Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario

Wouter Biesmans[†], Jonas Vanthornhout[⋆], Jan Wouters[⋆], Marc Moonen[†], Tom Francart[⋆], Alexander Bertrand[†]

***Abstract*— Recent research has shown that it is possible to detect which of two simultaneous speakers a person is attending to, using brain recordings and the temporal envelope of the separate speech signals. However, a wide range of possible methods for extracting this speech envelope exists. This paper assesses the effect of different envelope extraction methods with varying degrees of auditory modelling on the performance of auditory attention detection (AAD), and more specifically on the detection accuracy. It is found that sub-band envelope extraction with proper power-law compression yields best performance, and that the use of several more detailed auditory models does not yield a further improvement in performance.**

I. Introduction

Humans are able to focus on a particular auditory stimulus while filtering out all other stimuli, which is known as the cocktail party effect. Recently, it has been shown that, by recording brain activity of a person that is presented an audio mixture of two simultaneous speakers and asked to attend to only one of them, it is possible to detect which of the speakers was attended to [1]–[3]. This auditory attention detection (AAD) paradigm opens up new research possibilities in the fields of neuroscience, audiology and signal processing. One possible future real-world application would be to incorporate AAD in hearing prostheses (HPs), such as hearing aids or cochlear implants. Adding some EEG sensors to a HP would then allow to beamform towards the attended speaker, as opposed to a fixed - but suboptimal - beamforming in the frontal direction as often done in current HPs. This way, the HP will always enhance the attended speech rather than the sound coming from the frontal direction, which may as well be noise or an unattended speaker.

AAD has been shown to be feasible based on high density intra-cranial measurements such as electrocorticography (ECoG) [1] as well as scalp measurements such as magnetoencephalography (MEG) [2] and electroencephalography (EEG) [3], where the latter is the only practical modality for mainstream wearable applications. In particular it has been shown that attention in single-trial EEG recordings of about 60 seconds can be detected [3].

In short, AAD is performed by correlating the envelope of each individual speech signal separately with a reconstructed envelope. Envelope reconstruction is performed by filtering the EEG signals with a spatio-temporal filter or decoder. If the decoder is designed to maximize correlation of its output with the attended speech envelope, the highest correlation value is assumed to correspond to the attended speaker.

Different methods can be used to extract the envelope from the individual speech signals. As such speech envelopes are desired to be highly correlated with the neural representation of speech in the auditory cortex, it is expected that envelopes obtained through increasing detail of auditory modelling will result in increased performance of the subsequent AAD.

Some simple options for envelope extraction are broad-band full-wave rectification followed by low-pass filtering (as often used in electronics), squaring followed by low-pass filtering (to obtain long-term power averages), or taking the absolute value of the Hilbert transform (representing the mathematical envelope). These methods do not explicitly model the physiology of the auditory periphery.

More physiologically-motivated techniques include applying power-law amplitude compression as a very simple model for loudness growth [7], or preprocessing the speech signal in perceptually uniform frequency sub-bands after which an envelope for each sub-band is extracted. The latter technique models the behaviour of the basilar membrane in the inner ear.

Even more complex auditory models [4]–[6] can be used that model the full auditory periphery, including the outer to inner ear, basilar membrane, hair cells and possibly neuronal behaviour.

The goal of this paper is to investigate whether the choice of envelope extraction method significantly affects the detection accuracy of the AAD, and if so, which methods should be preferred. We show that some basic auditory modelling, such as calculating sub-band envelopes and applying power-law compression, results in increased performance, and that the use of several more detailed auditory models does not further improve performance.

The paper is organised as follows: Section II reviews the AAD procedure. Section III describes the different envelope extraction methods that will be analysed. Section IV describes the details of the behavioural experiment, as well as the details of the applied preprocessing. Section V discusses the results and finally Section VI concludes this paper.

