Adaptive Attention-driven Speech Enhancement for EEG-informed Hearing Prostheses

Neetha Das[1,2], Simon Van Eyndhoven[1], Tom Francart[2], and Alexander Bertrand[1]

***Abstract*— State-of-the-art hearing prostheses are equipped with acoustic noise reduction algorithms to improve speech intelligibility. Currently, one of the major challenges is to perform acoustic noise reduction in so-called cocktail party scenarios with multiple speakers, in particular because it is difficult -if not impossible- for the algorithm to determine which are the target speaker(s) that should be enhanced, and which speaker(s) should be treated as interfering sources. Recently, it has been shown that electroencephalography (EEG) can be used to perform auditory attention detection, i.e., to detect to which speaker a subject is attending based on recordings of neural activity. In this paper, we combine such an EEG-based auditory attention detection (AAD) paradigm with an acoustic noise reduction algorithm based on the multi-channel Wiener filter (MWF), leading to a neuro-steered MWF. In particular, we analyze how the AAD accuracy affects the noise suppression performance of an adaptive MWF in a sliding-window implementation, where the user switches his attention between two speakers.**

I. Introduction

People with hearing impairment often have difficulties to understand speech in noisy environments, leading to social isolation and decreased quality of life. This can be partly overcome by embedding digital signal processing algorithms in auditory prostheses such as hearing aids or cochlear implants. By using an array of microphones, it is possible to apply beamforming techniques that exploit spatial characteristics of the acoustic scenario to extract the sound from a target direction while reducing background noise from other directions. Advanced beamformers, such as the multi-channel Wiener filter (MWF) [1] optimally suppress noise in any acoustic scenario by continuously measuring the statistics of the background noise and adapting the beamformer coefficients accordingly.

The MWF relies on the fact that a speech signal contains many pauses, during which the noise statistics can be measured. The MWF uses a voice activity detection (VAD) mechanism to identify the segments during which the speaker is silent. However, in a so-called cocktail-party scenario, the background noise contains interfering speakers, such that a fundamental problem emerges: which speaker is the listener actually attending to, i.e., which speech signal should be tracked by the VAD, and which speech signal(s) should be treated as noise. We will refer to this problem as auditory attention detection (AAD). In practice, without extra knowledge about the listener's intentions, the AAD problem can only be tackled in a heuristic or pragmatic way, e.g., by selecting the speaker with highest intensity, or by selecting the speaker closest to the frontal direction (requiring speaker localization). However, such a system will often choose the wrong speaker, in which case the MWF will be adapted to suppress the attended speech, rather than enhancing it.

Several studies have demonstrated that auditory attention can be decoded from electroencephalography (EEG) recordings in a two-speaker scenario [2]–[5]. In [3], [4], it is suggested that hearing prostheses could be extended with chronic and discreet EEG recording technology [6]–[8], to form so-called neuro-steered hearing prostheses. Such devices would allow to track auditory attention based on EEG recordings, and then adapt the beamforming algorithm to steer towards the identified speaker of interest. A first proof of concept for such an EEG-informed noise reduction algorithm has been described in [9] for a batch-mode (non-adaptive) version of MWF.

In this paper, we investigate whether the accuracy of EEG-informed AAD allows to *adaptively* steer an MWF-based beamformer to extract the attended speaker from the microphone recordings of a binaural hearing prosthesis. To this end, we use a sliding-window implementation of the least-squares based AAD algorithm originally proposed in [2]. We assume that the envelopes of the two speech sources are available, which are used as inputs to the AAD algorithm, as well as voice-activity tracks for the MWF. In practice, these speech envelopes can be wirelessly transmitted to the hearing prosthesis by a multimedia device that produces the speech signals through its loudspeakers [10] or by an external reference microphone [11]. If this is not possible, the envelopes have to be extracted from the microphone recordings at the hear prosthesis, e.g., using techniques such as in [12]. The latter is beyond the scope of this study, but was treated in [9]. The impact of imperfect or noisy speech envelopes to perform AAD has also been investigated in [5], where it was found that the AAD performance is quite robust to uncorrelated noise in the envelopes.

[1] KU Leuven - Dept. Electrical Engineering, Stadius Center for Dynamical Systems, Signal Processing, and Data Analytics. Correspondence to: Neetha.Das@esat.kuleuven.be.

[2] KU Leuven - Department of Neurosciences, Research Group Experimental Oto-rhino-laryngology.

