‘Are you even listening?’ - EEG-based detection of absolute auditory attention to natural speech with application to neuro-steered hearing devices

Arnout Roebben[1,*], Nicolas Heintz[1,2], Simon Geirnaert[1,2], Tom Francart[2] and Alexander Bertrand[1,3]

Abstract

In this study, we use electroencephalography (EEG) recordings to perform absolute auditory attention detection (aAAD), i.e., determine whether a subject is actively listening to a presented speech stimulus or not. More precisely, we aim to discriminate between an active listening condition, and a distractor condition where subjects passively listen to the speech stimulus while performing another cognitive task. To this end, we re-use an existing EEG dataset where the subjects watch a silent movie as a distractor condition, and introduce a new EEG dataset with two other distractor conditions (silently reading a text and performing arithmetic exercises). We focus on two EEG features, namely neural envelope tracking (NET) and spectral entropy (SE). We find significantly higher NET and lower SE in the active listening condition compared to the distractor conditions, which for the SE is the reverse of what was previously found for an active listening versus passive listening condition (without any distractors). In addition, aAAD is used in the context of a selective auditory attention decoding (sAAD) task, where the goal is to decode to which of two competing speakers the subject is attending, which is a core task in the context of so-called neuro-steered hearing devices. We show that evaluating sAAD performance only on segments of active listening improves sAAD performance when detecting these active listening segments as having higher NET, whereas the reverse trend is observed when detecting these segments as having lower SE. We conclude that NET is a more reliable metric for aAAD as it is consistently higher for the active listening condition, whereas the relation of the SE between the active listening and passive listening conditions seems to depend on the nature of the distractor task. Consequently, NET shows the most promise for aAAD and to detect auditory inattentive segments in neuro-steered hearing devices.

Index Terms

Absolute auditory attention detection, Selective auditory attention decoding, EEG, Neural envelope tracking, Spectral entropy, Neuro-steered hearing device, Brain-computer interface.

I. Introduction

The human auditory system is able to focus on a single speaker of interest while filtering out the speech of competing speakers in a so-called cocktail party scenario. This process is often referred to as selective (auditory) attention. Recently, there has been

[1]KU Leuven, Dept. Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.

[*]arnout.roebben@esat.kuleuven.be

[2]KU Leuven, Dept. of Neurosciences, ExpORL. Herestraat 49 bus 721, B-3000 Leuven, Belgium

[3]Leuven.AI - KU Leuven institute for AI

a wide interest in decoding this selective auditory attention based on neural activity, which is often referred to as (selective) auditory attention decoding ((s)AAD) [1]–[4]. One possible use case for such an sAAD algorithm is to control the speech enhancement algorithm of a hearing device, such that it is able to decide which speaker should be enhanced and which speakers should be suppressed [5].

Selective auditory attention can be decoded from, e.g., electroencephalography (EEG) recordings, either based on the principle of neural envelope tracking (NET) [3], [4], or based on differences in the features extracted from general neural activity [1], [6]–[8]. The former exploits correlations between the EEG and the amplitude envelope of the attended speech stimulus [6], where higher correlation coefficients are observed for the attended speaker than for the unattended one(s) [1]. Alternatively, the latter aims to decode speaker-dependent features from the EEG alone (without using the stimulus), for example features that relate to the spatial location of the attended speaker [3], [4].

In a neuro-steered hearing device, these sAAD decoders consequently assume the subject is actively listening to any one of the competing speakers. However, such decoders should not make decisions whenever the subject is *not* actively listening to any of the speech stimuli to avoid arbitrary speaker selections. Such functionality would require an algorithm that can discriminate between active and passive listening conditions from EEG data, which is the main focus of this paper. Discriminating between active and passive listening would also be useful in time-adaptive decoders in order to only update the decoder coefficients when the subject is actively listening [9], [10]. Furthermore, besides the application of neuro-steered hearing devices, this functionality could be useful as an objective, general tool to measure the state of active listening during auditory EEG experiments.

While decoding selective attention to competing speakers has been widely studied, the literature about decoding active versus passive listening is less extensive. In [11], the envelope tracking of subjects actively attending an auditory stimulus was shown to be significantly higher than that of subjects watching a silent movie, while ignoring the auditory stimulus. To quantify the level of (active) listening across both conditions, linear decoders reconstructed the speech envelope from the EEG [11], and the resulting correlation between the reconstructed envelope and the ground truth envelope was shown to be significantly higher when actively focusing on the auditory stimulus. In [12], a similar scenario was investigated using the peak cross-correlation between the EEG and the speech envelope as a measure for envelope tracking. However, no difference across the active versus passive listening conditions was found in this study.

In [13], the spectral entropy (SE) of the normalised EEG power spectral densities (PSDs) was used to quantify the level of sustained attention to an auditory stimulus. In this study, subjects were instructed to attend to an auditory stimulus, without any distractor conditions. The SE was then used to discriminate between high and low active listening segments, which are assumed to arise naturally as the subject's focus might vary during the task. This SE characterises the uncertainty in distribution and, as a consequence, higher SE levels map to less regularity and predictability. Previously, it has been shown experimentally that the SE was able to correctly predict anaesthetic depth [14], [15], respiration movements [15], sleep stages [15], and imagined finger movement [15], possibly due to changes in EEG regularity and predictability [14], [15]. Similarly, the SE allowed to discriminate subjects in rest from subjects performing mental

arithmetic [16], and subjects in rest from subjects fixated on flashing patterns [17]. In [13], it was found that training neural decoders on high SE segments resulted in an improved decoding performance when evaluating on the full validation set. However, no performance difference was found between evaluating the decoders on high versus low SE validation segments. Corresponding to these results, higher SE levels were hypothesised to correspond to higher auditory attention levels.

In this paper, we aim to detect whether or not a subject actively listens to a single speaker and compare across different distractor conditions. To this end, we introduce a new EEG dataset with 10 subjects, wherein subjects are asked to either actively listen to a speech stimulus or to ignore it while silently reading a text or solving arithmetic exercises. Next to this dataset, we reuse a dataset from [11] in which the distractor condition consists of watching a silent movie [11]. We then investigate whether both NET and SE can distinguish between the active listening condition and any of these distractor conditions. Finally, we combine this with an sAAD task with two competing speakers, in which the attended and unattended speaker needs to be discriminated. Segments labelled as passive listening by the NET and SE features are removed at validation time, attempting to take out segments in which the subject is not paying attention to any of the speakers, hence for which no truthful sAAD decision can be made.

In what follows, the modulation of the active listening towards a single auditory stimulus, i.e., detecting when a subject is in a state of (in)active listening, will be referred to as *absolute auditory attention detection (aAAD)*. The selective attention in a multispeaker scenario, i.e., decoding to whom the subject is listening, will be referred to as *selective auditory attention decoding (sAAD)*, conform the literature [5]. Fig. 1 illustrates the difference between both scenarios, and shows how to combine them within the same framework.