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 Research Council CoE PFV/10/002 (OPTEC), OT/14/119 and BOF/STG-14-005, Research Project FWO nr. G.0662.13 'Objective mapping of cochlear implants', IWT O&O Project nr. 110722 'Signal processing and automatic fitting for next generation cochlear implants'. The scientific responsibility is assumed by its authors.

[†] 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

II. Auditory Attention Detection Procedure

Our goal is to detect which of two simultaneous speakers a subject is attending to by reconstructing the attended speech envelope $S_{att}(t)$ from the C-channel EEG measurement $M(t,c)$, where t is the discrete time index and c is the channel index. Similar to [2], [3] reconstruction is achieved by means of a spatio-temporal decoder $D(n,c)$ as follows:

$$\tilde{S}_{att}(t) = \sum_{n=0}^{N-1} \sum_{c=1}^{C} D(n,c) M(t+n,c) \tag{1}$$

In words, the attended speech envelope is reconstructed as a weighted sum of all C EEG channels as well as $N-1$ time-delayed versions of all of these EEG channels. The weights are contained in the decoder matrix $D \in \mathbb{R}^{N \times C}$ and can, for example, be determined by minimizing a least-squares error objective function:

$$\tilde{D} = \arg\min_{D} E[|\tilde{S}_{att}(t) - S_{att}(t)|^2], \tag{2}$$

where $E[.]$ denotes expected value.

It is interesting to observe that this objective function yields the same solution (up to an irrelevant scalar) as when one would design $\tilde{D}$ such that $\tilde{S}(t)$ and $S(t)$ are maximally correlated:

$$\tilde{D} \sim \arg\max_{D} E[\frac{\tilde{S}_{att}(t) S_{att}(t)}{|\tilde{S}_{att}(t)||S_{att}(t)|}]. \tag{3}$$

By introducing the vectors

$$\mathbf{m}_c(t) = [M(t,c)\ M(t+1,c) \cdots M(t+N-1,c)]^T \in \mathbb{R}^{N \times 1} \tag{4}$$

$$\mathbf{m}(t) = [\mathbf{m}_1(t)^T\ \mathbf{m}_2(t)^T \cdots \mathbf{m}_C(t)^T]^T \in \mathbb{R}^{NC \times 1}, \tag{5}$$

i.e. by simulating time-lags as additional, time-shifted EEG channels, we can rewrite (1) as

$$\tilde{S}_{att}(t) = \mathbf{d}^T \mathbf{m}(t) \tag{6}$$

where $\mathbf{d} \in \mathbb{R}^{NC \times 1}$ represents D in vector format, i.e. with all of its columns stacked.

The optimal decoder can be derived by setting the derivative of (2) with respect to the elements of D to zero, or by solving (3) using Lagrange multipliers after reformulating the denominator as a constraint, resulting in:

$$\tilde{\mathbf{d}} = R^{-1}\mathbf{r}, \tag{7}$$

where $R = E[\mathbf{m}(t)\ \mathbf{m}(t)^T] \in \mathbb{R}^{NC \times NC}$ is the EEG autocorrelation matrix and $\mathbf{r} = E[\mathbf{m}(t)\ S_{att}(t)] \in \mathbb{R}^{NC \times 1}$ is a vector containing the cross-correlations of the attended speech envelope and the (time-delayed) EEG signals.

An alternative procedure for calculating a suitable decoder is based on a Generalized Eigenvalue Decomposition (GEVD) that maximizes correlation of the reconstructed envelope with the attended speech envelope while minimizing it with the unattended speech envelope [2]. Either method could be used for this study, but we chose the first because of its simplicity.

AAD is performed in two stages. In the first stage, the decoders are trained using the EEG signals and the attended speech envelope as in (7). As the behavioural experiment results in multiple measurement trials per subject, decoders are always trained using a subject-specific leave-one-out cross validation (see section IV for more details). When multiple trials are used to construct the decoder, this is implemented by concatenating (not averaging [1]) their respective EEG signals and speech signals over time, and using the concatenated signals to estimate the correlation matrix R and cross-correlation vector $\mathbf{r}$. This concatenation results in an improved estimate of the correlation matrices and is less arbitrary than averaging decoders $\tilde{\mathbf{d}}$ obtained by each trial individually.