This work was carried out at the ESAT and ExpORL Laboratory of KU Leuven, in the frame of KU Leuven Special Research Fund BOF/STG-14-005, OT/14/119, and CoE PFV/10/002 (OPTEC), IUAP nr. P7/19 (DYSCO), and has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 637424). The scientific responsibility is assumed by its authors. The experiments were approved by the KU Leuven ethical committee and all subjects (and/or their legal guardian) signed an informed consent form.

II. Problem statement and algorithms

A. Problem statement

We consider a hearing prosthesis equipped with a microphone array with M microphones, observing an acoustic scene with two speech sources. The m-th microphone signal is described in the frequency domain as

$$y_m(\omega) = x_{1,m}(\omega) + x_{2,m}(\omega) + n_m(\omega)\,,\ m = 1,\ldots,M \quad (1)$$

where $x_{1,m}$ and $x_{2,m}$ denote the signal components corresponding to speaker 1 and 2, respectively (as observed at microphone m), n_m denotes the noise picked up by microphone m, and ω denotes the frequency variable. $x_{1,m}$ and $x_{2,m}$ consist of speech signals that are filtered by an acoustic transfer function from their respective source to microphone m, capturing the effect of propagation through the acoustic environment and the head. By stacking the M microphone signals in an M-channel signal $\mathbf{y}(\omega) = [y_m(\omega) \ldots y_M(\omega)]^T$, we can write (1) as

$$\mathbf{y}(\omega) = \mathbf{x}_1(\omega) + \mathbf{x}_2(\omega) + \mathbf{n}(\omega) \quad (2)$$

where $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{n}$ denote the stacked versions of the corresponding signal components in (1).

Our goal is to design a multi-channel filter or beamformer $\mathbf{w}(\omega)$ for each frequency ω, which extracts the attended speech source from $\mathbf{y}(\omega)$, i.e., the filtered output signal $z(\omega) = \mathbf{w}(\omega)^H\mathbf{y}(\omega)$ should be an estimate of the attended speech signal, which can be either speaker 1 or speaker 2 (superscript H denotes the conjugate transpose operator).

In the next two subsections, we will briefly explain how (a) the beamformer $\mathbf{w}(\omega)$ is computed, and (b) how the attended speaker can be detected based on EEG signals.

B. MWF-based speaker extraction

In practice, the filter $\mathbf{w}(\omega)$ as well as the filtering operation $\mathbf{w}(\omega)^H\mathbf{y}(\omega)$ are computed for a discrete set of frequencies $\omega_1,\ldots,\omega_{max}$ in the time-frequency domain, e.g., based on a short-time Fourier transform (STFT). In the sequel, we will omit the frequency variable ω for the sake of conciseness.

Assuming without loss of generality that the listener attends to the first speaker, our goal is to estimate the signal x_{1r} as observed in an arbitrarily pre-selected reference microphone r. We compute $\mathbf{w}$ such that z is as close as possible to $x_{1,r}$ in linear minimum mean squared error (LMMSE) sense, i.e.,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E\{|x_{1,r} - z|^2\} = \arg\min_{\mathbf{w}} E\{|x_{1,r} - \mathbf{w}^H\mathbf{y}|^2\} \quad (3)$$

where $E\{\cdot\}$ denotes the expectation operator. The solution of (3) is given by the MWF [1]:

$$\hat{\mathbf{w}} = \mathbf{R}_{yy}^{-1}\mathbf{R}_{x_1x_1}\mathbf{e}_r \quad (4)$$

where $\mathbf{R}_{yy} = E\{\mathbf{y}\mathbf{y}^H\}$, $\mathbf{R}_{x_1x_1} = E\{\mathbf{x}_1\mathbf{x}_1^H\}$, and $\mathbf{e}_r$ denotes the r-th column of an $M\times M$ identity matrix, which selects the column of $\mathbf{R}_{x_1x_1}$ that corresponds to the reference microphone.

