EEG-based Auditory Attention Decoding Using Unprocessed Binaural Signals in Reverberant and Noisy Conditions*

Ali Aroudi, Simon Doclo

***Abstract*— To decode auditory attention from single-trial EEG recordings in an acoustic scenario with two competing speakers, a least-squares method has been recently proposed. This method however requires the clean speech signals of both the attended and the unattended speaker to be available as reference signals. Since in practice only the binaural signals consisting of a reverberant mixture of both speakers and background noise are available, in this paper we explore the potential of using these (unprocessed) signals as reference signals for decoding auditory attention in different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy). In addition, we investigate whether it is possible to use these signals instead of the clean attended speech signal for filter training. The experimental results show that using the unprocessed binaural signals for filter training and for decoding auditory attention is feasible with a relatively large decoding performance, although for most acoustic conditions the decoding performance is significantly lower than when using the clean speech signals.**

I. Introduction

In complex listening conditions the human auditory system has a remarkable ability to separate a speaker of interest from a mixture of speakers and background noise [1]. Recent studies on the neural activity of the auditory system have shown that cortical responses are correlated with the envelope of the attended speech signal [2], [3]. Based on this finding, a method for decoding auditory attention from single-trial EEG recordings has been proposed in [4]. Moreover, an extensive research effort has recently focused on investigating how to use auditory attention decoding (AAD) as part of a brain computer interface, e.g., for controlling a hearing aid [5], [6], [7], [8], [9].

The AAD method proposed in [4] aims to reconstruct the attended speech envelope from the EEG recordings using a spatio-temporal filter. During the training step, the clean attended speech signal is used to train the filter coefficients by minimizing the least-squares error between the attended speech envelope and the reconstructed envelope. In the decoding step, the clean speech signals of both the attended and the unattended speaker are required as *reference* signals. However, in practice only the binaural signals, consisting of a mixture of the attended and the unattended speech signals and influenced by head-related transfer functions, reverberation and background noise, are available. Although many acoustic signal processing algorithms are available to reduce background noise and perform blind source separation [10], [11], in this paper we explore the potential of using the unprocessed binaural signals for AAD, both in the decoding step as well as in the training step, for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy).

For an acoustic scenario comprising two competing speakers and diffuse noise at different SNRs and reverberation times, 64-channel EEG responses with 18 participants were recorded. The experimental results show that for all acoustic conditions it is possible to decode auditory attention using the binaural signals as reference signals with a relatively large decoding performance, although -as expected- for most conditions the decoding performance is significantly lower than when using the clean speech signals as reference signals. In addition, when using the binaural signals as reference signals, for most conditions there is no significant difference between using the clean speech signals and using the binaural signals for training the filter coefficients.

II. Auditory Attention Decoding

In this section the least squares method used for decoding auditory attention is presented. In Section II-A the different acoustic conditions used for recording EEG responses are defined. In Sections II-B and II-C the training and evaluation steps are discussed, both using either the clean or the binaural signals.

A. Acoustic scenario

Consider an acoustic scenario comprising two competing speakers and background noise in a reverberant environment, where the ongoing EEG responses of a listener to these acoustic stimuli are recorded (cf. Fig. 1). The binaural signals at the ears hence consist of a mixture of both clean speech signals $s^1[n]$ and $s^2[n]$, where n denotes the discrete time index, incorporating head filtering effects, reverberation and background noise. The signal at the m-the ear, with $m = 1$ denoting the left ear and $m = 2$ denoting the right ear, can then be written as

$$y_m[n] = \sum_{j=1}^{2} \underbrace{h_m^j[n] * s^j[n]}_{x_m^j[n]} + v_m[n], \tag{1}$$

with $h_m^j[n]$ the acoustic impulse response between the j-th speaker and the m-th ear, $*$ the convolution operation, $x_m^j[n]$

*This work was supported in part by the PhD Program "Signals and Cognition" and the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG). A. Aroudi and S. Doclo are with the Department of Medical Physics and Acoustics, University of Oldenburg, Germany.

A. Aroudi and S. Doclo are with the Department of Medical Physics and Acoustics, University of Oldenburg, Germany.

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

the speech component of the j-th speaker at the m-th ear, and $v_m[n]$ the background noise component at the m-th ear. The speech component $x_m^j[n]$ consists of an anechoic component $x_m^{j,an}[n]$ (encompassing the anechoic head-related impulse response [1]) and a reverberant component $x_m^{j,re}[n]$, i.e.

$$x_m^j[n] = x_m^{j,an}[n] + x_m^{j,re}[n]. \tag{2}$$

Using (2), the signal at the m-th ear can be rewritten as

$$y_m[n] = \sum_{j=1}^{2} x_m^{j,an}[n] + \sum_{j=1}^{2} x_m^{j,re}[n] + v_m[n].$$

For notational conciseness the time index n will be omitted in the remainder of this paper.