In the second stage, for each trial the trained decoder $\tilde{\mathbf{d}}$ is used to reconstruct the attended speech envelope $\tilde{S}_{att}(t)$ from the EEG signals (cfr. (1)). Correlation values of the reconstructed envelope with the envelope of both speech signals are then calculated and compared. The speech envelope that has the highest correlation with the reconstructed envelope is classified as the attended speech. As a performance measure, the detection accuracy can then be calculated as the fraction of detections that are performed correctly, across all trials.

Note that in (2) we defined a decoder $\tilde{D}$ that attempts to reconstruct the attended speech envelope $S_{att}(t)$. We can also define a decoder that reconstructs the unattended speech envelope $S_{unatt}(t)$ analogously, but we found that this unattended decoder, unlike the attended decoder, is very ear-specific, i.e. it can only successfully reconstruct unattended speech that is presented at the same ear as in the trials that were used to train the decoder. Thus for real-world applications the attended decoder is more interesting as it is more generally applicable.

III. Envelope Extraction Methods

The goal of this paper is to assess the effect of different speech envelope extraction methods, with varying degrees of auditory modelling, on the performance of AAD. It is to be expected that methods that model the auditory periphery in more detail, approach the actual neural representation (as measured by the EEG) more closely, even though none of them account for the higher level processing that takes place in the brain stem and auditory cortex. It is not clear however how significant and therefore relevant the effect of such a more accurate representation is for our application, i.e. AAD. This section will describe the different methods for extracting a speech envelope $S(t)$ that are assessed in section V.

Starting from the broadband speech signal, four simple methods are examined with no or little regard for physiological correctness. The first method, referred to as 'abs', calculates the absolute value (= full-wave rectification) of the speech signal and then applies low-pass filtering. This method is often used in analogue electronics and is computationally very efficient. The second method, 'hilbert',

[1] It is noted that, although each experiment consists of several trials, we do not average the EEG data over trials, i.e., this is a single-trial method. Furthermore, not all trials use the same speech signal (see section IV).

calculates the envelope as the amplitude of the (complex) Hilbert transform of the speech signal. In the case of a modulated sine wave, this results in the modulating signal. The third method, 'square', squares the speech signal before integrating the signal over some time frame (equivalent to low-pass filtering). This results in a signal that can be thought of as the long-term energy envelope of the speech signal.

Whereas we can think of the first two methods to have no amplitude compression (linear methods, $\alpha = 1$), the third involves a quadratic compression (or expansion rather, $\alpha = 2$). However, the actual relation between perceptual loudness growth and stimulus amplitude $a(t)$ is often modelled through a power-law (i.e. $S(t) = a(t)^\alpha$) compression with exponent $\alpha = 0.6$ [7], which reflects compression in the auditory periphery and thereafter, and is implemented as a fourth method, 'powerlaw'. Finally, as an alternative to power-law compression, we have also investigated the effect of using a logarithmic compression as in [8].

In the auditory pathway, the speech signal is first split into frequency sub-bands by the basilar membrane before any envelope extraction really takes place. To model this in a simple way, each of the aforementioned four broadband methods is also applied to sub-band signals, obtained by filtering the speech signal by a zero-phase gammatone filter bank [9]. The filter bank contains 17 perceptually uniform gammatone filters, with center frequencies ranging from 156 Hz to 4911 Hz and each with an equivalent rectangular bandwidth (ERB) equal to 1.5. In each of these 17 bands, sub-band envelopes are extracted by each of the four methods above. These sub-band envelopes are then added together again, as the measured EEG also contains some combination of these electrical responses represented by sub-band envelopes. As an alternative to a simple summation of the envelopes, we have also investigated the use of an additional frequency weighting based on the band importance functions used in the speech intelligibility index (SII) [10], before summing the envelopes.