Our research is complementary to, yet distinct from, [11] and [13]. First, we focus on natural speech within a broader scope of conditions, as opposed to [13], where no distractor conditions were present, and to [11], where only a silent movie distractor condition was used and where only artificial, standardised sentences were used as speech stimuli at validation time, requiring little semantic processing in the brain [11]. Second, we compare both the NET and SE features to discriminate between the active listening condition and distractor conditions, whereas the previous studies mostly focused on either one of them and consequently do not have such a direct comparison. Finally, we apply aAAD to the sAAD task by only evaluating the sAAD performance on segments where the subject is signalled to listen actively.

We will demonstrate that higher SE does not necessarily correspond to a higher auditory attention (as hypothesised in [13]), i.e., SE shows different trends depending on the choice of the alternative (passive listening) condition. We will also demonstrate that the NET metric shows consistent behaviour, i.e., it is always higher in the active listening condition compared to the passive listening condition, and is therefore the better choice for aAAD.

This paper is structured as follows. First, we describe the algorithmic methodology to extract the NET and SE features for aAAD, and explain how to use these to detect active versus passive listening conditions. Thereafter, we briefly review the sAAD task and methodology. The datasets, data processing, and experimental setup are subsequently described. Finally, we discuss the corresponding results and conclusions.

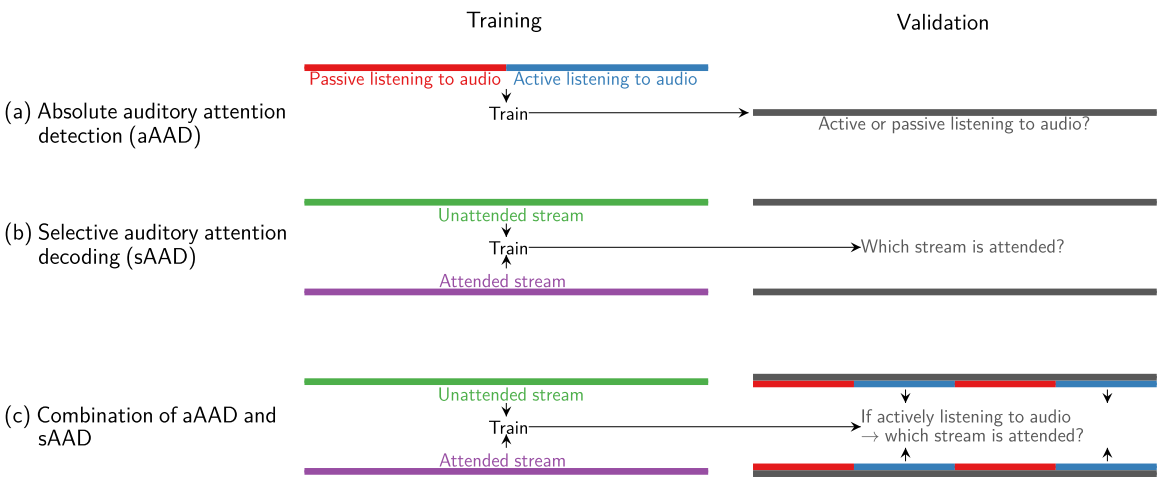

Fig. 1. Overview of absolute auditory attention detection (aAAD), selective auditory attention decoding (sAAD) and their combination. (a) In aAAD, the subject is attending a single auditory stimulus, either in a state of attention or inattention to the speech (i.e., active or passive listening). At training time, data from these two conditions are leveraged to train the features and classifiers. At validation time, these are used to assess whether the subject is actively listening or not. (b) Contrarily, in sAAD, multiple auditory stimuli are simultaneously present at each given moment. At training time, the data are again leveraged to train the features and classifiers. At validation time, however, the sAAD features and classifiers are used to discriminate to which stimulus the subject is attending. (c) Both setups can be combined, where the sAAD performance is evaluated only on segments where active listening is signalled, removing segments of auditory inattention.

II. Absolute auditory attention detection

To detect whether a subject is in a state of active listening, i.e., perform aAAD, we will extract two features from the EEG:

1) **Neural envelope tracking (NET):** In the presence of a speech stimulus, the neural response tracks certain characteristics of that speech stimulus, such as its envelope, i.e., the slow variations over time [1], [6]–[8], [13]. This envelope tracking is hypothesised to be more strongly present when the subject is more attentive to the speech [1], [11].
2) **Spectral entropy (SE):** Different tasks or conditions result in different neural activity across the brain [4], [13], [14], [16]–[18]. These differences in neural activity can, e.g., be quantified using the SE of the EEG recorded at a particular scalp location. The SE is hypothesised to quantify the predictability or regularity in the EEG activity, as will be detailed infra [13]–[17]. In [13], it was hypothesised that the SE is higher in active listening compared to passive listening conditions.

The NET feature is discussed first, and the SE feature thereafter.

A. Neural envelope tracking

To evaluate the degree of envelope tracking, a decoder consisting of a linear spatio-temporal filter is applied to the EEG signals to reconstruct the envelope of the speech stimulus. This is commonly achieved by minimising the squared error between the original and reconstructed speech envelope [1], [2], [11]. The degree of envelope tracking can subsequently be assessed by computing the correlation coefficient $\rho(.)$ between the reconstructed and ground truth envelope.

Let $y(t)$ represent the target speech envelope at sample time t, and $x_c(t)$ the sample of the c-th EEG channel at sample time t. The goal is to reconstruct the T subsequent target speech envelope samples $\mathbf{y} = \begin{bmatrix} y(0) & y(1) & \cdots & y(T-1) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{T\times 1}$ from channel-concatenated and time-lagged EEG signals $X \in \mathbb{R}^{T\times LC}$, defined as:

$$\begin{aligned} X &= \begin{bmatrix} X_1 & X_2 & \cdots & X_C \end{bmatrix} \\ X_c &= \begin{bmatrix} x_c(0) & x_c(1) & \cdots & x_c(L-1) \\ x_c(1) & x_c(2) & \cdots & x_c(L) \\ \vdots & \vdots & & \vdots \\ x_c(T-1) & 0 & \cdots & 0 \end{bmatrix}, \end{aligned} \tag{1}$$

where C denotes the number of EEG channels, T the number of training samples, and L the number of time lags. This reconstruction is achieved by designing a decoder $\hat{\mathbf{d}} \in \mathbb{R}^{LC\times 1}$ that minimises the mean squared error between the ground truth envelope $\mathbf{y}$ and the reconstructed envelope $X\mathbf{d}$ [1], [9]:

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmin}} \, ||\mathbf{y} - X\mathbf{d}||_2^2, \tag{2}$$

of which the solution equals [1], [9]:

$$\hat{\mathbf{d}} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\mathbf{y} \tag{3a}$$

$$= R_{xx}^{-1}\mathbf{r}_{xy}. \tag{3b}$$

Herein, $R_{xx} \in \mathbb{R}^{LC\times LC}$ and $\mathbf{r}_{xy} \in \mathbb{R}^{LC\times 1}$ respectively denote the estimated EEG auto-correlation matrix and EEG-envelope cross-correlation vector.

