Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario

Wouter Biesmans[†], Neetha Das[†⋆], Tom Francart[⋆], Alexander Bertrand[†]

***Abstract*—This paper considers the auditory attention detection (AAD) paradigm, where the goal is to determine which of two simultaneous speakers a person is attending to. The paradigm relies on recordings of the listener's brain activity, e.g., from electroencephalography (EEG). To perform AAD, decoded EEG signals are typically correlated with the temporal envelopes of the speech signals of the separate speakers. In this paper, we study how the inclusion of various degrees of auditory modelling in this speech envelope extraction process affects the AAD performance, where the best performance is found for an auditory-inspired linear filter bank followed by power law compression. These two modelling stages are computationally cheap, which is important for implementation in wearable devices, such as future neuro-steered auditory prostheses. We also introduce a more natural way to combine recordings (over trials and subjects) to train the decoder, which reduces the dependence of the algorithm on regularization parameters. Finally, we investigate the simultaneous design of the EEG decoder and the audio subband envelope recombination weights vector using either a norm-constrained least squares or a canonical correlation analysis, but conclude that this increases computational complexity without improving AAD performance.**

***Index Terms*—Neuro-steered auditory prostheses, cocktail party, auditory attention, EEG processing, speech envelope, auditory models.**

I. Introduction

The human auditory system has the remarkable ability to attend to one speaker and ignore the others in a so-called cocktail party scenario with multiple simultaneous speakers. Since the effect was first described in 1953 [2], it has been a topic of ongoing research in the fields of neuroscience and audiology. It was demonstrated in [3] that speech spectrograms that were reconstructed from cortical responses to a multi-speaker stimulus reveal spectral and temporal features of the attended speaker, as if the unattended speakers weren't there. This is very interesting from a neuroscientific point of view as it provides a new research tool that can help to understand and map the human auditory processing. Furthermore, it allows to detect to which speaker a subject is attending in a cocktail party scenario [4]–[6]. This auditory attention detection (AAD) paradigm might lead to a breakthrough for auditory prostheses (APs) such as hearing aids and cochlear implants. As it stands, current state-of-the-art APs employ beamforming, fixed or adaptive, to enhance a signal from one direction and suppress the rest. However, the system does not know to which signal the listener intends to attend. Therefore, the integration with an AAD system to steer the beamforming algorithm to the attended speaker would be of great benefit. AAD has successfully been applied to electrocorticography [4], magnetoencephalography (MEG) [6] and EEG [5], [7], [8]. When aiming for application of AAD in portable, mainstream devices such as a AP however, EEG is the only practical non-invasive modality. Although wearable EEG devices are currently still quite bulky, significant progress has been made towards unobtrusive wearable EEG solutions [9]–[14].

Different multi-channel approaches have been proven to be successful at performing AAD. In [8], robust features that are relevant for classification are extracted from the neural measurements, and then used to train a classifier. In [15], attention is tracked with a high temporal resolution using a state-space modelling approach. Another approach, which is currently more popular for AAD, relies on stimulus reconstruction in which a spatio-temporal linear decoder is first trained, and then used to reconstruct the envelope of the attended speaker's speech signal from the multi-channel neural measurements. The decoder can be trained using either a least-squares (LS) estimation error objective function [5], [7], or by maximizing a cross-correlation ratio using a generalized eigenvector decomposition [6]. Once such a decoder has been trained, it can be applied to other neural recordings, after which the reconstructed speech envelope can be compared to the actual speech envelopes through Pearson's correlation coefficient. A final classification then marks the speaker corresponding to the envelope with highest correlation as the attended speaker.

In this paper, we follow the LS approach proposed in [5] which is a popular method for AAD, mainly because of its simplicity and computational efficiency, while at the same time being very effective, as shown in several studies [7], [16]–[18]. However, we introduce a more natural way of combining the data for training the decoder, in which we solve a single LS problem over the entire data set, rather than averaging over a multitude of per-trial LS solutions. This different training methodology not only results in better AAD performance, but also reduces the sensitivity with respect to a regularization parameter, up to a point where the latter can be fully eliminated if sufficient training data is available.

A second goal of this paper is to investigate whether it is possible to improve AAD performance by including knowledge of the auditory periphery into the speech envelope extraction step. When a sound arrives at the ear, it is first filtered by the middle ear, followed by complex non-linear processing in the inner ear where the sound wave is converted into a series of spikes in the auditory nerve. Thereafter this

The work of W. Biesmans was supported by a Doctoral Fellowship of the Research Foundation - Flanders (FWO). This research work was carried out at the ESAT and ExpORL Laboratories of KU Leuven, in the frame of KU Leuven Special Research Fund BOF/STG-14-005 and OT/14/119. The scientific responsibility is assumed by its authors. A conference precursor of this manuscript has been published in [1].

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

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

spike train is processed by the brainstem, midbrain, auditory cortex and higher cortical areas. The reconstructed envelope used in the current study is probably derived from neural activity originating from the auditory cortex. While we currently only have limited knowledge of cortical processing of speech, there are good models available of the processing that takes place in the auditory periphery (outer and inner ear). As the spike trains in the auditory nerve serve as the input to the auditory cortex, it makes sense to include a model of the auditory periphery in the AAD processing chain.

Bearing in mind that the computational power in auditory prostheses is limited, we aim to optimize the computational complexity versus AAD performance trade-off, by including speech envelope extraction methods that model the auditory periphery with gradually increasing precision. We start from the standard envelope extraction method used in [5], and gradually include pragmatic and computationally cheap auditory-inspired signal operations, such as amplitude compression and subband processing with an auditory filter bank, and assess their individual effect on the AAD performance. Finally, envelope extraction methods based on three well-established and computationally complex auditory models are examined.

Subband envelope methods in general pose an extra question: how should we recombine these subband envelopes into one envelope, i.e., which weights should be given to each subband? As there is no obvious way to do this, one might benefit from an algorithm that determines the optimal recombination. Canonical correlation analysis (CCA) and bimodal LS provide a framework for obtaining optimal envelope weight vectors. We apply these algorithms to the best-performing subband envelope extraction method to see if they result in any performance increase.

The different methods are assessed using experimental data with EEG recordings from 16 subjects. We note that these are all new subjects, and this excludes the 7 subjects from our pilot study which was published as a conference precursor in [1], in which a different measurement protocol was used.

The paper is organised as follows: in Section II we start by reviewing the basic AAD procedure used in this paper, introducing the training methodology and detailing the evaluation strategy. In Section III we then discuss the different auditory-inspired speech envelope extraction methods. As subband envelopes provide us with an additional challenge, we describe two extended AAD procedures in Section IV that also take care of an optimal recombination of several subband envelopes into one envelope. In Section V, we provide details of the experiment design and the key processing parameters. In Section VI, we evaluate the effect of the different envelope extraction methods on AAD performance. In Section VII, we discuss the implications of these results for future application in APs, and we elaborate on remaining open problems. Finally, we draw conclusions in Section VIII.

II. Auditory Attention Detection

In this section, we review the basic AAD procedure, covering the training and detection process in more detail.

A. Problem statement

For the remainder of this paper we assume that for each test subject we have a set of K measurements, referred to as trials, available. Each trial consists of a C-channel EEG recording and the corresponding attended and unattended speech signals, which were simultaneously presented to the subject during the recording of the trial. Every trial is assumed to have the same length, which we will define later. We use $M(t, c)$ to denote the C-channel EEG recording, where t is the discrete time or sample index and c the channel index. The temporal envelopes are obtained by extracting the envelopes from both speech signals (attended and unattended) and are denoted by $s_a(t)$ and $s_u(t)$ respectively. We use the index k to indicate recordings from the k-th trial when appropriate, and a tilde to refer to an estimated variable rather than the real one (e.g. $\tilde{s}_a(t)$).