For extraction of even more realistic envelopes we refer to three complete, well-known auditory models. The first, 'Yang' [4], decomposes the speech signal into 128 sub-bands, and models the non-linear response of the hair cells and accounts for lateral inhibition. The second and third auditory models, 'Meddis' [5] and 'Zilany' [6], model even more complex auditory mechanics such as efferent feedback and neuron behaviour. The output of these models that will be used in this paper are the instantaneous neuron firing rates. As these models offer free choice as to which neurons' firing rates are calculated, 17 neurons corresponding to the same 17 frequencies as the gammatone filter bank are used. As the output signals of these last two models correspond to neuron firing rates, we weigh them by the neuronal density [11], [12] at their respective frequencies before they are summed to again form one 'envelope'.

Note that low-pass filtering or integration mentioned as a last step in some of the methods is ignored, as it is followed by stricter bandpass filtering later on (see section IV), as a simple approximation of further cortical processing.

IV. Experimental procedures

A. Experiment

Three parts of two Dutch short-stories, read by different male speakers, were selected. The parts were chosen to last approximately 100 seconds after truncation of possible silences to 300 ms. In each trial, the subject was presented one part of each of the two stories simultaneously through insert phones (Etymotic ER3A), to simulate a so-called cocktail party scenario. The subject was asked to attend to only one of both stories, and ignore the other.

The required direction of attention (left or right), as well as the direction from which each individual speaker was presented, was frequently alternated between trials.

In condition 1, each of the two earphone signals contained a single (and different) speaker, presented at 57 dBA. In condition 2, the stimuli presented to each ear were processed by head-related transfer functions (HRTFs) first, simulating a more realistic scenario in which each speaker is simulated to be spatially located either 90 degrees left or right from the frontal direction of the subject. For this condition the stimulation level was only 48dBA.

Each subject attended each part of each story in each ear twice for each of the conditions. This resulted in (2 stories x 3 parts x 2 ears x 2 conditions x 2 repetitions =) 48 trials per subject.

The effects of these different conditions (attended direction and speaker, with or without HRTFs) are not examined here, but as each decoder was trained using 47 out of these 48 trials, it is assured that the calculated decoders are not overfitted towards one speaker, one ear, or the HRTF versus dry-speech condition.

Seven normal-hearing (verified by audiometry) subjects between 20 and 26 years old participated in the experiment. During the experiment their EEG was continuously measured using a 64-channel BioSemi ActiveTwo system in an electromagnetically shielded and soundproof room. The 64 EEG electrodes were placed on the subjects' head according to the international 10 - 20 system. It should be noted that although all seven subjects were native Dutch speakers, two of them were raised bilingually and showed significantly worse detection accuracies compared to the other five: they were the only subjects for which detection accuracies did not always rise above chance level.

B. Processing

The raw EEG signals were bandpass filtered between 2 and 9 Hz and downsampled from the original 8192 Hz sample-rate to 40 Hz. Every digital filter used in the processing has a zero-phase response.

Each of the six single-speaker audio parts was low-pass filtered at 5 kHz (cfr. transducer cut-off frequency) before extracting the envelope using one of the methods described in section III. After envelope extraction, the resulting envelope was processed identically as the EEG, i.e. bandpass filtered between 2-9 Hz, and downsampled to a 40 Hz sample-rate.

For the decoder, time-lags from 0 to 250 ms were used which, at a 40 Hz sample rate, corresponds to $N = 11$ time-lags. For each trial, a different decoder was trained,

using all 47 other trials from the same subject (cfr. leave-one-out cross validation). In this way the decoder is trained independently of the dataset it will be applied to, in order to avoid overfitting, which is indeed a concern as the decoder is high-dimensional. For the evaluation stage (i.e. the actual AAD), only the first 60 seconds of each trial were used to compute the correlation coefficient with the two speech envelopes.