At validation time, this decoder is applied to the validation data that were held out during training. The EEG data $X^{(val)} \in \mathbb{R}^{T_v\times LC}$ are then used to reconstruct the speech envelope $\hat{\mathbf{y}} \in \mathbb{R}^{T_v\times 1}$ of length T_v:

$$\hat{\mathbf{y}} = X^{(val)}\hat{\mathbf{d}}. \tag{4}$$

The correlation $\rho(\hat{\mathbf{y}}, \mathbf{y}^{(val)})$ between this reconstructed envelope $\hat{\mathbf{y}}$ and the ground truth envelope $\mathbf{y}^{(val)} \in \mathbb{R}^{T_v\times 1}$, computed over T_v samples, is hypothesised to be higher when actively listening [11]. Therefore, we use this stimulus correlation as a first feature to discriminate between active and passive listening. As per [11], the Spearman correlation coefficient, a rank-ordered Pearson correlation, is used in this work [19].

B. Spectral entropy

SE is a measure for the uncertainty of a random variable by characterising the peakedness of the probability density function [20]. Herein, large SE values map to high uncertainty and hence lower predictability. Since power spectral densities (PSDs) after normalisation are non-negative and sum to one like probability density functions, the SE can also be applied to these normalised PSDs in the frequency domain in order to describe the peakedness of these PSDs [14], [16]. This so-called SE can be hypothesised to be used to leverage spectral differences, where higher SE levels correspond to less predictability in the neural response. Let $\mathcal{P}_{x_c}(f) \in \mathbb{R}$ represent the normalised PSD

estimate of the EEG signal $x_c(t)$ in channel c, then the SE in channel c and across the frequency band $f_1 - f_2$ is defined as:

$$\mathrm{SE}_c\left[f_1 - f_2\right] = -\sum_{f=f_1}^{f_2} \mathcal{P}_{x_c}(f) \log_2\left(\mathcal{P}_{x_c}(f)\right). \tag{5}$$

Previously, this SE allowed to the predict anaesthetic depth [14], [15], respiration movements [15], sleep stages [15], and imagined finger movement [15]. In addition, the SE allowed to discriminate between subjects in rest from subjects performing mental arithmetic [16], and subjects in rest from subjects fixated on flashing patterns [17]. It was hypothesised, that this predictive and discriminative power arises from the regularity and predictability differences in the EEG activity as characterised by the SE. In [13], higher SE values were hypothesised to correlate with higher levels of auditory attention, as training decoders on the high SE segments outperformed the decoders trained on the low SE segments when evaluating on the full validation set [13].

III. Application to selective auditory attention decoding

As opposed to detecting when the subject is attentive or inattentive to a specific speech stimulus, in the sAAD setting, we aim to decode to which speaker the subject is attending in a multi-speaker scenario [1], [2], [5]. This corresponds to a selective auditory attention setup, as illustrated in Fig. 1.

In an sAAD framework, similar neural decoders can be used to reconstruct the envelope of the attended speaker [1], [2], [5]. Here, the decoder $\hat{\mathbf{d}}_{sAAD} \in \mathbb{R}^{LC\times 1}$ is computed using the envelope $\mathbf{y}_a \in \mathbb{R}^{T\times 1}$ of the attended speaker, while the envelope(s) of the unattended speaker(s) is not used during training [1], [2], [5], [9]:

$$\hat{\mathbf{d}}_{sAAD} = \underset{\mathbf{d}_{sAAD}}{\operatorname{argmin}} = ||\mathbf{y}_a - X\mathbf{d}_{sAAD}||_2^2. \tag{6}$$

At validation time, the output of the decoder is correlated with the speech envelopes of all speakers over T_v samples. The speaker that yields the highest correlation $\rho(.)$ between the reconstructed envelope $\hat{\mathbf{y}} = X^{(val)}\hat{\mathbf{d}}_{sAAD}$ and the speaker envelope $\mathbf{y}_i^{(val)} \in \mathbb{R}^{T_v\times 1}$ (for speaker i) is decoded as the attended one [1], [2], [9]. In the case of two speakers, the decision process can be described as:

$$\begin{aligned} &\text{If } \rho\Big(\hat{\mathbf{y}},\, \mathbf{y}_1^{(val)}\Big) > \rho\Big(\hat{\mathbf{y}},\, \mathbf{y}_2^{(val)}\Big), \text{ Speaker 1 attended} \\ &\text{If } \rho\Big(\hat{\mathbf{y}},\, \mathbf{y}_1^{(val)}\Big) < \rho\Big(\hat{\mathbf{y}},\, \mathbf{y}_2^{(val)}\Big), \text{ Speaker 2 attended.} \end{aligned} \tag{7}$$

As illustrated in Fig. 1, both aAAD and sAAD can be combined. By only evaluating the sAAD performance on those segments where the subject is signalled to be in a state of active listening, the auditory inattentive segments are removed. We hypothesise that removing these auditory inattentive segments in which the subject is not paying auditory attention will improve the sAAD decision process.

IV. Experimental procedures

To validate our methods, a new dataset was recorded, which is complementary to the datasets of [11] and [21], which will also be used in our study. The dataset descriptions are given first and the corresponding data processing thereafter. The experimental setup is subsequently formulated, as well as the hypothesis tests.

A. Datasets

In this work, we utilise three datasets: Dataset I is a newly recorded dataset that deals with subjects attending to an auditory stimulus versus subjects ignoring that stimulus while focusing on silently reading a text or solving arithmetic exercises. Dataset II originates from [11] where subjects were either instructed to attend to an auditory stimulus or to ignore that stimulus while focusing on a silent movie. Finally, Dataset III is an sAAD dataset of [21].

1) Dataset I: **Goal:** The goal of this experiment is to investigate the neural differences between a setting where the subject is actively listening to a speech stimulus, and a distractor condition where the subject ignores the auditory stimulus while focusing on silently reading a text or solving arithmetic exercises.

Subjects: 10 Dutch-speaking subjects (5 male, 5 female), between 21 and 27 years old, participated in the experiment. These subjects were unpaid volunteers and signed an informed consent, approved by the Social-Societal Ethics Committee at KU Leuven. Experiments were conducted according to these regulations.