AAD can be achieved through an envelope reconstruction approach: a decoder is designed which reconstructs the attended speech envelope from the multi-channel EEG recordings. It has been shown that a linear, spatio-temporal decoder is capable of adequately reconstructing the attended speech envelope such that the reconstructed envelope resembles the attended speech envelope $s_a(t)$ more than the unattended speech envelope $s_u(t)$ [4]–[6]. This resemblance is quantified by Pearson's correlation coefficient and is used for the final AAD: the speech envelope that correlates best with the reconstructed envelope is ultimately classified as the attended speech envelope.

The stimulus reconstruction defining the decoder $D \in \mathbb{R}^{N_l \times C}$ can be expressed as follows:

$$\tilde{s}_a(t) = \sum_{n=0}^{N_l-1} \sum_{c=1}^{C} D(n, c) M(t + n, c). \tag{1}$$

Here n denotes the time lag index, with time lags ranging from 0 to $N_l - 1$ samples. The spatio-temporal nature of the decoder is expressed through the channel index c and the time lag index n, and allows the attended envelope at sample time t to be reconstructed as a weighted sum of all of the C EEG channels at time t, as well as future sample times $t + n$. The time lags account for the physical delay between the presentation of the auditory stimulus and the moment it is actually processed by the brain. It has been found in [5], that time lags up to 250 ms are most effective at reconstructing the envelope, which was also verified in our data.

B. Design of the decoder

The decoder D can be determined through optimization of a well-chosen objective function, for example by minimizing the expected value ($E[.]$) of the squared error between the estimated and the actual attended speech envelope as in [5]. Another sensible approach would be to maximize the Pearson correlation coefficient between both. Up to an irrelevant scalar, as we show below, both are in fact equivalent, i.e.:

$$\tilde{D} = \arg\min_D \; E[|\tilde{s}_a(t) - s_a(t)|^2], \tag{2}$$

$$\sim \arg\max_D \; \frac{E[\tilde{s}_a(t) s_a(t)]}{\sqrt{E[\tilde{s}_a^2(t)] E[s_a^2(t)]}}. \tag{3}$$

For ease of notation we define vectors $\mathbf{m}_c(t) \in \mathbb{R}^{N_l}$, containing all N_l time lags of channel c, and $\mathbf{m}(t) \in \mathbb{R}^{N_l C}$, containing all time lags of each of the C channels:

$$\mathbf{m}_c(t) = [M(t,c)\ M(t+1,c) \cdots M(t+N_l-1,c)]^T \quad (4)$$

$$\mathbf{m}(t) = [\mathbf{m}_1(t)^T\ \mathbf{m}_2(t)^T \cdots \mathbf{m}_C(t)^T]^T. \quad (5)$$

Equation (1) can then be rewritten as

$$\tilde{s}_a(t) = \mathbf{d}^T \mathbf{m}(t), \quad (6)$$

where $\mathbf{d} \in \mathbb{R}^{N_l C}$ is the vectorized version of D. This new notation is used in the remainder of this text.

Substituting equation (6) into equation (2) results in a standard linear minimum mean squared error (LMMSE) problem. Its analytical solution is well-known and can be obtained by setting the derivative with respect to the entries of $\mathbf{d}$ equal to zero, resulting in:

$$\tilde{\mathbf{d}} = R^{-1}\mathbf{r}_{\mathbf{m}s}, \quad (7)$$

where $R = E[\mathbf{m}(t)\mathbf{m}(t)^T] \in \mathbb{R}^{N_l C \times N_l C}$ is the autocorrelation matrix of the EEG recordings, and $\mathbf{r}_{\mathbf{m}s} = E[\mathbf{m}(t)s_a(t)] \in \mathbb{R}^{N_l C}$ is a vector containing the cross-correlations of the attended speech envelope and the (time-lagged) EEG channels.

To show the equivalence between (2)-(3), we reformulate (3) as a maximization of the numerator while replacing the denominator by a norm-constraint. With the notation introduced above, this leads to the equivalent problem:

$$\tilde{\mathbf{d}} \sim \arg\max_{\mathbf{d}}\ \mathbf{r}_{\mathbf{m}s}^T\mathbf{d} \qquad \text{s.t.} \quad \mathbf{d}^T R\mathbf{d} = 1, \quad (8)$$

where the factor $\sqrt{E[s_a^2(t)]}$ in (3) is omitted as it is independent of $\mathbf{d}$. This reformulation can then be solved using Lagrange multipliers, resulting in a scaled version of (7).

C. Training

In practice, the true autocorrelation matrix R and cross-correlation vector $\mathbf{r}_{\mathbf{m}s}$ are of course unknown, but can be estimated from the measurements through their sample estimates, denoted as $\tilde{R}$ and $\tilde{\mathbf{r}}_{\mathbf{m}s}$. This effectively transforms the LMMSE problem from (2) into a LS problem.

As the dimension of the decoder is large and conditions between trials might vary slightly, over-fitting to the specific data used for training is a real concern. To overcome this, a cross validation approach should be taken where decoders are only applied to EEG recordings that were not used to construct the decoder. In this paper, we use a subject-specific leave-one-out cross validation. This means that the data from all $K-1$ other trials from the same subject are used in the design of the decoder to decode trial k. To emphasize this, we use the subscript $-k$ to denote the decoder $\tilde{\mathbf{d}}_{-k}$ for the k-th trial.

To combine data from multiple trials in the design of the decoder, it is common practice to construct a preliminary set of decoders $\tilde{\mathbf{d}}_k = \tilde{R}_k^{-1}\tilde{\mathbf{r}}_{\mathbf{m}s,k}$ for $k = 1 \ldots K$, using only the data from a single trial [5], [7], [16], [17]. Then preliminary decoders from all trials but trial k can be averaged to obtain a decoder $\tilde{\mathbf{d}}_{-k}$ that is used to decode trial k:

$$\tilde{\mathbf{d}}_{-k} = \frac{1}{K-1}\sum_{\substack{i=1 \\ i\neq k}}^{K} \tilde{\mathbf{d}}_i, \quad k = 1 \ldots K, \quad (9)$$

where i is used as another trial index. If a subject-independent decoder is to be trained, further averaging can be done across subjects to find a so-called 'grand-average' decoder [5], [7].

Although, mathematically speaking, this is a rather arbitrary way of combining the data from multiple trials or subjects, this method has been proven to be successful in several papers [5], [7], [16], [17]. However, typically single trials are rather short (e.g. 60 seconds) compared to the large dimension of $\tilde{R}_k$. This contributes to $\tilde{R}_k$ being ill-conditioned or even (in extreme cases) rank-deficient, which is a problem when evaluating equation (7). The obtained decoders are very sensitive to perturbations on the training data, such that all trials may generate very different solutions. In this case, a simple averaging of the different decoders may prove ineffective.