$\mathbf{R}_{yy}$ can be directly estimated by means of temporal averaging (over different STFT frames). In this study, $\mathbf{R}_{yy}$ is initialized as $\mathbf{R}_{yy}[0] = 10^{-6}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and an updating procedure with a forgetting factor $0 \ll \lambda < 1$ is used, i.e.,

$$\mathbf{R}_{yy}[k] = \lambda\mathbf{R}_{yy}[k-1] + (1-\lambda)\mathbf{y}[k]\mathbf{y}[k]^H \quad (5)$$

where k denotes the STFT frame index, after removing the frames in which the attended speaker is silent. The latter are used to populate the 'interference-only' correlation matrix $\mathbf{R}_{vv}$, which only contains contributions from the unattended speaker $\mathbf{x}_2$ and the noise $\mathbf{n}$, and for which we use a similar updating scheme as in (5). Assuming independence between all sources, the matrix $\mathbf{R}_{x_1x_1}$ in (4) can then be estimated as $\mathbf{R}_{x_1x_1} = \mathbf{R}_{yy} - \mathbf{R}_{vv}$. However, we used a more robust estimation of $\mathbf{R}_{x_1x_1}$ based on a generalized eigenvalue decomposition of $\mathbf{R}_{yy}$ and $\mathbf{R}_{vv}$, as proposed in [1] and [13] (details omitted). Due to the continuous updating of the above correlation matrices, the MWF $\hat{\mathbf{w}}$ is an adaptive filter.

To distinguish between segments in which the attended speaker is active or silent, we need a speaker-dependent VAD mechanism which only triggers when the attended speaker is active, e.g., based on envelope demixing strategies [9], [12]. In this study, we make abstraction of this problem and assume that the speech envelopes of the two speakers are available, where the VAD tracks are computed by simply thresholding these envelopes.

C. Auditory attention detection

To detect to which of both speakers a listener is attending, EEG data is recorded simultaneously with the microphone signals. Let $r_n[t]$ denote the signal in the n-th EEG channel at sample time t, and let $s_1[t]$ and $s_2[t]$ denote the speech envelope of speaker 1 and 2, respectively. For training data with a known attended speaker, we design a linear EEG decoder that reconstructs the attended speech envelope (say, s_1) from the EEG data in LMMSE sense [2]–[5]:

$$\min_{d_{n\tau}} E\left\{\left|s_1[t] - \sum_{\tau=0}^{T-1}\sum_{n=1}^{N} d_n[\tau]\,r_n[t+\tau]\right|^2\right\} \quad (6)$$

where the $d_n[\tau]$'s define the decoder weights over T time lags and N channels. By stacking all $d_n[\tau]$'s in a vector $\mathbf{d}$, and similarly stacking the corresponding samples $r_n[t+\tau]$ (over all $\tau = 0,\ldots,T-1$ and $n = 1,\ldots,N$) in the vector $\mathbf{r}[t]$, then (6) can be rewritten as

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} E\{|s_1[t] - \mathbf{d}^T\mathbf{r}[t]|^2\} \quad (7)$$

such that the optimal EEG decoder is found as

$$\hat{\mathbf{d}} = \mathbf{R}_{rr}^{-1}\mathbf{c}_{rs_1} \quad (8)$$

where $\mathbf{R}_{rr} = E\{\mathbf{r}[t]\,\mathbf{r}[t]^T\}$, and $\mathbf{c}_{rs_1} = E\{\mathbf{r}[t]s_1[t]\}$. The decoder $\hat{\mathbf{d}}$ can be computed from (8) using training data in which the attended speaker is known. As in [4], (8) is computed once over the entire training data set, instead of over individual trials followed by an averaging of the per-trial decoders (as originally proposed in [2]). If sufficient training data is available, the former avoids to tune a regularization parameter, and generally results in more accurate decoders [4].

To perform AAD on new data, the trained EEG-decoder $\hat{\mathbf{d}}$ is applied on new EEG recordings, and its output signal $p[t] = \hat{\mathbf{d}}^T \mathbf{r}[t]$ is then correlated to the speech envelopes of speaker 1 and speaker 2, where the attended speaker is selected as the one with the highest correlation coefficient. As we target an adaptive algorithm, we compute two time-dependent Pearson correlation coefficients $\rho_{s_1p}[t]$ and $\rho_{s_2p}[t]$ between the EEG decoder output p and the envelope of speakers 1 and 2, respectively, over a sliding window of L seconds. The window includes the current sample at time t and $L-1$ previous samples from the past L seconds. At a given time t, speaker 1 is selected as the attended speaker if $\rho_{s_1p}[t] \geq \rho_{s_2p}[t]$, whereas speaker 2 is selected if $\rho_{s_1p}[t] < \rho_{s_2p}[t]$.

Similar to [2]–[5], we assume in this pilot study that the speech envelopes s_1 and s_2 are known, e.g., provided by an external device or remote microphone [10], [11]. If this is not the case, then the envelopes have to be extracted from the microphone signals in $\mathbf{y}$, as in [9], [12], which is beyond the scope of this paper.