Equipment: A 24-channel Smarting mobile EEG recording system was utilised [22]. The experiment was conducted in a non-radio frequency shielded room and the scalp of the subjects was treated with electroconductive gel while fitting the EEG cap. Raw EEG data were converted into MATLAB-compatible files using OpenVibe software [23].

Presentation structure: The speech stimuli consisted of children stories in Dutch [24]. The experiment consisted of four phases, and a different story was used in each phase. The narrators were all male, except in phase two. An overview of the presentation structure within each phase can be found in Fig. 2. In the first phase, the subjects were instructed to listen to a story ('Bianca en Nero') for 20 minutes, referred to as 'Audio' in Fig. 2. To engage the subjects, they were told upfront that a questionnaire, concerning the content of that auditory fragment, had to be filled in at the end of the phase.

In the second phase, the subjects again had to listen to a story ('Ver van het kleine paradijs') for 11 minutes, after which a questionnaire had to be filled in about the content of the fragment. However, during this second phase, specific tasks were displayed on a computer screen at certain times to manipulate the attention to the speech. A first task consisted in solving as many arithmetic exercises (e.g., $1012 - 448 = 564$) as possible within one minute, in order to reduce the attention to the speech. The subjects were informed that no questions in the questionnaire would originate from those parts of the story during which the arithmetic exercises had to be conducted. As a second task, the subjects were informed that they should listen even more attentively for one minute since questions were certainly to be asked about this following minute of the fragment, requiring increased attention to the speech. The presentation structure of this phase 2 is shown in Fig. 2, where 'Mathematics' refers to segments where subjects were instructed to solve arithmetic exercises and where 'Question' refers to segments where subjects were informed that a question had to be answered about the next minute in the story. The third phase used the same setup as the second phase, but the presentation structure was changed as shown in Fig. 2 and a different story stimulus was used ('Milan').

In a final, fourth phase, the subjects were instructed to focus on silently reading a text while ignoring the auditory stimulus ('Eline'), a condition referred to as 'Text' in Fig. 2. This text was provided as a printed document, and consisted of a Dutch version of the short story 'Vergif' written by Roald Dahl [25]. As before, subjects were informed in

advance to fill in a questionnaire at the end of the fragment. In this phase, however, the questionnaire related solely to the content of the text, resulting in a no-attention condition to the speech. Between each of the four phases, a short pause was inserted to give the subjects time to fill in the corresponding questionnaires. The stories were played through the stereo speakers of a laptop. At the start of the experiment, the subjects could determine the volume themselves, such that they were comfortable during the experiments. To keep the subjects motivated throughout the experiment, the subjects were informed that a gift card was to be handed out to the three best-scoring participants on the combined results of the questionnaires and arithmetic exercises. No preprocessing of the auditory stimuli took place before presenting them to the subjects.

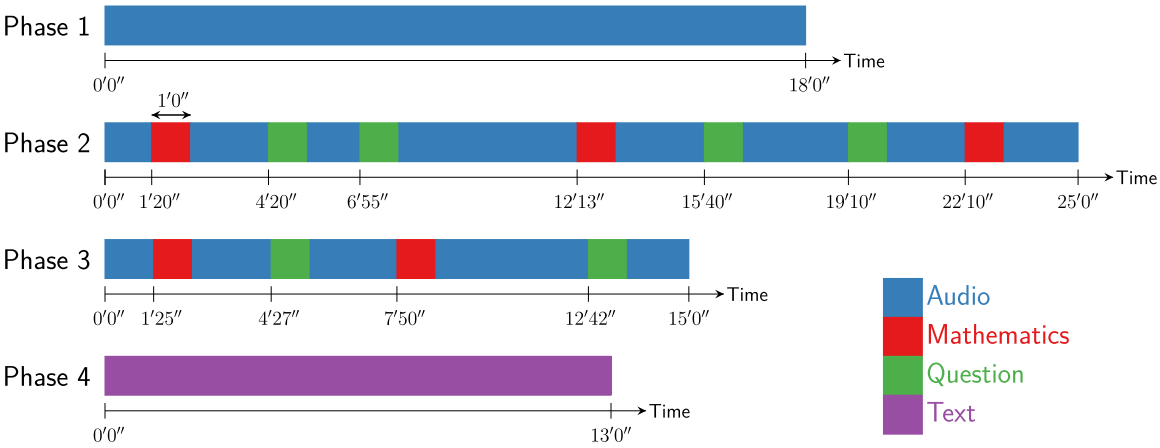

Fig. 2. The presentation structure of Dataset I. During phase 1, the subjects were instructed to attend an auditory stimulus and answer questions about the content afterwards ('Audio'). During phase 2 and phase 3, this structure was interleaved with specific tasks to be completed in order to modulate the subjects' state of auditory attention. During the 'Mathematics' task, subjects needed to solve arithmetic exercises, reducing the auditory attention. During the 'Question' task, subjects needed to increase their auditory attention as a question certainly was to be asked about that part of the fragment. During phase 4, the subjects were instructed to silently read a text while ignoring the auditory stimulus.

2) Dataset II: Dataset II is a subset of the dataset used in [11]. This subset consists of 7 normal-hearing subjects, who all participated in two experiments. In the first experiment, subjects were instructed to attend a Dutch continuous auditory stimulus (story 'Milan'). In the second experiment, the subjects were instructed to attend a silent, subtitled cartoon movie while ignoring an auditory stimulus. For each subject, about 15 minutes of data are available for both conditions. The EEG data consist of 64 channels. The interested reader can find a more detailed description of this dataset in [11].

3) Dataset III: Dataset III corresponds to the sAAD dataset of [21]. Herein, 16 normal-hearing subjects were exposed to a competing speaker scenario with two speakers, and were instructed to attend to one of the speakers while ignoring the other one. As auditory stimuli, four Dutch children's stories were used. In total, approximately 72 minutes of recorded data per subject are available, using a 64-channel EEG cap. A more detailed description of this dataset can be found in [2].

In the present study, we do not further consider the segments corresponding to the 'Question' condition of Dataset I (green segments in Fig. 2), to ensure a consistent active listening condition between Dataset I and Dataset II. Furthermore, the data of the

text reading and arithmetic exercise solving distractor conditions are considered as one distractor condition, and hence their data are concatenated to increase the amount of data. We justify this approach since we want to binary discriminate between the active listening condition and the passive listening condition, although the amount of data for the text reading distractor condition is larger.

B. Data processing

1) Preprocessing: The preprocessing framework is chosen the same as in [11], yet, with an additional muscle artefact removal step. The speech envelope is extracted from the audio data based on the procedure proposed in [2]. First, the raw audio data are filtered using a gamma-tone filterbank consisting of 28 filters, with centre frequencies between $50\,\mathrm{Hz}$ and $5\,\mathrm{kHz}$, spaced according to 1 equivalent rectangular bandwidth [26], [27]. The output signal of each filter is thereafter transformed with a power law, i.e., $|y_k(t)|^{0.6}$, $k = 1..28$. The signals are then linearly combined with equal weight and downsampled to $256\,\mathrm{Hz}$, with the built-in low-pass anti-aliasing filter in MATLAB 2021b, to extract the envelope.