To avoid rank-deficiency and improve conditioning, regularization is then typically applied to R_k:

$$\tilde{\mathbf{d}}_k = (\tilde{R}_k + \lambda z_k Q)^{-1}\tilde{\mathbf{r}}_{\mathbf{m}s,k}, \quad (10)$$

where λ is a relative regularization parameter, which is multiplied by z_k, the mean eigenvalue of $\tilde{R}_k$. This mean eigenvalue z_k can easily be calculated as the average of the diagonal elements of $\tilde{R}_k$. Q is the regularization matrix, typically chosen to be the identity matrix (corresponding to a ridge regression), penalizing the L2 norm of $\mathbf{d}$. In the case of AAD, Q is sometimes also chosen to be:

$$Q = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}. \quad (11)$$

This choice for Q penalizes the norm of the discrete derivative of $\mathbf{d}$ with respect to its entry index. This can be preferred if the spatio-temporal decoder is expected to be smooth in the temporal dimension.

Using such regularization schemes, AAD performance can be boosted to an acceptable level. However, disadvantages of regularization are that it adds bias, and that it requires a regularization parameter λ that needs to be tuned in order to find a good balance between sufficient generalization (to avoid over-fitting) and too much generalization (to avoid loosing predictive power). In the Appendix, we show that the performance is rather sensitive to the choice of this regularization parameter, when the decoder is indeed computed as an average of per-trial decoders.

In this paper, we use a more natural way to combine the recordings, by minimizing the following sum of LS objective functions:

$$\tilde{\mathbf{d}}_{-k} = \arg\min_{\mathbf{d}_{-k}} \frac{1}{K-1}\sum_{\substack{i=1 \\ i\neq k}}^{K} E[|\tilde{s}_{a,i}(t) - s_{a,i}(t)|^2], \quad (12)$$

$$\tilde{s}_{a,i}(t) = \mathbf{d}_{-k}^T\mathbf{m}_i(t). \quad (13)$$

This objective function can be maximized by plugging the average autocorrelation and cross-correlation vector from all but the k-th trial (respectively denoted by $\tilde{R}_{-k}$ and $\tilde{\mathbf{r}}_{\mathbf{m}s,-k}$)

into equation (7). Note that averaging sample autocorrelation and covariance matrices is equivalent to concatenating the recordings in time and training a single decoder based on this longer, concatenated recording. Again, if a subject-independent decoder is to be designed, the summation in equation (12) can be extended to sum over all subjects. However, in the sequel, we only consider subject-dependent decoders.

As more samples are available for the estimation of $\tilde{R}_{-k}$ compared to $\tilde{R}_k$, it is naturally better conditioned, in our case[1] removing the need for a regularization scheme. In addition, we show in the Appendix that, even if the optimal value of the regularization parameter is selected in (9)-(10) through a parameter sweep, the second option still yields significantly better results. Therefore, in the remainder of this paper, we always train a single LS decoder on the full training data set as in (12) rather than averaging single-trial decoders as in (9)-(10).

D. Detection details

Once the decoder $\tilde{\mathbf{d}}_{-k}$ has been trained based on the data from all trials except trial k, it can be used to reconstruct the attended speech envelope from the EEG recording of trial k, as in equation (1). Pearson correlation coefficients can then be calculated between this reconstructed speech envelope $\tilde{s}_a(t)$ and both real speech envelopes $s_a(t)$ and $s_u(t)$. We refer to these coefficients as the reconstruction accuracies, and denote them by r_a and r_u respectively. The speech envelope corresponding to the highest reconstruction accuracy is then naturally classified as the attended speech, e.g. if $r_a > r_u$, this results in a correct classification. This process is repeated for each trial. Note that the discriminative power of these correlation coefficients strongly depends on the trial length, which can be chosen post hoc (after the experiment).

III. Envelope Extraction Methods

The main focus of this paper is to evaluate whether it is possible to improve the AAD performance, by gradually including more knowledge of the auditory periphery in the envelope extraction process. In this section we describe the different methods for extracting such a speech envelope $s(t)$ from the speech signal $x(t)$. The AAD performance using these envelope extraction methods is evaluated in section VI. We start by describing some basic envelope extraction methods lacking any auditory motivation, and gradually increase the complexity of auditory modelling. Finally, some methods based on more accurate (but more complex) models of the auditory periphery are discussed. Concluding this section, we briefly motivate the choice of some filtering parameters relevant to the envelope extraction.

[1]We note that the necessity of regularization depends on two factors: the number of (independent) samples available to the LS problem (amount of training data), and the number of elements of the decoder (# unknowns). Thus, to avoid having to use regularization, our approach is to keep the sample rate and the number of time lags N_l as low as possible, while providing a maximal number of samples for training.

A. Basic envelope extraction

Basic envelope extraction methods are based on what we intuitively think of when considering envelopes of signals. The first method calculates the speech envelope by taking the absolute value $|x(t)|$ of the broadband signal $x(t)$ and low pass filtering the result. The process is known as full-wave rectification and is often used in electronics. We abbreviate this method as 'abs'.

As an alternative, one can compute the amplitude of the complex-valued analytic signal, which is often referred to as the mathematical envelope of a signal, as it results in the modulating signal when applied to a modulated sine wave. The analytic signal of a signal $x(t)$ can be constructed as $x(t) + jH(x(t))$, where $H(.)$ represents the Hilbert transform operator, which applies a 90 degrees phase shift to the original signal. We mention this method for completeness, but when applied to audio signals it results in nearly identical envelopes as the 'abs' method (after subsequent band pass filtering, see subsection III-E). As it is also computationally more complex, we omit this method in favour of the first.

In a second method we consider, the speech envelope is calculated as the long-term power average of the signal. As the long-term average can be obtained by integration, or equivalently, low pass filtering, we obtain the envelope by squaring the original signal and low pass filtering it afterwards. We refer to this method as 'square'.

B. Compressed envelopes

The human auditory system is not a linear system. For example the relation between intensity of the stimulus and the perceived loudness is less-than-linear. This compression results in a relative attenuation of higher amplitude signals, making it possible for the human ear to have a large dynamic range. The relationship between loudness, which is a perceptual measure of stimulus intensity, and the actual stimulus intensity has been studied extensively. Typical simple models used for this relation use either a power law relation, i.e. $|x(t)|^\beta$ with exponent[2] $\beta = 0.6$ [19], or a logarithmic relation [20], i.e. $\log(|x(t)| + \epsilon)$. Here ϵ is a small, positive number ensuring that the argument of the logarithm is strictly positive. We refer to these methods as 'p-law' and 'log' respectively.

Remark: the power law function is scale-invariant for positive scaling factors a, i.e. $|ax(t)|^\beta = a^\beta |x(t)|^\beta, \forall a \geq 0$. This means that normalization of the signal has no influence on the shape of the resulting envelope, which is desirable. The logarithmic function has a similar property: scaling of the argument only results in a DC offset in the envelope: $\log(|ax(t)|) = \log(|x(t)|) + \log(a), \forall a > 0$. As all envelopes are high pass filtered at a later stage, we can ignore this DC offset. Hence, for our application, the logarithmic function can also be considered scale-invariant.

C. Subband envelopes

In the auditory pathway, speech signals are split into frequency subbands by the basilar membrane in the cochlea

[2]Note that 'abs' and 'square' are effectively also power law methods with exponents chosen as respectively 1 and 2.

