label-preserving augmentation strategy was applied in which the two competing audio streams were swapped and the corresponding attention label was inverted. No augmentation was applied during evaluation to ensure unbiased performance estimation. For each dataset, accuracies were entered into a two-way repeated-measures ANOVA with within-subject factors decision window length (6 levels) and condition (2 levels), with Greenhouse–Geisser correction for sphericity. Per-window comparisons were performed as post-hoc paired t-tests, Holm–Bonferroni corrected across the six window lengths ($\alpha = 0.05$).

6. Results

The decoding performance of the proposed model on the DTU and KUL datasets is shown in Figures 2 and 3, respectively. Decoding performance consistently improved with increasing decision window length across all signal conditions. This was expected as longer windows provide more temporal information about the neural response and stimulus dynamics, leading to more reliable estimation of the attended stream, consistent with previous studies [13]. For the DTU dataset, the binaural mixture condition achieved mean accuracies of 55.69% (1 s), 59.45% (2 s), 66.21% (5 s), 72.28% (10 s), 79.21% (20 s), and 85.46% (40 s), demonstrating the feasibility of decoding auditory attention directly from binaural mixtures. The ear-separated condition yielded highly comparable performance, with mean accuracies of 55.56% (1 s), 59.45% (2 s), 66.73% (5 s), 73.13% (10 s), 79.75% (20 s), and 85.74% (40 s). A two-way repeated-measures ANOVA revealed a significant main effect of decision window length ($F(5, 85) = 243.3$, $p < 0.001$), but no significant effect of condition ($F(1, 17) = 0.69$, $p = 0.42$) and no significant interaction between condition and decision window length ($F(5, 85) = 0.63$, $p = 0.57$). This indicates that both representations provide comparable decoding performance under the proposed framework. For the KUL dataset, decoding performance under the binaural mixture and dichotic ear-separated conditions is shown in Figure 3. Similar to the DTU dataset, decoding accuracy increased with longer decision windows. The HRTF condition achieved mean accuracies of 54.37% (1 s), 58.05% (2 s), 64.58% (5 s), 70.18% (10 s), 77.53% (20 s), and 84.55% (40 s). The ear-separated condition yielded 54.79% (1 s), 58.03% (2 s), 62.98% (5 s), 67.90% (10 s), 73.60% (20 s), and 79.49% (40 s).

In contrast to the DTU dataset, a significant main effect of condition ($F(1, 15) = 6.24$, $p = 0.025$) and a significant interaction between condition and decision window length ($F(5, 75) = 5.99$, $p = 0.005$) were found besides the main effect of window length ($F(5, 75) = 222.3$, $p < 0.001$), indicating that the difference between conditions depended on decision window length. The two conditions performed comparably at the shortest windows (1 s: 54.37% vs 54.79%; 2 s: 58.05% vs 58.03%), while the "Mixture" condition increasingly outperformed the "Ear-separated" condition at longer windows, reaching an advantage of 5.1 percentage points at 40 s, suggesting that spatialized binaural cues provide additional benefit for auditory attention decoding when sufficient temporal context is available [22]. However, in post-hoc testing with Holm–Bonferroni correction across the six window lengths, no individual window reached significance (smallest adjusted $p = 0.067$ at 40 s).

7. Conclusions

In this work, we investigated spatial auditory attention decoding in a subject-independent setting using an end-to-end deep learning framework that jointly models EEG signals and Gammatone-based audio representations. We considered both ear-separated (clean) and binaural mixture conditions. Experimental results on the DTU and KUL datasets demonstrated that Gammatone filterbank representations enable reliable decoding of the attended spatial direction under both conditions. As expected, decoding accuracy (performance) consistently improved with increasing decision window length. For the DTU dataset, binaural mixture representations achieved performance comparable to ear-separated signals across all decision window lengths, with no statistically significant differences. For the KUL dataset, while both conditions showed a similar dependence on window length, the advantage of the binaural mixture condition increased significantly with decision window length, indicating that spatial cues can provide additional benefits when sufficient temporal context is available. Overall, these findings suggest that spatial auditory attention can

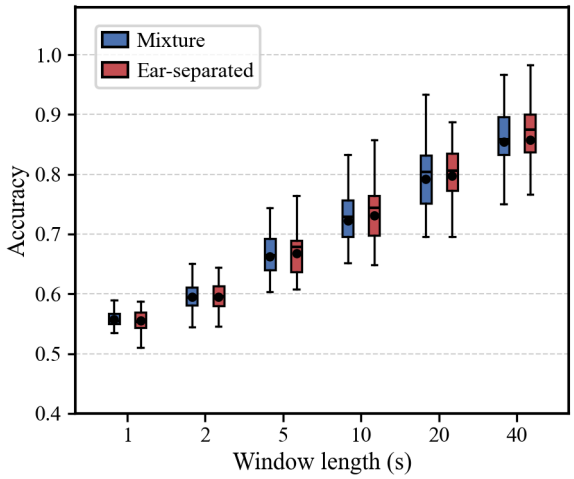

Figure 2: Left-right attention decoding accuracy for the DTU dataset for different decision window lengths. Boxplots show the performance for the HRTF-spatialized mixture (blue) and the ear-separated signals (red). The central line indicates the median across subjects, and circles denote the mean accuracy for each boxplot.