For the EEG recordings we will consider four different acoustic conditions, i.e. *anechoic, reverberant, noisy* and *reverberant-noisy*. Depending on the acoustic condition, the signal at the m-th ear obviously comprises different components. For the *anechoic* condition, it is equal to

$$x_m^{an} = \sum_{j=1}^{2} x_m^{j,an}, \tag{3}$$

for the *reverberant* condition it is equal to

$$x_m^{re} = x_m^{an} + \sum_{j=1}^{2} x_m^{j,re}, \tag{4}$$

for the *noisy* condition it is equal to

$$x_m^{no} = x_m^{an} + v_m, \tag{5}$$

and for the *reverberant-noisy* condition it is equal to

$$y_m = x_m^{re} + v_m. \tag{6}$$

It should be noted that in the experiments (cf. Section III) the positions of the attended and the unattended speaker are not always the same, i.e. sometimes the attended speaker is on the left side of the listener (and the unattended speaker is on the right side), while sometimes the attended speaker is on the right side (and the unattended speaker is on the left side). Therefore, we introduce the indices m_a and m_u, where m_a corresponds to the side of the attended speaker and m_u corresponds to the side of the unattended speaker. Due to the head filtering effect this implies that the broadband energy ratio between the attended speech component and the unattended speech component at the m_a-th ear is always larger than at the m_u-th ear.

B. Training step

In the training step, the attended speaker is assumed to be known and the attended speech envelope $e^a[i]$, with $i = 1 \ldots I$ the sub-sampled time index, is used for filter training. In most previous work, e.g. [4], [5], [6], [7], [12], the envelope of the clean attended speech signal s^a has been used. One of the goals of this paper is to investigate whether it is also possible to use the unprocessed binaural signals at the ears instead of the clean attended speech signal as *training* signal.

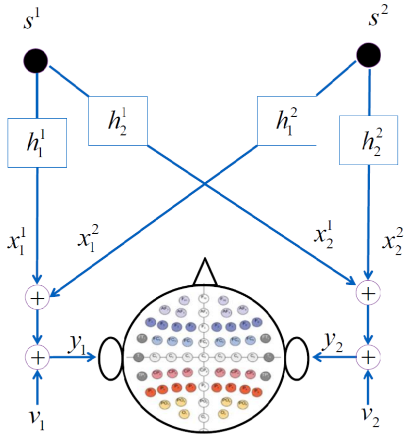


Fig. 1. The binaural acoustic configuration used for stimuli presentation in different acoustic conditions.

The AAD method proposed in [4] uses a spatio-temporal filter to estimate the attended speech envelope $\hat{e}^a[i]$ from C-channel EEG recordings $r_c[i]$ ($c = 1 \ldots C$) as

$$\hat{e}^a[i] = \sum_{c=1}^{C} \sum_{l=0}^{L-1} w_{c,l}\, r_c[i + \Delta + l], \tag{7}$$

with $w_{c,l}$ the l-th filter coefficient in the c-th channel, L the number of filter coefficients per channel, and Δ modeling the latency of the attentional effect in the EEG responses to the speech stimuli. In vector notation, (7) can be written as

$$\hat{e}^a[i] = \mathbf{w}^T \mathbf{r}[i], \tag{8}$$

with

$$\mathbf{w} = \left[\mathbf{w}_1^T\, \mathbf{w}_2^T \ldots \mathbf{w}_C^T\right]^T, \tag{9}$$

$$\mathbf{w}_c = \left[w_{c,0}\, w_{c,1} \ldots w_{c,L-1}\right]^T, \tag{10}$$

$$\mathbf{r}[i] = \left[\mathbf{r}_1^T[i]\, \mathbf{r}_2^T[i] \ldots \mathbf{r}_C^T[i]\right]^T, \tag{11}$$

$$\mathbf{r}_c[i] = \left[r_c[i+\Delta]\;\, r_c[i+\Delta+1] \ldots r_c[i+\Delta+L-1]\right]^T, \tag{12}$$

with $(.)^T$ the transpose operation. During the training step, the filter $\mathbf{w}$ is computed by minimizing the least-squares error between the attended speech envelope $e^a[i]$ (assumed to be known) and the reconstructed envelope $\hat{e}^a[i]$, regularized with the squared l_2-norm of the derivatives of the filter coefficients to avoid over-fitting [4], i.e.