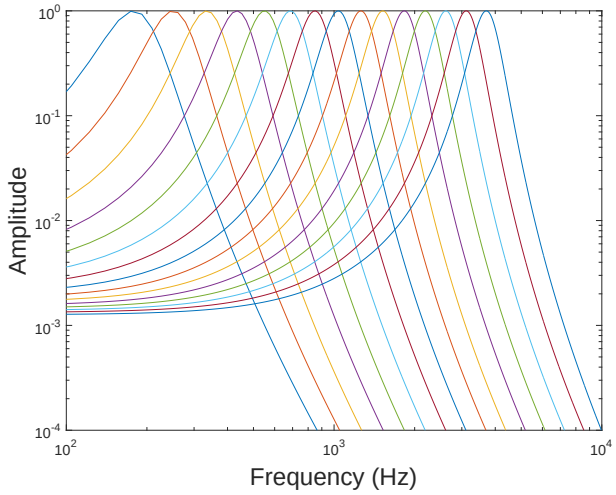

Fig. 1: Frequency response of the gammatone filterbank used to decompose the audio into frequency subbands.

before the actual envelope extraction process takes place. We mimic these auditory filters, e.g. by applying a gammatone filter bank [21], [22] to the audio signal $x(t)$. The filter bank contains $N_s = 15$ perceptually uniform gammatone filters, each with an equivalent rectangular bandwidth equal to 1.5, and center frequencies ranging from 150 Hz to 4 kHz (as the stimulus is also band-limited to 4 kHz, see Section V). Its frequency response is depicted in figure 1. Note that the filters become wider at higher frequencies, which reflects the fact that the human auditory system has a poorer spectral resolution in the higher frequencies. It was found that the actual number of subbands N_s is not critical for performance, as long as the filter widths are scaled accordingly and $N_s \geq 5$.

The aforementioned envelope extraction methods can then be applied to each of the N_s subband signals instead of the broadband signal, resulting in a vector $\mathbf{s}(t) \in \mathbb{R}^{N_s}$ of subband envelopes. Since the basic AAD procedure as described in section II requires just one envelope however, some recombination of these subband envelopes is necessary. We only consider weighted linear combinations: $s(t) = \mathbf{w}^T\mathbf{s}(t)$, where $\mathbf{w} \in \mathbb{R}^{N_s}$ is the envelope weight vector. Subband envelopes could, for example, be recombined with equal weights, or with weights determined by the respective band importance (BI) [23]. The latter is used in the calculation of the speech intelligibility index, and we employ the same here to weight the frequency subbands to increase the influence of the frequency bands that contribute most to speech intelligibility. We will compare this with a uniform (all-ones) weight vector $\mathbf{w}$, and additionally, in section IV, we will also define two methods to choose a set of envelope weights that are optimal in some mathematical sense.

Applying the gammatone filterbank before, and using the previously described envelope extraction methods to each of the subband signals, introduces a subband version of each. We mark these respective subband versions by the postfix 'sub', e.g. 'abs sub'. If the postfix 'sub' is used in the sequel, we refer to the case where the subbands are added with equal weights unless stated otherwise.

D. Auditory models

For even more detailed models of the auditory periphery, we refer to three well-known auditory models [24]–[26].

The first model, 'Yang' [24], is available through the NSL Matlab Toolbox, and processes the audio in three stages. The analysis stage models the cochlear filters by applying a wavelet transform, to decompose the signal into 128 subbands. The transduction stage is applied to each subband individually and models the dynamics of the hair cells. Hair cells are the auditory system's transducers, transforming the mechanical vibrations into electrical activity. Finally, the reduction stage reduces the amount of information by extracting only certain spectral features, using a lateral inhibitory network. As the output consists of 128 subband 'envelopes', we recombine these, by adding them using uniform weights.

The second model, 'Meddis' [26], is implemented in the MATLAB Auditory Periphery (MAP) toolbox. It is more complex than 'Yang', both from a modelling perspective as from a computational point of view. As a result, envelope extraction cannot be performed in real-time. Therefore the model, in its current form, might not be directly amenable for wearable devices. Still it is interesting to see how it performs. It models the auditory periphery through 9 modules, each tied to a specific physiological process and its outputs each represent physically measurable values such as the instantaneous neuron firing rate at user-defined frequencies.

The third and final model, 'Zilany' [25], models the auditory periphery using a phenomenological approach. The main stages comprise a forward control path, outer and inner hair cell sections and a synapse model. In its current implementation however, it is even more computationally demanding than 'Meddis'.

Both 'Meddis' and 'Zilany' provide instantaneous firing rates of the auditory nerves at neurons corresponding to user-defined frequencies as a possible output. We use these firing rates at the same 15 center frequencies as were previously used for the gammatone filter bank as subband 'envelopes'. As the output is present in the form of neuronal firing rates, they are weighted by their respective neuronal densities [27], [28] before being added together to form a single envelope.

E. Filtering

It has been shown that speech envelope and EEG recordings correlate best within the δ and θ band frequencies [4]. We therefore digitally band pass filter both the speech envelopes and the EEG recordings between 2 and 9 Hz. Because of this, whenever any low pass filtering or integration occurs as a last step in the envelope extraction process, it is skipped, as it is redundant. All filtering is performed using a linear phase filter, where the same filter is applied to the speech envelopes and the EEG recordings.

IV. Bimodal AAD Procedure

With the notion of subband envelopes that was introduced previously, a new question arises: can AAD performance be further improved by recombining subband envelopes through

a mathematically optimal weight vector instead of the arbitrary or physiologically motivated options that we proposed before? In this section, we provide two mathematically optimal methods for simultaneously obtaining both an EEG decoder $\mathbf{d}$ and an envelope weight vector $\mathbf{w}$. In this case data from two modalities (envelope and EEG domain) are used simultaneously, hence the term bimodal.

Optimality should of course first be defined by some objective function. Two suitable objective functions, similar to those from the basic AAD processing, are considered. As in equation (3), we can choose to maximize the Pearson correlation between the estimated attended speech envelope $\tilde{s}_a(t)$ and the true attended speech envelope $s_a(t)$, which now is a recombination of multiple attended speech subband envelopes contained in $\mathbf{s}_a(t)$. The second option is to minimize the LS estimation error between the two, similar to equation (2). The difference now is that in both objective functions the audio weights vector $\mathbf{w}$ is included as an additional optimization variable. The objective function (3) then becomes:

$$\tilde{\mathbf{d}}, \tilde{\mathbf{w}} = \arg\max_{\mathbf{d},\mathbf{w}} \frac{E[(\mathbf{d}^T\mathbf{m}(t))(\mathbf{w}^T\mathbf{s}_a(t))]}{\sqrt{E[(\mathbf{d}^T\mathbf{m}(t))^2]E[(\mathbf{w}^T\mathbf{s}_a(t))^2]}}. \quad (14)$$

This objective function also appears in canonical correlation analysis (CCA) [29], and its solution corresponds to the first canonical weight vectors. The solution can also be derived using Lagrange multipliers after reformulating the denominator of (14) as two norm constraints. The optimal decoder and envelope weight vector are then found as:

$$\tilde{\mathbf{d}} = \text{GEVec}_1(R_{\mathbf{ms}}R_{\mathbf{s}}^{-1}R_{\mathbf{ms}}^T, R) \quad (15)$$

$$\tilde{\mathbf{w}} = \text{GEVec}_1(R_{\mathbf{ms}}^T R^{-1} R_{\mathbf{ms}}, R_{\mathbf{s}}), \quad (16)$$

where $\text{GEVec}_1(A, B)$ is used to denote the principal generalized eigenvector of the matrix pencil (A, B), $R_{\mathbf{s}} = E[\mathbf{s}_a(t)\mathbf{s}_a^T(t)] \in \mathbb{R}^{N_s \times N_s}$ is the subband speech envelopes' covariance matrix, and $R_{\mathbf{ms}} = E[\mathbf{m}(t)\mathbf{s}_a^T(t)] \in \mathbb{R}^{C \times N_s}$ is a matrix containing the cross-correlations between each (time-lagged) EEG channel and each subband envelope.