III. Experiment

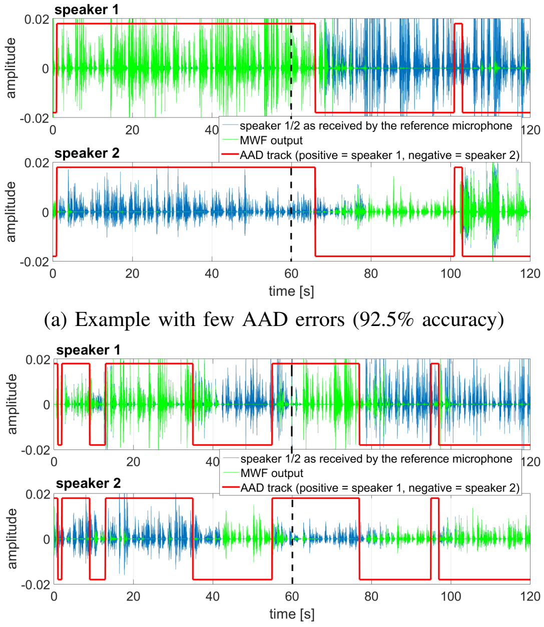
EEG recordings from 16 normal-hearing subjects were collected in a previous study (for details, we refer to [4]). During the experiment, the subjects listened to two simultaneously active speech sources, which were presented at 60dBA using insert phones. The experiment consisted of several trials (26 minutes in total), where the subject was asked to switch attention between left and right ear across trials. The speech sources were filtered with head-related transfer functions (HRTFs) to mimic realistic sound perception from sources impinging on in-the-ear microphones. After filtering, these simulated microphone recordings were downsampled from 44.1 kHz to 8 kHz to reduce the computational load.

We split the EEG recordings into segments of 60s each. We then paired up several segments to generate a set of 72 test frames, for each subject, of 120s each, with attention on speaker 1 during the first 60s and attention on speaker 2 during the last 60s. Hence each test frame contained a switch of attention from speaker 1 to speaker 2 after 60s. For each test frame, a decoder was trained by solving the least-squares problem in (8) over all data not included in the current test frame. The speech envelopes $s_1[t]$ and $s_2[t]$ were computed by powerlaw compression (with power 0.6) on the raw speech signals and downsampling to 20 Hz, followed by band-pass filtering between 2 and 9 Hz, as in [4]. The decoder output $p[t]$ was correlated to $s_1[t]$ and $s_2[t]$ over a sliding window of length L (measured in seconds) at 20 Hz sampling rate with a window shift of 1s. We evaluated the performance for $L = 10$, $L = 20$ and $L = 30$s. Zero-padding was applied at the beginning of each test frame to initialize the sliding window. This resulted in 120 different AAD decisions at a rate of 1Hz over the full length of the test frame. These AAD decisions were used to decide which VAD track (of which speaker) was fed to the MWF, where the VAD track could switch every second depending on the AAD decision at that time point. The MWF was applied on $M = 6$ microphone signals, which were synthesized using HRTFs from a binaural hearing prosthesis with 2x3 microphones [14]. The MWF operated on STFT frames of 256 samples (32 ms), in a weighted overlap-add procedure with 50% overlap, and with a forgetting factor $\lambda = 0.9905$. This value of λ corresponds to approximately a memory retention of 4 seconds for the MWF which was found to be a good choice to stabilize the output SNR against spurious switches in the VAD track due to erroneous AAD decisions. In addition, at instances where a switch in attention was detected, the speech and noise correlation matrices of the MWF were reset to initial conditions effectively forgetting all information before the switch in attention, and hence leading to a faster recovery. In order not to reset the MWF each time a spurious switch in the AAD track is found, we applied a median filter of 11s over the binary signal with AAD decisions. The correlation matrices were only reset when there was a switch in this median output signal. This approach ensured that a switch in attention was taken into account only if it was consistent for at least 5 seconds.

IV. Results

Fig. 1a shows an example of a single trial of 120s where the attention switches from speaker 1 to speaker 2 after 60s (for $L = 20$s). The two plots show the two speech signals as observed at the reference microphone (in blue) and at the output of the MWF (in green). Note that the MWF is computed on the speech mixture as observed at the microphone array, but for the sake of intelligibility, the speech components $\mathbf{x}_1$ and $\mathbf{x}_2$ are fed separately into the resulting MWF to generate the two plots in green, i.e., the actual MWF output signal is the sum of both (green) signals. Similarly, the sum of the blue signals yields the reference microphone signal. The red plot shows the AAD decisions over time, where a positive value corresponds to speaker 1 and negative to speaker 2. Fig. 1a and fig. 1b show trials with high and low AAD accuracy respectively.