The EEG data are first also downsampled to $256\,\mathrm{Hz}$. Next, muscle and eye artefacts are removed using a multichannel Wiener filter (MWF) approach according to [28], [29]. Since the MWF is a data-driven filter, the data for all conditions for the same subject are filtered using the same MWF in order to avoid condition-dependent preprocessing. This approach requires artefact annotation, for which heuristic detection mechanisms are utilised [11], [30], further specified in Appendix. The EEG data are subsequently re-referenced to the Cz-channel. Both the EEG and envelope are either bandpass filtered using a Chebyshev type 2 filter with cutoff frequencies tailored to the delta band ($0.5-4\,\mathrm{Hz}$) (for the NET calculation), or passed through unfiltered (for the SE calculation), as the SE frequency selection will be performed directly on the PSD in correspondence to (5). Finally, both signals are downsampled to $128\,\mathrm{Hz}$ and periods of longer silence ($>0.25\,\mathrm{s}$) are removed.

Additionally, for Dataset I, the EEG data are first linearly detrended to compensate for the strong baseline drift. The first downsampling before artefact removal is dropped for Dataset III, as the EEG data in this case are only available at 128 Hz and highpass filtered above $0.5\,\mathrm{Hz}$ [21].

2) Hyperparameters: The following hyperparameters are adhered to:

- **NET:** The lag value L is chosen equal to $L = 64$, corresponding to a time window of 500 ms, to capture the relevant neural response [8], [11]. For the two 64-channel EEG datasets (Dataset II and Dataset III), the correlation matrix in (3b) is of dimension $LC \times LC$ where $LC = 4096$, which is quite large. For the sake of numerical stability and to avoid overfitting, we, therefore, apply L2-regularisation $R_{xx}^{(reg)} = \alpha R_{xx} + \beta I_{LC\times LC}$, wherein α, β are computed according to the method presented in [31], of which a MATLAB implementation is available [32].
- **SE:** In correspondence to [13], the SE is calculated in a network of frontal, parietal-occipital and occipital channels (Fp1, Fpz, Fp2, AF7, AF3, AFz, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2 for the 64-channel EEG cap and Fp1, Fp2, F7, F8, Fz, O1 and O2 for the 24-channel EEG cap) on a frequency range spanning the alpha ($8-13\,\mathrm{Hz}$) and beta ($13-30\,\mathrm{Hz}$) bands, i.e., setting $f_1 = 8\,\mathrm{Hz}$ and $f_2 = 30\,\mathrm{Hz}$ in (5). The average of the SE values over these

channels is utilised as an active listening feature. The PSD estimate is calculated using the multitaper spectral analysis using 7 Slepian tapers with a frequency spacing of $0.5\,\mathrm{Hz}$ [33].

C. Experimental setup

1) Absolute auditory attention detection: To study how the NET and SE features vary across the different active listening and distractor conditions, a 10-fold cross-validation is conducted per subject on Dataset I and Dataset II, where the folds are split chronologically. Regarding the NET, the decoder is trained solely on the active listening data of the left-in folds in order to reconstruct the speech envelope when attending to the speech stimulus, whereas regarding the SE, no training phase is required. The left-out fold is partitioned into windows, quantifying the amount of time given to decide whether the subject is in a state of active listening. To this end, window lengths of $5\,\mathrm{s}$, $10\,\mathrm{s}$, $20\,\mathrm{s}$, $30\,\mathrm{s}$ and $60\,\mathrm{s}$ are used.

Subsequently, the classification accuracy of the features, defined as the ratio of the number of correctly predicted windows and the total amount of windows, is analysed to discriminate between high versus low attention windows. To this end, an equal amount of active and passive listening windows are present. A linear discriminant analysis (LDA) classifier is used, although other classifiers could be used as well [34]. This LDA classifier is trained using a nested 10-fold cross-validation procedure, where the folds are split chronologically. As for the NET, the covariance matrix estimate in the LDA classifier is regularised by adding a weighted identity matrix to the estimate according to [31].

2) Combination of aAAD and sAAD: Next, we aim to analyse whether aAAD can be incorporated into an sAAD framework. A 10-fold cross-validation is conducted on the sAAD task of Dataset III, where the folds are split chronologically and the left-out folds are split into $60\,\mathrm{s}$ windows. The performance of these sAAD decoders is evaluated on a proportion of the left-out folds spanning the $x\%$, $x = 0, \ldots, 100$, of the segments signalled by the NET and SE features as having the highest absolute auditory attention. Regarding the NET, this selection of active listening segments is performed by applying the sAAD decoders on the left-out folds, computing the maximum of the correlation between the decoder output $\hat{\mathbf{y}}$ and both speech envelopes $\mathbf{y}_1$ and $\mathbf{y}_2$ $\left(\max\left(\rho\left(\hat{\mathbf{y}}, \mathbf{y}_1^{(val)}\right), \rho\left(\hat{\mathbf{y}}, \mathbf{y}_2^{(val)}\right)\right)\right)$ over $60\,\mathrm{s}$ windows, and selecting the segments which have the $x\%$ highest values of this feature. Regarding the SE, the $x\%$ lowest SE values on the left-out folds are selected in correspondence to the results on Dataset I and Dataset II as will be detailed infra. We hypothesise that by evaluating the sAAD performance only when subjects are signalled to be actively listening, the sAAD performance will increase as segments are removed where the subjects are not actively listening to any speech stimulus and no truthful sAAD decision can be made.

D. Hypothesis tests

Hypothesis tests are performed using the two-sided Wilcoxon signed rank test [35] and the resulting p-values are corrected for multiple comparisons using the Benjamini-Hochberg correction [36]. To assess the statistical significance of slopes, a linear regression model is fitted, and the t-test [37] on the coefficient corresponding to the slope is performed under the null hypothesis that the corresponding coefficient is zero. All hypothesis tests are performed with respect to a significance level $\alpha = 0.05$.

V. Results

A. Absolute auditory attention detection

Fig. 3 shows the individual data points and the per-subject averages, across windows and folds, for the cross-validation on Dataset I and Dataset II for the $10\,\mathrm{s}$ and $30\,\mathrm{s}$ windows. The two-sided Wilcoxon signed rank test on the per-subject averages indicates a significant difference between the active listening and distractor conditions for both the NET and SE features $(\max(p) = 0.047)$. Herein, the NET correlations attain higher values in the active listening than in the distractor conditions, whereas the SE attains lower values in the active listening than in the distractor conditions. For the NET, these results are consistent with [11], where higher NET correlations were found for subjects actively listening than for subjects passively listening while watching a silent movie. Nevertheless, while [11] used artificial, standardised sentences, we used natural speech and introduced two new distractor tasks. For the SE, the reverse trend is observed to [13], where higher SE values were found to correlate with active listening than with passive listening without any distractor.