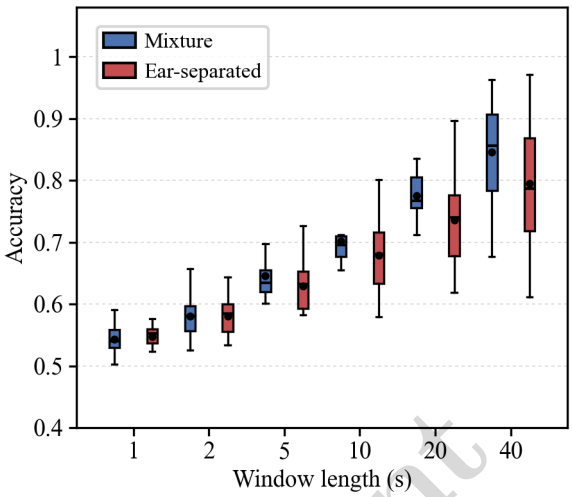

Figure 3: KUL dataset decoding accuracy for different decision window lengths. Boxplots show the performance for the HRTF (blue) and dichotic (red).

be decoded from binaural mixtures without explicit source separation, while also highlighting that the contribution of spatial information may depend on the dataset characteristics and temporal integration window. This supports the development of more realistic and deployable AAD systems.

References

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990. DOI: 10.7551/mitpress/1486.001.0001.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953. DOI: 10.1121/1.1907229.

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multitalker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012. DOI: 10.1038/nature11020.

[4] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Natl. Acad. Sci.*, vol. 109, no. 29, pp. 11 854–11 859, 2012. DOI: 10.1073/pnas.1205381109.

[5] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving precise temporal processing properties of the auditory system using continuous stimuli," *J. Neurophysiol.*, vol. 102, no. 1, pp. 349–359, 2009. DOI: 10.1152/jn.90896.2008.

[6] J. A. O'Sullivan, A. J. Power, N. Mesgarani, *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *J. Neural Eng.*, vol. 11, no. 5, p. 056 001, 2014. DOI: 10.1088/1741-2560/11/5/056001.

[7] A. Aroudi and S. Doclo, "Cognitive-driven binaural LCMV beamformer using EEG-based auditory attention decoding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 406–410. DOI: 10.1109/ICASSP.2019.8683635.

[8] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-based attention-driven speech enhancement for noisy speech mixtures using n-fold multi-channel wiener filters," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2017, pp. 1660–1664. DOI: 10.23919/EUSIPCO.2017.8081390.

[9] J. O'Sullivan, Z. Chen, S. A. Sheth, G. McKhann, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multispeaker environments without access to separated sources," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2017, pp. 1644–1647. DOI: 10.1109/EMBC.2017.8037155.

[10] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Sci. Adv.*, vol. 5, no. 5, eaav6134, 2019. DOI: 10.1126/sciadv.aav6134.

[11] E. Ceolini, J. Hjortkjær, D. D. E. Wong, *et al.*, "Brain-informed speech separation for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117 282, 2020. DOI: 10.1016/j.neuroimage.2020.117282.

[12] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding using unprocessed binaural signals," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2017, pp. 484–488. DOI: 10.1109/EMBC.2017.8036867.

[13] N. D. T. Nguyen, H. Phan, S. Geirnaert, K. Mikkelsen, and P. Kidmose, "AADNet: An end-to-end deep learning model for auditory attention decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 33, pp. 2695–2706, 2025. DOI: 10.1109/TNSRE.2025.3587637.

[14] H. Li, K. Chen, and B. U. Seeber, "Auditory filterbanks benefit universal sound source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 181–185. DOI: 10.1109/ICASSP39728.2021.9414105.

[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. DOI: 10.1016/j.neuroimage.2017.04.026.

[16] S. A. Fuglsang, D. D. E. Wong, and J. Hjortkjær, *EEG and audio dataset for auditory attention decoding*, Zenodo, 2018. DOI: 10.5281/zenodo.1199011.

[17] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995. DOI: 10.1121/1.412407.

[18] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, 2017. DOI: 10.1109/TNSRE.2016.2571900.

[19] N. Das, T. Francart, and A. Bertrand, *Auditory attention detection dataset KULeuven*, Zenodo, 2019. DOI: 10.5281/zenodo.4004271.

[20] 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," *J. Neural Eng.*, vol. 21, no. 1, p. 016 017, 2024. DOI: 10.1088/1741-2552/ad2214.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2019.

[22] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *J. Neural Eng.*, vol. 13, no. 5, p. 056 014, 2016. DOI: 10.1088/1741-2560/13/5/056014.