Based on experiments, it was found that choosing $R_{\mathbf{s}} = I_{N_s}$, which corresponds to infinite regularization, yields as good results as any other choice of the regularization parameter. However, it avoids having to estimate $R_{\mathbf{s}}$, which is especially convenient in a practical setting. It is noted that CCA with one of the covariance matrices set to the identity matrix is equivalent to Orthonormalized Partial LS (OPLS), a variant of PLS [30].

The second option, which we refer to as 'bimodal LS', corresponds to the cost function (2) which is now altered to:

$$\tilde{\mathbf{d}}, \tilde{\mathbf{w}} = \arg\min_{\mathbf{d},\mathbf{w}} E[|\mathbf{d}^T\mathbf{m}(t) - \mathbf{w}^T\mathbf{s}_a(t)|^2], \quad (17)$$

$$\text{s.t.} \quad \| \begin{bmatrix} \mathbf{d} \\ \mathbf{w} \end{bmatrix} \|_2^2 = 1.$$

The norm constraint on $\begin{bmatrix} \mathbf{d} \\ \mathbf{w} \end{bmatrix}$ is necessary to avoid the trivial solution. After writing the two-norm in full, it can be seen that the solution can be found as:

$$\begin{bmatrix} \mathbf{d} \\ \mathbf{w} \end{bmatrix} = \text{EVec}_{\text{min}}(T) \quad (18)$$

$$T = \begin{bmatrix} R & -R_{\mathbf{ms}} \\ -R_{\mathbf{ms}}^T & R_{\mathbf{s}} \end{bmatrix}, \quad (19)$$

where $\text{EVec}_{\text{min}}(A)$ denotes the eigenvector of the matrix A corresponding to the smallest eigenvalue. Unlike in (2)-(3), both objective functions now provide us with different solutions. However, it can be shown that if the smallest eigenvalue of T is small compared to the eigenvalues of both R and $R_{\mathbf{s}}$, both solutions are approximately equivalent (proof omitted).

To limit our search space for the best performing envelope extraction method, we only apply 'CCA' and 'bimodal LS' to the best performing subband envelope procedure from section III, to see if any further performance improvement can be obtained.

V. Experimental Procedures

To be able to evaluate the performance of the different envelope extraction methods proposed above, we set up an AAD experiment, which we describe in this section.

A. Set-up of the experiment

1) Goal: The experiment was designed to mimic a cocktail party scenario in which the subject listens to two simultaneous speakers at two distinct spatial locations, and attempts to attend to only one of them while ignoring the other.

2) Subjects: 16 normal hearing subjects (verified by audiometry) between 17 and 30 years old participated in the experiment, 8 of them were male, 8 were female. All of them (and/or their legal guardian) signed an informed consent form approved by the KU Leuven ethical committee.

3) Equipment: During the entire experiment, 64-channel EEG was recorded using a BioSemi ActiveTwo system. The electrodes were placed on the head according to international 10-20 standards. The experiment took place in a soundproof, electromagnetically shielded room, and auditory stimuli were presented to the subjects using insert phones (Etymotic ER3A) at 60dBA. As the insert phones' transducer has a cut-off frequency of 4 kHz, all audio signals were low pass filtered at 4 kHz as well.

4) Stimulus structure: As audio material, four Dutch short stories [31], narrated by different speakers (all male), were selected. Silences were truncated to 500 ms, and the resulting audio was divided into two 'tracks', one of which was to be attended by the subject while the other was to be ignored. Each track consisted of four story 'parts', lasting approximately six minutes each. After presenting one part to the subject (a 'presentation'), some multiple choice questions about the content of the attended story part appeared on a screen. These questions were intended to keep the subject engaged in the task and the answers were not used further in this study. After four story parts, the subject was offered an extended break.

5) Presentation structure: After the break, the same stimuli were presented, but the subject was asked to attend to the other track. After a second break, the subject was then asked to attend to each part of the first track again. This time however, only the first two minutes of each part was presented, without questions in between. This was repeated two more times, such that the first two minutes of each part of the first track was attended four times in total. These so-called 'repetitions' were kept at a minimum as they are perceived as boring and might result in attention loss, and were included for a specific purpose in a related study. However, it is important to note that we do not exploit these repetitions to, e.g., improve signal-to-noise ratio by averaging them. The EEG analysis in this paper is performed on a single-trial basis, although a subset of the stimuli appears multiple times in the data set. Summarizing, the subject first attended eight unique story parts of six minutes each, before listening to the first two minutes of each part of the first track for three more times, totalling twelve repetitions. This brings the grand total at 8 x 6 minutes + 12 x 2 minutes = 72 minutes of recorded EEG per subject.

6) Presentation mode: In order to design a more general decoder that works in different conditions, two conditions were varied evenly in between every presentation: the ear to which the attended track was presented, and the acoustic processing of the speech signals. Either 'dry' speech was offered, i.e. each speaker was presented to a different ear, or speech signals were processed by (dead room) head-related transfer functions, simulating a more realistic listening scenario in which the speakers are spatially located 90 degrees to the left and the right of the subject. In this case the stimuli of both ears contain both speech signals, albeit with different intensities and delays. The order of presentation of both condition types was balanced over the different subjects.

B. Data Processing

The recorded EEG is band pass filtered between 2 and 9 Hz, and down-sampled to 20 Hz. The auditory stimuli that were presented to the subjects are sampled at 8 kHz (as their frequency content only ranges up to 4 kHz). For both the attended and the unattended speech signals, envelopes are extracted using the different methods detailed in section III. Afterwards these envelopes are band pass filtered between 2 and 9 Hz, and downsampled to 20 Hz as well.

Trials are then created by chopping the data into pieces of equal length. Most studies on AAD employ 60 second trials. For the main analysis of this paper however, i.e. the comparison of the different envelope extraction methods, we deliberately choose a shorter trial duration of 30 seconds, because it makes the differences between methods become more apparent. Furthermore, it also means that more trials (144 instead of 72) can be evaluated, resulting in more power for the statistical tests. Nonetheless, to allow for a comparison with literature, we also provide some results with 60 second trials.

For each of the 16 subjects, a decoder D was then trained for each trial, using the data of every other trial of this subject as described in subsection II-C. Note that it can be expected that all decoders for a subject are very similar, because of the leave-one-trial-out approach. The decoders are applied to the EEG recordings, resulting in 144 reconstructed speech envelopes per subject (one for each trial). Pearson correlations with both attended and unattended speech envelopes are calculated for each trial, and compared (see subsection II-D for details). This results in either a correct (1) or wrong (0) detection for each trial. Thus, for each tested envelope extraction method, a binary detection result vector $\mathbf{q}$ of length 2304 (16 subjects * 144 trials per subject = 2304 total trials) is obtained.

These can either be used as an input for a statistical test (see subsection V-C), or averaged to obtain the method's so-called 'detection accuracy', which is used as the main performance parameter.

C. Permutation test

To evaluate whether a method a performed significantly better than method b, permutation tests were used [32]. The test statistic S was calculated as the sum of the differences between the binary result vectors of both methods: $\mathbf{q}_{(a-b)} = \mathbf{q}_a - \mathbf{q}_b$, $S = \sum_j \mathbf{q}_{(a-b)}(j)$, where j is the vector entry index. Note that a large positive value of the test statistic S implies that method a is more accurate than method b, whereas a large negative value implies the opposite. A value of S that is close to zero implies that both methods perform similarly. To test for statistical significance, the test statistic S was then compared to its estimated cumulative density function (CDF) under the null hypothesis of no difference between the methods. This CDF was estimated by repeatedly ($n = 100\ 000$) randomly permuting each of the 16 subjects' results among the two methods, and re-evaluating the test statistic.