These feature values correspond to classification accuracies using the LDA classifier as shown in Fig. 4. Both the NET and the SE features outperform chance level, computed as the upper bound of a 95% one-sided confidence interval of a binomial distribution with success rate 0.5. Both features are subsequently able to discriminate between the active and passive listening condition, wherein the NET visually seems to suffer more from shorter window lengths than the SE.

B. Combination of aAAD and sAAD

Fig. 5 shows the sAAD accuracy evaluated on the $x\%$, $x = 0, \ldots, 100$, highest active listening segments signalled as the highest NET correlation and lowest SE segments, in correspondence with the results on Dataset I and Dataset II. As hypothesised, when evaluating the sAAD performance on the highest NET correlation segments, the sAAD accuracy increases as the proportion of the validation set decreases, corresponding to keeping the $x\%$ highest attention windows. This trend is, furthermore, significant $(\max(p) = 7.00 \cdot 10^{-7})$. However, evaluating sAAD performance on the lowest SE segments, contrary to the results on Dataset I and II, shows a significant negative trend when decreasing the proportion of the validation set used $(\max(p) = 0.0023)$. This trend is different from the SE results on Dataset I and Dataset II, but consistent with the results of [13].

VI. Discussion

As shown in Fig. 3 and Fig. 4, both the NET and SE features yield significant differences between the active listening and distractor conditions. Nevertheless, while the NET correlations are consistently higher in the active listening condition than in the distractor conditions in line with [11], this relation is inconsistent for the SE. Indeed, contrary to the findings in [13], where a higher SE is found for the active listening condition than for a passive (not performing any task) condition, the results on Dataset I and Dataset II show a higher SE on the movie watching, arithmetic exercise solving and text reading distractor conditions (with passive listening) than on the active listening condition without such distractor task. In agreement with these results the sAAD accuracy

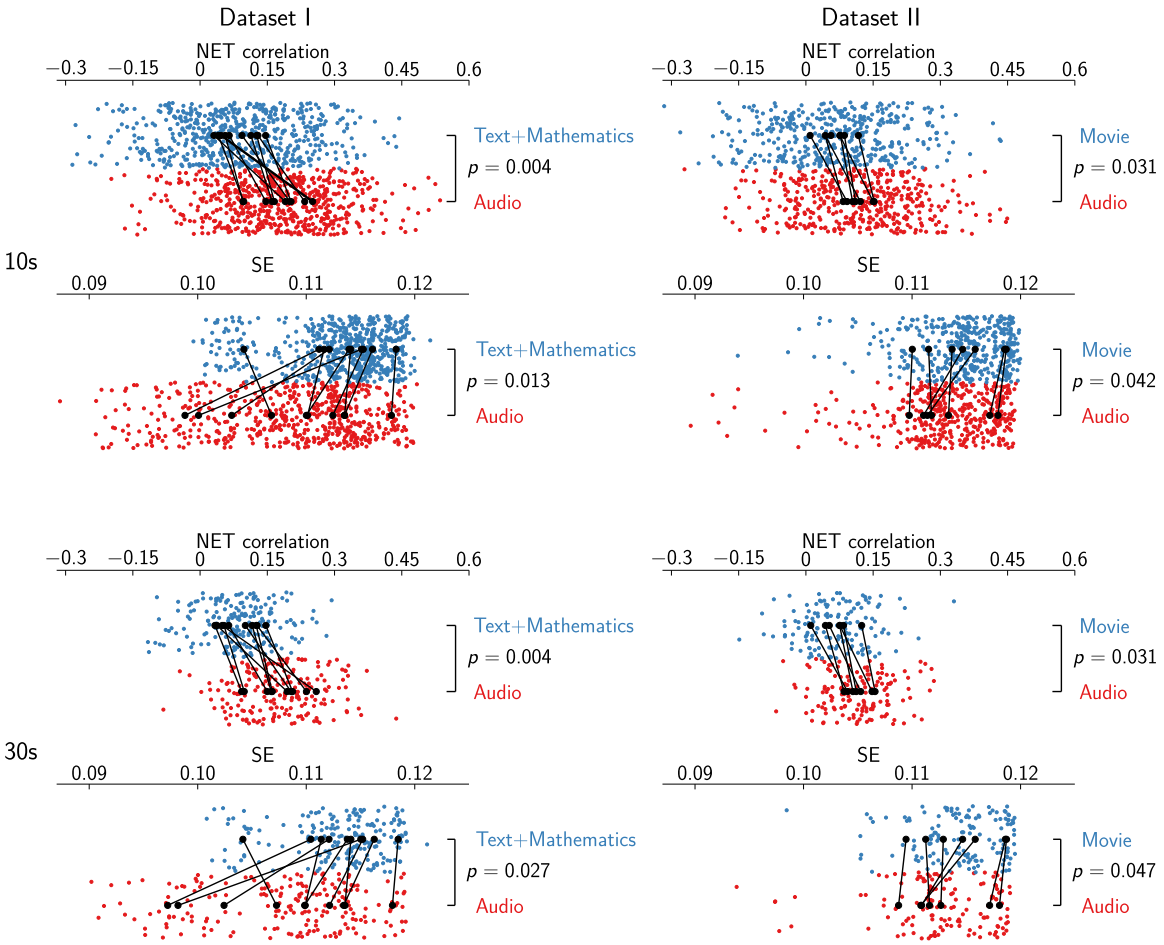

Fig. 3. The neural envelope tracking (NET) correlations and spectral entropy (SE) on Dataset I and Dataset II. The NET correlation is significantly higher when actively listening ('Audio') than when passively listening during the movie watching ('Movie'), arithmetic exercise solving ('Mathematics') and text reading ('Text') distractor conditions. Contrarily, the SE in a network of frontal, parietal-occipital and occipital channels on a frequency range spanning the alpha and beta bands is significantly lower when actively listening than when passively listening during a distractor condition. This experiment has been repeated for window lengths of 10 s and 30 s. Red and blue dots represent the individual data points, whereas the black dots represent the data points per subject averaged across windows and folds. Lines connect the per-subject averages and p-values are noted on the right of the data.

increases when evaluating on a reduced portion of high NET correlation segments. However, the sAAD accuracy decreases when evaluating on a reduced portion of low SE segments.