V. Results and Discussion

The detection accuracy of AAD using the different envelope extraction methods detailed in section III is shown in figure 1. To not overload the figure, the logarithmic compression and the SII-based frequency weighting are not shown. However, we briefly note that neither resulted in a significant improvement of AAD performance compared to the competing methods in figure 1 (power-law and unweighted sub-band summation, respectively). As inter-subject differences are large, detection accuracies for each subject (symbols) are provided along with the average detection accuracy over all subjects (bars).

To compare the performance of the different methods pair-wise, we applied one-tailed Wilcoxon signed-rank tests ($\alpha < 0.05$, no multiple comparisons correction) for paired samples to the subject-specific detection accuracies. Both 'square' methods performed significantly worse than most other methods. In addition, both 'powerlaw' methods and 'Zilany' performed significantly better than the 'abs' and 'hilbert' broadband methods. The 'powerlaw' sub-band and 'Zilany' method even performed significantly better than the 'Yang' and 'Meddis' method. Other pairwise comparisons did not reach significance.

Applying the same test to compare all grouped broadband results with their respective sub-band results, shows that sub-band methods significantly ($p < 0.01$) outperform their broadband counterparts (cfr. coloured lines in figure 1).

From these results we can conclude that using knowledge of the auditory periphery pays off, at least up to a certain degree. The 'powerlaw' method, inspired by a simple model of loudness growth, has the best performance of the 4 'simple' methods. Additionally sub-band envelope extraction methods resulted in higher detection accuracies than broadband methods.

Finally, it seems that for AAD, calculating sub-band envelopes and applying power-law compression results in similar or better performance than the 3 more advanced auditory models. This could be explained by the fact that neither simple methods nor the full auditory models account for higher level processing in the brainstem and auditory cortex, which might render detailed modelling of the auditory periphery useless.

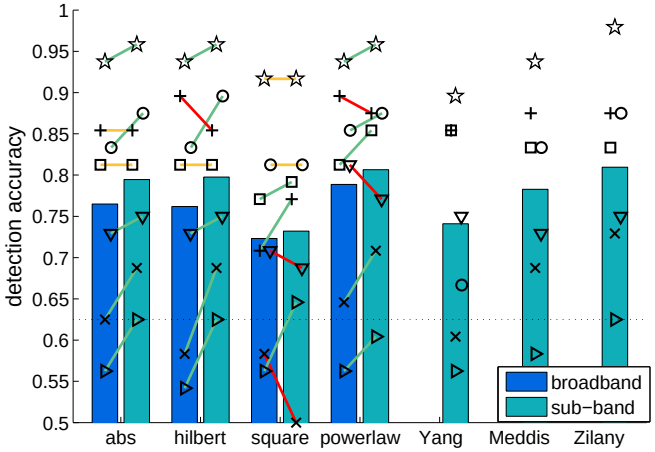

Fig. 1: Mean (bars) and individual subject (symbols) detection accuracies for each of the different envelope extraction methods. Coloured lines connect each subject-specific broadband performance with its corresponding sub-band performance. The dotted black line at 62.5% indicates the detection accuracy which is only 5% likely to be surpassed by chance, based on a binomial distribution ($p = 0.5$, $n = 48$, $\alpha = 0.05$).

VI. Conclusion

We have shown that inclusion of some basic auditory modelling in speech envelope extraction, such as a power-law compression and/or the use of an auditory inspired filter bank, significantly improves the performance of AAD. However, the use of several more detailed models of the full auditory periphery did not further increase performance.

Acknowledgment

The authors would like to thank Andreas Prokopiou for his help in setting up the speech models, and Simon Van Eyndhoven and Duygu Ataman for their help in the data collection.

References

[1] 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.

[2] 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.

[3] 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*, p. bht355, 2014.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

[12] D. D. Greenwood, "A cochlear frequency-position function for several species29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.