We hypothesize, based on the preliminary results in our conference precursor[3] [1], that 'p-law sub' is the best-performing method and therefore compare its performance pair-wise with every other method. To account for the multiple comparisons, an adjustment to the individual rejection criteria is made using the Holm-Bonferroni method [33].

VI. Results

In the sequel, we assume a uniform subband weighting when referring to 'subband' methods, unless explicitly stated otherwise (a comparison with other weighting methods is reported further on). Figure 2 shows the subject-specific (circles) and mean (bars) AAD accuracy for the different envelope extraction methods. The exact (experiment-wide) values, along with the p-values resulting from a comparison with the 'p-law sub' method as discussed in subsection V-C are also shown in table I. An asterisk '*' in the last column indicates statistical significance at $\alpha = 0.05$ significance level. Even after a Holm-Bonferroni correction, 'p-law sub' performs significantly better than all other methods, although 'log sub' and 'abs sub', two other simple subband methods, come close in performance. Another interesting result is that in our current paradigm, the complex models ('Yang','Meddis', and 'Zilany') yielded no improvement over the more basic envelope extraction methods.

[3]It should be noted that [1] was based on a different set of measurements, independent of the measurements used in this manuscript

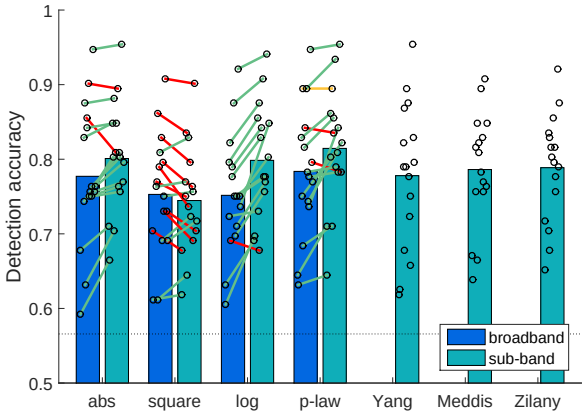

Fig. 2: Mean (bars) and individual subject (circles) detection accuracies for each of the different envelope extraction methods for a trial length of 30s. The dotted black line at 57% indicates the subject-specific detection accuracy which is only 5% likely to be surpassed by chance, based on a binomial distribution (success rate = 0.5, number of trials = 144).

Method	Detection accuracy (%)	p-value	Holm p-value
abs	77.7	<0.001	<0.001*
abs sub	80.1	0.008	0.023*
square	75.3	<0.001	<0.001*
square sub	74.5	<0.001	<0.001*
log	75.2	<0.001	<0.001*
log sub	79.9	0.034	0.034*
p-law	78.4	<0.001	0.002*
p-law sub	**81.5**	//	//
Yang	77.8	0.006	0.023*
Meddis	78.6	0.001	0.004*
Zilany	78.9	0.011	0.023*

TABLE I: Results for the different envelope extraction methods

As we noted before, 'abs', 'p-law' and 'square' are all power law variants with different values of the exponent β. Instead of limiting the analysis to just these a priori chosen instances, we also varied β between 0.1 and 2 for both a broadband and subband approach. Figure 3 depicts the result, showing the evolution of the average AAD accuracy as a function of the power law exponent β. From this figure it can be seen that indeed a choice for β of lower than 1 seems appropriate. The optimum in the figure is rather broad and achieved for the subband approach with values for β between 0.2 and 0.8. Another important observation here is that for most values of β, except for sub-optimally high instances (like β=2 in the square law), the subband method clearly outperforms the broadband method. This was also the case when using a logarithmic compression (see table I). It can therefore be concluded that dividing the speech signal in its subband components before extracting envelopes is an important factor for improving AAD performance.

The bimodal AAD procedures were applied using all basic envelope extraction methods. To provide an example, figure 4 shows the weights that were given to each subband envelope (p-law subbands) for each of the 16 subjects using 'CCA' (left), and the corresponding subband signal power fractions,

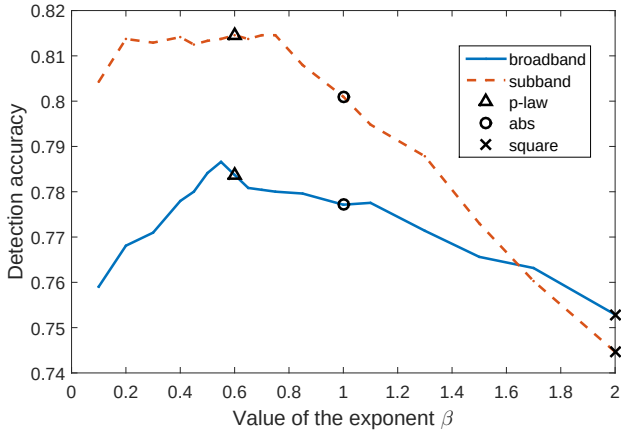

Fig. 3: Performance of the broadband and subband powerlaw envelope extraction method for different values of the exponent β.

Envelope	X	Y	p-value	X(sign)Y
abs	uniform weights	CCA	0.412	<
		band importance	0.140	>
		bimodal LS	0.418	<
square	uniform weights	CCA	0.135	<
		band importance	0.004	>*
		bimodal LS	0.142	<
log	uniform weights	CCA	0.324	<
		band importance	0.128	>
		bimodal LS	0.068	>
p-law	uniform weights	CCA	0.316	>
		band importance	0.056	>
		bimodal LS	0.166	>

TABLE II: Results for the different subband weighting methods

obtained by multiplying the weights and the power fraction of the respective subband (right). The figure shows that weights are consistent across subjects and that the subband envelopes corresponding to the lowest frequencies contribute most to the eventual envelope.

Of both bimodal procedures, 'CCA' yielded a detection accuracy equal to 81.2 %, compared to 79.3% for 'bimodal LS' when applied to p-law subbands. However, this difference in performance was found not to be statistically significant (p-value 0.191). The same holds for 2 of the 3 other basic envelope methods, where 'CCA' only yields significantly better performance compared to 'bimodal LS' when applied to log envelopes (p-values: abs 0.375, log 0.018*, square 0.250). The performance of the 'CCA' approach was however found to be significantly better when applied to the p-law envelopes as compared to the log envelopes (p-value 0.034).

The performance of all the considered subband weighting methods (CCA, bimodal LS and band importance weighting) were also compared to a uniform (all-ones) weighting. The results are shown in table II, where an asterisk '*' in the last column indicates statistical significance at $\alpha = 0.05$ significance level. Uniform weighting was found to be either not significantly different or significantly better than the non-uniform methods.