As the SE characterises the uncertainty in distribution on a network of frontal, parietal-occipital and occipital channels on a frequency range spanning the alpha and beta bands, our hypothesis is that the SE does not detect the absolute auditory attention, but rather reflects the differences in cognitive resources spent on each task (either active listening or a distractor condition). Indeed, in Dataset I and Dataset II, the passive listening conditions are all coupled to specific distractor conditions (movie watching, arithmetic exercise solving and text reading) that require mental resources. The relation for SE between active and passive listening, consequently, seems to depend on the specific distractor task, reducing its relevance as a generic feature for absolute auditory attention detection.

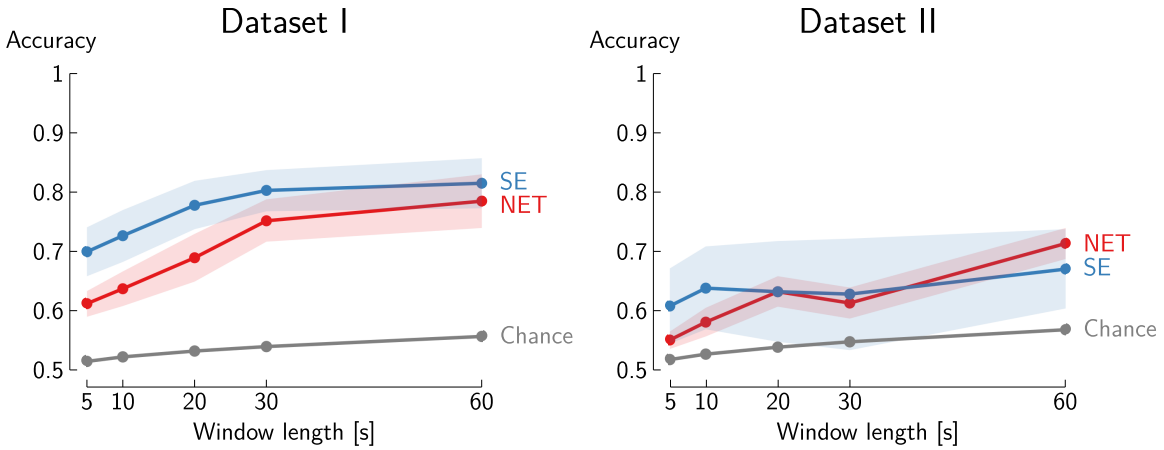

Fig. 4. Classification accuracies of the absolute auditory attention detection (aAAD) task on Dataset I and Dataset II using the neural envelope tracking (NET) correlations and the spectral entropy (SE). The mean accuracy across the subjects is marked in bold and the standard deviation is shown as shading. On both datasets, the methods yield accuracies above chance level.

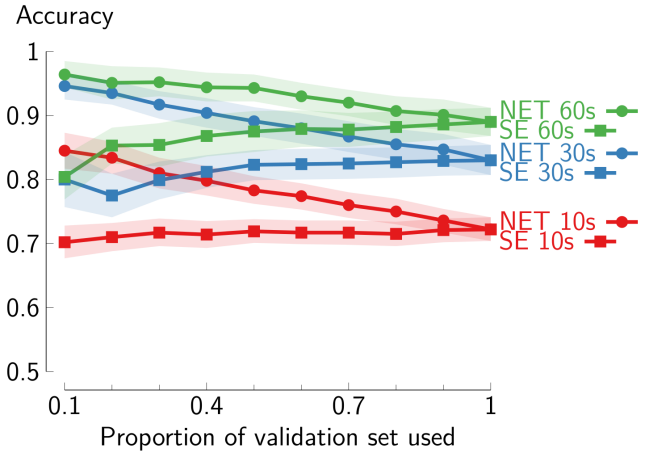

Fig. 5. Selective auditory attention decoding (sAAD) accuracies on the $x\%$, $x = 0, \ldots, 100$, highest active listening segments, as signalled by the highest neural envelope tracking (NET) correlation and lowest spectral entropy (SE) segments conform the findings on Dataset I and Dataset II. By selecting the highest NET correlation segments and only evaluating sAAD performance on these segments, the sAAD accuracy increases with a decreasing proportion of the validation set used, whereas the reverse is observed when selecting low SE segments. This trend is present for all the $10\,\mathrm{s}$, $30\,\mathrm{s}$ and $60\,\mathrm{s}$ window lengths.

Nevertheless, the SE does prove useful in discriminating between two distinct, predefined tasks (e.g., active listening versus watching a movie) whenever both tasks require a different amount of cognitive resources. However, the relative change of the SE between two conditions will depend on the nature of the task(s) in both conditions, and is not necessarily (directly) related to the state of active versus passive listening. This hypothesis is, furthermore, consistent with the predictive power of the SE in dedicated tasks, such as for anaesthetic depth [14], [15], respiration movements [15], sleep stages [15], and imagined finger movement [15], and consistent with the discriminative power in dedicated tasks, such as between subjects in rest and subjects performing mental arithmetic [16], and between subjects in rest and subjects fixated on flashing patterns [17]. The NET

correlations, on the contrary, do not suffer from this phenomenon of relative change between conditions since this technique is directly tied to the task of discriminating active versus passive listening by explicitly trying to reconstruct features of the presented speech stimulus, which is expected to become less accurate when the subject is not actively attending the speech. As a result, although the NET correlation accuracies seem lower than the SE classification in Fig. 4, the NET correlations seem better tailored to the aAAD and sAAD tasks.

In Fig. 4, the results on Dataset I attain higher accuracies than in Dataset II, possibly due to differences in distractor conditions, or data quality. Furthermore, although the experiment of Dataset I was conducted in a non-radio frequency shielded room, both features significantly discriminate between the active and passive listening conditions. This indicates that this technology is viable in everyday environments, outside controlled radio frequency shielded laboratories.

VII. Conclusion

In this paper, we have estimated the modulation of active listening to speech, a task referred to as absolute auditory attention detection (aAAD). Both neural envelope tracking (NET) and spectral entropy (SE) features were used to perform this aAAD task. To this end, we have introduced a new dataset containing an active listening condition, as well as distractor conditions during which the subject silently reads a text or solves arithmetic exercises. Next to this new dataset, we have also used an existing dataset where the distractor condition consisted of watching a silent movie. In both datasets and all distractor conditions, we have found that the NET and SE features are both able to discriminate between active listening versus distractor (i.e., passive listening) conditions. Whereas the NET increases from the active listening condition to the distractor conditions, the alpha and beta band SE in the frontal, parietal-occipital, and occipital channels decreases from the active listening condition to the distractor conditions. Only evaluating sAAD performance on segments of high NET shows an increased selective auditory attention decoding (sAAD) accuracy, whereas evaluating on segments of low SE shows the reverse trend. Thus, likely, the SE rather relates to the cognitive load required for each task than the actual absolute auditory attention, whereas the NET correlations directly relate to (in)attention to the stimulus and are consistent across several datasets and tasks. The NET, thus, seems the better option to estimate absolute auditory attention, and appears to be more suited to use in conjunction with an sAAD task in neuro-steered hearing devices.