Fig. 2 shows the difference between the SNR at the reference microphone (input) and the output of the MWF (0dB corresponds to no difference). The SNR is here defined as the ratio between the power of the attended and the unattended speaker. The plot shows the medians over all subjects and all trials for $L = 10$, $L = 20$ and $L = 30$s, with an attention switch after 60s. The black curve shows the SNR improvement when the MWF is always fed with the VAD track of the attended speaker, which corresponds to a scenario with an instantaneous AAD with perfect accuracy. The error bars (for the red curve) show the first and third quartiles for $L = 20$s, which seems to provide the best trade-off between a quick recovery after an attention switch along with a good SNR improvement and fewer AAD errors that may steer the MWF towards the wrong speaker. Note that the SNR improvement becomes negative at 60s, as the MWF is then still suppressing the attended speaker due to the sudden switch in attention.



(a) Example with few AAD errors (92.5% accuracy)

(b) Example with many AAD errors (63.3% accuracy)

Fig. 1. Two different examples of MWF-based speaker extraction with an attention switch at 60s.

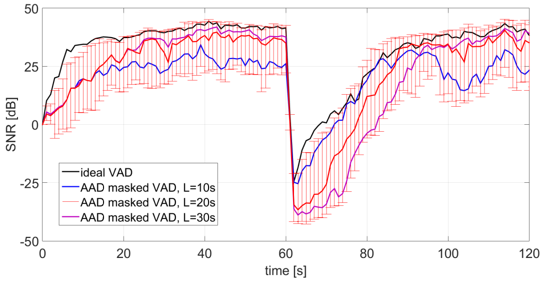


Fig. 2. SNR improvement as a function of time for different values of L (median over all subjects and all trials, with error bars representing first and third quartiles for $L = 20$s).

V. Discussion

Fig. 1a demonstrates that an AAD-informed MWF can extract the attended speaker and that brief AAD errors are not problematic, although they may slightly steer the MWF towards the other speaker (the latter is observed at 10s in Fig. 1b). This can be avoided by using a larger forgetting factor λ, with the drawback that the MWF will then adapt more slowly to changes in attention or in the acoustic scene.

From Fig. 1b, we conclude that a high AAD accuracy is crucial to obtain a stable output. Indeed, when the AAD starts failing regularly after 35s, the unattended speaker leaks to the MWF output, up to a point where it even dominates the output. Such temporary effects have an influence on the overall SNR improvement in Fig. 2, but are not visible due to the averaging. Poor trials as in Fig. 1b are not an exception, and they actually appear quite often in our study. Since the window length L has a large impact on the AAD accuracy, higher values will result in more accurate AAD and a lower probability for the MWF to steer towards the wrong speaker, at the cost of poorer time resolution to track changes in the attention. Finally, it is noted that other effects might play a role which are not taken into account here, such as the subject temporarily being distracted by the unattended speaker. Only an online closed-loop implementation and behavioral assessment can determine whether the MWF enhances the correct speaker (almost) all the time.

VI. Conclusions

We have demonstrated that EEG-informed AAD allows to adaptively steer an MWF, to extract the attended speaker in a two-speaker scenario. We found that a high AAD accuracy is crucial in order to stabilize the MWF and steer it to the correct speaker. When more advanced methods can ensure a robust and accurate AAD, more stable results may be expected. Also, other challenges such as extraction of speech envelopes and reducing the AAD decision delay need to be tackled before a real system can be realized.

References

[1] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230 – 2244, Sep. 2002.

[2] J. O'Sullivan et al., "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–706, 2015.

[3] 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.

[4] 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," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 2016, accepted for publication.

[5] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Influence of noisy reference signals on selective attention decoding," in *Proc. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, Italy, 2015.

[6] 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," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, Nov 2012.

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

[8] 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, no. 16743, 2015.

[9] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," 2016. [Online]. Available: http://arxiv.org/abs/1602.05702

[10] A. Geusens, A. Bertrand, B. Cornelis, and M. Moonen, "Multi-channel noise reduction in hearing aids with wireless access to an external reference signal," in *Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012.

[11] J. Szurley, A. Bertrand, B. Van Dijk, and M. Moonen, "Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal," *IEEE Trans. Audio Speech and Language Processing*, 2016.

[12] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas USA, March 2010, pp. 85–88.

[13] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[14] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–10, 2009.