Finally, we present performance of the best-performing envelope method ('p-law sub') for different trial lengths. Detection accuracy on trials of 60 seconds was equal to 87.5 %. This is comparable to results mentioned in literature

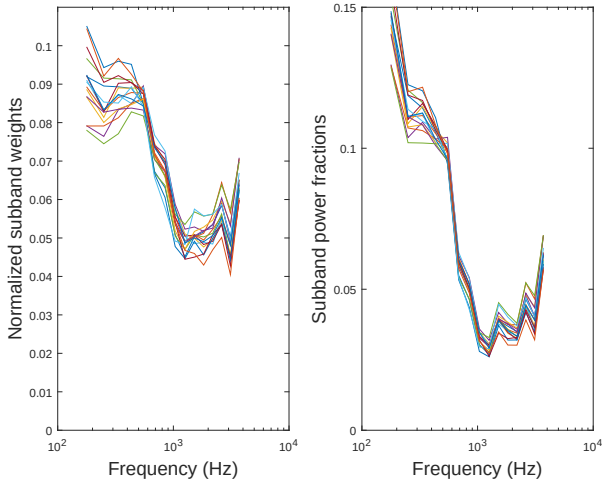

Fig. 4: Normalized subband weights (left), obtained using 'CCA', and the relative signal power contribution of each subband to the eventual envelope (right). Each line represents results for a different subject.

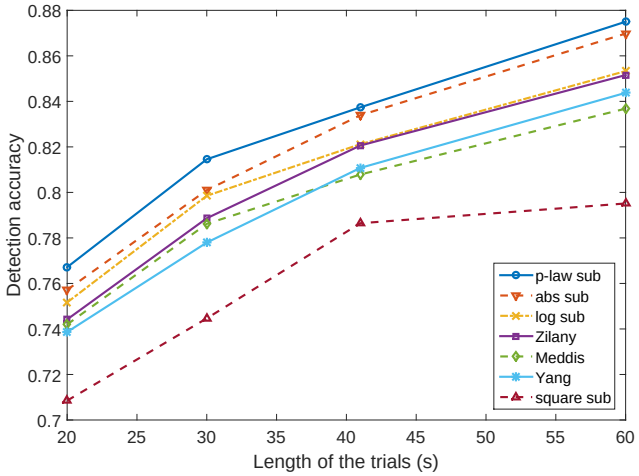

Fig. 5: Evolution of AAD accuracy as a function of trial length for the different subband envelope extraction methods.

[5], [7], especially when considering that our experiment and subsequently our decoders were more general, spanning multiple presentation modes (switching attended ear as well as switching acoustic conditions for each subject). For a more general picture, figure 5 shows the evolution of the AAD accuracy for the different subband envelope extraction methods (with broadband methods omitted for clarity of the figure) as a function of the trial length (20, 30, 40 and 60 seconds were evaluated).

VII. Discussion

A. Implications for application in auditory prostheses

Computational efficiency is an important factor that should be kept in mind when designing an AAD algorithm for future neuro-steered APs. As can be seen from figure 2 and table I, a simple auditory model based on a power law in combination with a gamma tone filterbank yields better results than more complex auditory models ('Yang','Meddis', and 'Zilany'). This is good news, as this shows that it is possible to have a good AAD performance without the methods being computationally taxing on the implementation. Similarly, we also observe that the same advantage holds for using equal weights to recombine subband envelopes in comparison to the more computationally complex 'CCA' or 'bimodal LS' approach. Indeed, from table II, we see that non-uniform weighting shows no significant improvement over uniform weighting, even when the weights are optimized with CCA or bimodal LS. Therefore, for the sake of computational complexity, uniform weighting is preferred in real applications.

B. Future steps towards neuro-steered auditory prostheses

In this paper, we have investigated how auditory models influence AAD performance, showing that the trade-off between complexity and performance is not crucial and does not lead to strong dilemmas. However, there are other complexity-vs-performance trade-offs when considering AAD for neuro-steered APs that still need to be explored, such as the choice of number of EEG channels, number of time lags, and trial length. A long trial length reduces the variance on the correlation estimates and hence improves AAD accuracy (see figure 5), but also reduces the time resolution in the AAD decision. The effect of reducing the number of channels has been investigated in [7]. There is also a need to look into the effect of the acoustic environment such as reverberation and background noise, the potential to improve accuracy when training the subjects, and the effects of closed-loop feedback. Another important consideration towards the design of neuro-steered APs is the extraction of speech envelopes from the speech mixtures recorded with the APs' local microphones. Multiplicative non-negative independent component analysis (M-NICA) [34] has been shown to extract speech envelopes at a low computational cost to support an AAD-assisted noise reduction algorithm [35]. Other techniques such as adaptive beamforming can also be used for unmixing speech signals, keeping in mind the limitations on computational cost, constraints on acceptable latencies etc. Finally, it is noted that the demixing process that extracts the individual speech envelopes will never work perfectly and will inevitably result in some residual noise and cross-talk in the demixed envelopes. In [18], the effect of noisy envelopes on AAD performance was investigated, and it was found that, to some extent, decoding performance is robust to noisy reference signals.

VIII. Conclusion

AAD has the potential to take AP technology a step forward, by allowing to enhance the actual attended speaker, while adapting to the acoustic scenario and shifting auditory attention. We have proposed an 'all-at-once' methodology for training the decoder, in which we solve a single LS problem over the entire data set, rather than averaging over a multitude of per-trial LS solutions as in the existing literature on AAD. We have shown that this may result in better AAD performance, while also reducing the sensitivity with respect to a regularization parameter. The main goal of this paper was to investigate whether AAD performance could be improved by adding some auditory-inspired modifications to the envelope extraction process. We have shown that performing

the envelope extraction on subband signals rather than the broadband audio signal, mimicking the physiology of the auditory periphery, and adding a non-linear power law amplitude compression significantly improved AAD performance. Furthermore, we have shown that using complex models of the auditory periphery did not yield as good results as the simpler proposed methods. We have shown that using a mathematically optimal subband envelope weight vector, based on bimodal LS and CCA optimization methods, did not outperform the heuristic choice of equal weights for our specific dataset.

ACKNOWLEDGMENTS

The authors would like to thank Andreas Prokopiou for his help in preparing the speech models, Jonas Vanthornhout for his help in setting up the experiment, and Michael Hofmann for his suggestions regarding the statistical testing procedure. We would also like to thank all test subjects for their participation in the experiment.

APPENDIX
COMPARISON OF DECODER TRAINING METHODS

In Subsection II-C, two approaches for training the decoders were discussed. In the first approach, preliminary decoders were trained on the data of each single trial, and later averaged, necessitating one of two possible regularization strategies [5], [7], [16], [17]. The second approach, which minimizes a sum of LS objective functions, calculates decoders directly, based on a concatenation of the data from all but one trial.

Figure 6 shows the performance of both these approaches, denoted as (1) and (2) respectively, on trials with a length of 30 seconds (the same trials as described in Section V). For both regularization strategies (minimum norm and smoothness regularization), the performance evolution is shown as a function of the regularization parameter λ. Note that the regularization parameter λ is a relative one, scaled by the mean diagonal entry z of the EEG autocorrelation matrix R (see equation (10)), ensuring it to be independent of an irrelevant scaling of the data. The performance for no regularization ($\lambda = 0$) is not explicitly shown in the figure, but is the same as for the negligible $\lambda = 10^{-5}$ in the case of method (2).

The optimal value of λ for the first approach can be read from the figure to be $\lambda = 0.01$, corresponding to an AAD accuracy of 80.2% when using minimum norm regularization. As performance decreases significantly for larger and smaller values of λ, optimal tuning of this regularization parameter λ is key. This poses a problem, as a λ that is optimal for a specific dataset is not guaranteed to be optimal for another dataset. For the second approach however, the figure shows that any non-negligible regularization decreases its AAD accuracy. Regularization is therefore to be avoided with this approach, yielding an AAD accuracy of 81.5%.