Appendix

Segments are annotated as eye artefacts if the total power in the frontal channels (Fp1, AF7, AF3, Fpz, Fp2, AF8, AF4 and Afz for the 64-channel EEG cap, and Fp1, Fp2 and Fpz for the 24-channel EEG cap.) is higher than 5 times the baseline power in those channels, according to the procedure of [11].

Regarding the muscle artefact detection, the EEG signal is first filtered using a zero-phase Chesbyshev type 2 filter bandpass filter with cutoff frequencies between $20\,\mathrm{Hz}$ and $60\,\mathrm{Hz}$. Segments are thereafter annotated as muscle artefacts if the total power of this filtered EEG signal in the channels at the side of the head (AF7, F7, F5, FT7, FC5, T7, C5, TP7, CP5, P7, P5, P9, PO7, AF8, F6, F8, FC6, FT8, C6, T8, CP6, TP8, P6, P8,

PO8 and P10 for the 64-channel EEG cap, and F7, T7, CP5, F8, T8, TP9 and TP10 for the 24-channel EEG cap) is higher than 60 times the baseline power in those channels [30].

ACKNOWLEDGEMENTS

We thank Ir. Linsey Dewit-Vanhaelen and Ir. Elly Brouckmans for writing the protocol and performing the experimental recordings for the new Dataset I. We thank the authors of [11] (Dr. Jonas Vanthornhout, Dr. Lien Decruy and Prof. Tom Francart) for granting us access to Dataset II.

This research is funded by Aspirant Grant 1S31522N (for N. Heintz) from the Research Foundation - Flanders (FWO), a PDM mandate from KU Leuven (for S. Geirnaert, No. PDMT1/22/009), a junior postdoctoral fellowship fundamental research from the FWO (for S. Geirnaert, No. 1242524N), FWO project nr. G081722N, Internal Funds KU Leuven IDN project IDN/23/006, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895), and the Flemish Government (AI Research Program). The scientific responsibility is assumed by its authors.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

[1] 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, Jul. 2015.

[2] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, May 2017.

[3] 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, p. e56481, Apr. 2021.

[4] 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, May 2021.

[5] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, 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, Jul. 2021.

[6] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, Jul. 2012.

[7] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. 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, Mar. 2013.

[8] K. C. Puvvada and J. Z. Simon, "Cortical Representations of Speech in a Multitalker Auditory Scene," *Journal of Neuroscience*, vol. 37, no. 38, pp. 9189–9196, Sep. 2017.

[9] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE journal of biomedical and health informatics*, vol. 25, no. 10, pp. 3955–3966, Oct. 2021.

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

[11] J. Vanthornhout, L. Decruy, and T. Francart, "Effect of Task and Attention on Neural Tracking of Speech," *Frontiers in Neuroscience*, vol. 13, p. 977, 2019.

[12] Y.-Y. Kong, A. Mullangi, and N. Ding, "Differential modulation of auditory responses to attended and unattended speech in different listening conditions," *Hearing Research*, vol. 316, pp. 73–81, Oct. 2014.

[13] D. Lesenfants and T. Francart, "The interplay of top-down focal attention and the cortical tracking of speech," *Scientific Reports*, vol. 10, no. 1, p. 6922, Dec. 2020.

[14] H. Viertiö-Oja, V. Maja, M. Särkelä, P. Talja, N. Tenkanen, H. Tolvanen-Laakso, M. Paloheimo, A. Vakkuri, A. Yli-Hankala, and P. Meriläinen, "Description of the Entropy™ algorithm as applied in the Datex-Ohmeda S/5™ Entropy Module," *Acta Anaesthesiologica Scandinavica*, vol. 48, no. 2, pp. 154–161, 2004.

[15] I. Rezek and S. Roberts, "Stochastic complexity measures for physiological signal analysis," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 9, pp. 1186–1191, 1998.

[16] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano, "Quantification of EEG irregularity by use of the entropy of the power spectrum," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 3, pp. 204–210, Sep. 1991.

[17] D. Lesenfants, D. Habbal, C. Chatelle, A. Soddu, S. Laureys, and Q. Noirhomme, "Toward an Attention-Based Diagnostic Tool for Patients With Locked-in Syndrome," *Clinical EEG and Neuroscience*, vol. 49, no. 2, pp. 122–135, Mar. 2018.

[18] A. Belyavin and N. A. Wright, "Changes in electrical activity of the brain with vigilance," *Electroencephalography and Clinical Neurophysiology*, vol. 66, no. 2, pp. 137–144, Feb. 1987.

[19] J. C. F. de Winter, S. D. Gosling, and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," *Psychological Methods*, vol. 21, no. 3, pp. 273–290, Sep. 2016.

[20] E. R. Dougherty, *Random Processes for Image Signal Processing*. Bellingham: Wiley-IEEE Press, Nov. 1998.

[21] N. Das, T. Francart, and A. Bertrand, "Auditory attention detection dataset kuleuven (1.0.0) [data set]," https://zenodo.org/record/3377911, 2019.

[22] "Mobile EEG for Neuroscience Reseach - mbt | mBrainTrain," Jan. 2023. [Online]. Available: https://mbraintrain.com/

[23] J. Lindgren, "Converting .ov files to Matlab," Dec. 2015. [Online]. Available: http://openvibe.inria.fr/converting-ov-files-to-matlab/

[24] deBuren, "Radioboeken voor kinderen," 2007. [Online]. Available: https://soundcloud.com/deburen-eu/sets/radioboeken-voor-kinderen

[25] R. Dahl, *Alle verhalen*. Amsterdam: Meulenhof, Sep. 2013.

[26] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, Oct. 1995.

[27] P. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening, Modern Acoustics and Signal Processing*. Berlin: Springer, Jan. 2013, pp. 33–56.

[28] B. Somers, T. Francart, and A. Bertrand, "Github repository: MWF toolbox for EEG artifact removal," Jun. 2023. [Online]. Available: https://github.com/exporl/mwf-artifact-removal

[29] ——, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering*, vol. 15, no. 3, p. 036007, Jun. 2018.

[30] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *NeuroImage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[31] 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, Feb. 2004.

[32] ——, "Honey, I Shrunk the Sample Covariance matrix," 2014. [Online]. Available: http://ledoit.net/honey_abstract.htm

[33] B. Babadi and E. N. Brown, "A Review of Multitaper Spectral Analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1555–1564, May 2014.

[34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., ser. Springer Series in Statistics. New York: Springer, 2009.

[35] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[36] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[37] Student, "The Probable Error of a Mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.