Overall, the figure shows that concatenating measurements (2) rather than averaging decoders from single trials (1) yields higher AAD accuracy, with the additional benefit of not needing any regularization to boost performance. Comparing the AAD binary detection results for the optimal regularization settings for both approaches (as defined above) using a permutation test (see Subsection V-C), yields a p-value equal to 0.022, indicating indeed a statistically significant difference between the two approaches. From this we conclude that optimizing the sum of LS errors objective function is to be preferred over the averaging of single-trial LS solutions.

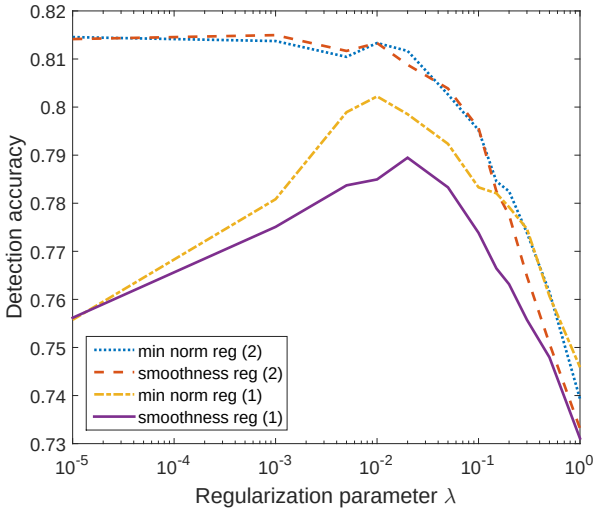

Fig. 6: Performance of the two training approaches and two regularization strategies, as detailed in Subsection II-C, for different values of the regularization parameter λ.

REFERENCES

[1] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand, "Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario.," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, IEEE, 2015.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[4] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, *et al.*, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[5] 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, 2015.

[6] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. National Academy of Sciences*, vol. 109, no. 29, pp. 11854–11859, 2012.

[7] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, 2015.

[8] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a "cocktail party"," *Journal of neural engineering*, vol. 11, no. 4, p. 046015, 2014.

[9] A. Casson, S. Smith, J. Duncan, and E. Rodriguez-Villegas, "Wearable EEG: what is it, why is it needed and what does it entail?," in *EMBC*, pp. 5867–5870, 2008.

[10] D. Looney, P. Kidmose, C. Park, M. Ungstrup, M. Rank, K. Rosenkranz, and D. Mandic, "The in-the-ear recording concept: User-centered and wearable brain monitoring," *Pulse*, vol. 3, pp. 32–42, Nov 2012.

[11] D. Looney, C. Park, P. Kidmose, M. Rank, M. Ungstrup, K. Rosenkranz, and D. Mandic, "An in-the-ear platform for recording electroencephalogram," in *EMBC*, pp. 6882–6885, Aug 2011.

[12] M. G. Bleichner, M. Lundbeck, M. Selisky, F. Minow, M. Jäger, R. Emkes, S. Debener, and M. De Vos, "Exploring miniaturized EEG electrodes for brain-computer interfaces. an EEG you do not see?," *Physiological Reports*, vol. 3, Apr. 2015.

[13] A. Bertrand, "Distributed signal processing for wireless eeg sensor networks," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 23, no. 6, pp. 923–935, 2015.

[14] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Scientific reports*, vol. 5, 2015.

[15] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, "Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, pp. 906–917, 2016.

[16] M. J. Crosse, H. A. ElShafei, J. J. Foxe, and E. C. Lalor, "Investigating the temporal dynamics of auditory cortical activation to silent lipreading," in *IEEE/EMBS International Conference on Neural Engineering, 2015.*, IEEE, 2015.

[17] T. Lauteslager, J. O'Sullivan, R. B. Reilly, E. C. Lalor, *et al.*, "Decoding of attentional selection in a cocktail party environment from single-trial EEG is robust to task," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 1318–1321, IEEE, 2014.

[18] A. Aroudi, B. Mirkovic, M. de Vos, and S. Doclo, "Auditory attention decoding with EEG recordings using noisy reference signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016.

[19] S. S. Stevens, "The measurement of loudness," *The Journal of the Acoustical Society of America*, vol. 27, no. 5, pp. 815–829, 1955.

[20] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and hearing*, vol. 29, no. 2, pp. 139–157, 2008.

[21] R. D. Patterson, M. H. Allerhand, and C. Giguere, "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, 1995.

[22] P. Sondergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 33–56, Berlin, Heidelberg: Springer, 2013.

[23] A. S.3.5-1997, "American national standard methods for calculation of the speech intelligibility index," tech. rep., Acoust. Soc. America, June 1997.

[24] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 824–839, 1992.

[25] M. S. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, 2009.

[26] R. Meddis, W. Lecluyse, N. R. Clark, T. Jürgens, C. M. Tan, M. R. Panda, and G. J. Brown, "A computer model of the auditory periphery and its application to the study of hearing," in *Basic Aspects of Hearing*, pp. 11–20, Springer, 2013.

[27] H. Spoendlin and A. Schrott, "The spiral ganglion and the innervation of the human organ of corti," *Acta Oto-laryngologica*, vol. 105, no. 5-6, pp. 403–410, 1988.

[28] D. D. Greenwood, "A cochlear frequency-position function for several species: 29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[29] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[30] L. Sun, S. Ji, S. Yu, and J. Ye, "On the equivalence between canonical correlation analysis and orthonormalized partial least squares.," in *IJCAI*, vol. 9, pp. 1230–1235, 2009.

[31] deBuren, "Radioboeken voor kinderen." http://www.radioboeken.eu/kinderradioboeken.php?lang=NL, 2007. [Online; accessed: 30-March-2015].

[32] E. J. G. Pitman, "Significance tests which may be applied to samples from any populations," *Supplement to the Journal of the Royal Statistical Society*, vol. 4, no. 1, pp. pp. 119–130, 1937.

[33] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[34] A. Bertrand and M. Moonen, "Blind separation of non-negative source signals using multiplicative updates and subspace projection," *Signal Processing*, vol. 90, no. 10, pp. 2877–2890, 2010.

[35] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *Accepted for publication in IEEE Transactions on Biomedical Engineering*, 2016.

Wouter Biesmans received his M.S. degree in Electrical Engineering with Summa Cum Laude from KU Leuven, Belgium in 2013. From 2013-2015, he researched EEG-based auditory attention detection at the Dept. of Electrical Engineering, KU Leuven. In 2015 he started working in industry, developing digital signal processing algorithms for smartphone sensors.

Neetha Das completed her M.S in Signal Processing from Nanyang Technological University, Singapore in 2013, and her B. Tech. in Electronics and Communication from College of Engineering, Trivandrum, India in 2009. She is currently a PhD student at the Dept. of Electrical Engineering (ESAT) and the Dept. of Neurosciences, KU Leuven, Belgium.

Tom Francart , born 1981 received the M.S. and Ph.D. degrees in engineering from the University of Leuven, KU Leuven, Leuven, Belgium, in 2004 and 2008, respectively. Since 2013 he is a research professor at KU Leuven. His research interests include sound processing for auditory prostheses, binaural hearing and objective measures of hearing.

Alexander Bertrand (M'08) received the Ph.D. degree in Engineering Sciences from KU Leuven, Belgium in 2011. He was a Visiting Researcher at UCLA, Imec-NL Eindhoven, and UC Berkeley. Since 2014, he is an assistant professor at the Dept. of Electrical Engineering (ESAT) of KU Leuven. His research involves signal processing algorithm design for high-density sensor arrays with a focus on biomedical applications.