$$J(\mathbf{w}) = \frac{1}{I} \sum_{i=1}^{I} \left(e^a[i] - \mathbf{w}^T \mathbf{r}[i]\right)^2 + \beta \mathbf{w}^T \mathbf{D} \mathbf{w}, \tag{13}$$

with

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}, \tag{14}$$

and β the regularization parameter. The filter minimizing the regularized cost function in (13) is equal to [5], [8]

$$\mathbf{w} = (\mathbf{Q} + \beta\mathbf{D})^{-1}\,\mathbf{q}, \tag{15}$$

with the correlation matrix $\mathbf{Q}$ and the the cross-correlation vector $\mathbf{q}$ equal to

$$\mathbf{Q} = \frac{1}{I}\sum_{i=1}^{I}\left(\mathbf{r}\,[i]\,\mathbf{r}^{T}\,[i]\right),\ \mathbf{q} = \frac{1}{I}\sum_{i=1}^{I}\left(\mathbf{r}\,[i]\,e^{a}\,[i]\right). \tag{16}$$

For each acoustic condition, the set of EEG recordings is segmented into T_{tr} trials. The correlation matrix and the cross-correlation vector corresponding to trial t are denoted as $\mathbf{Q}_t$ and $\mathbf{q}_t$, respectively, where t denotes the trial index. The filter to decode trial t is computed as

$$\tilde{\mathbf{w}}_t = \left(\tilde{\mathbf{Q}}_t + \beta\mathbf{D}\right)^{-1}\tilde{\mathbf{q}}_t, \tag{17}$$

with $\tilde{\mathbf{Q}}_t$ the average correlation matrix of trial t, computed by averaging all correlation matrices except $\mathbf{Q}_t$ (so-called leave-one-out averaging), and $\tilde{\mathbf{q}}_t$ the average cross-correlation vector of trial t, computed by averaging all cross-correlation vectors except $\mathbf{q}_t$.

Since EEG responses are recorded for different acoustic conditions, in this paper we will consider several training conditions (tc) and training signals for computing the filter $\tilde{\mathbf{w}}_t$, i.e. $tc = an$ using EEG responses in the *anechoic* condition and $x^{an}_{m_a}$ in (3) as training signal, $tc = re$ using EEG responses in the *reverberant* condition and $x^{re}_{m_a}$ in (4) as training signal, $tc = no$ using EEG responses in the *noisy* condition and $x^{no}_{m_a}$ in (5) as training signal, and $tc = rn$ using EEG responses in the *reverberant-noisy* condition and y_{m_a} in (6) as training signal.

C. Evaluation step

To decode to which speaker a listener attended during trial t, first an estimate of the the attended speech envelope $\hat{e}^a_t$ is computed using the (trained) filter $\tilde{\mathbf{w}}_t$ in (17), i.e.

$$\hat{e}^a_t = (\tilde{\mathbf{w}}_t)^T\,\mathbf{r}_t, \tag{18}$$

with $\mathbf{r}_t$ the EEG recordings of trial t. Based on the attended and the unattended correlation coefficients,

$$\rho^a_t = \rho\left(e^a_t, \hat{e}^a_t\right),\ \rho^u_t = \rho\left(e^u_t, \hat{e}^a_t\right), \tag{19}$$

with e^u_t the unattended speech envelope, it is then decided that auditory attention has been correctly decoded when $\rho^a_t > \rho^u_t$. The decoding performance p is defined as the percentage of correctly decoded trials over all considered trials and over all participants.

To compute the correlation coefficients in (19), different *reference* signals can be used to compute the attended and the unattended speech envelopes e^a_t and e^u_t, respectively. In most previous work, e.g. [4], [12], it has been assumed that the clean attended and unattended speech signals s^a and s^u are available, which is quite unrealistic in practice. Therefore, one of the goals of this paper is to investigate the decoding performance when using the (unprocessed) binaural signals as *reference* signals in different conditions, i.e. $x^{an}_{m_a}$ and $x^{an}_{m_u}$ in the *anechoic* condition, $x^{re}_{m_a}$ and $x^{re}_{m_u}$ in the *reverberant* condition, $x^{no}_{m_a}$ and $x^{no}_{m_u}$ in the *noisy* condition, and y_{m_a} and y_{m_u} in the *reverberant-noisy* condition.

In this paper we will investigate the decoding performance for several evaluation conditions (ec), with p_{ec}, $ec \in \{an, re, no, rn\}$ denoting the decoding performance corresponding to a specific evaluation condition. Please note that all analyses in this paper are performed with the same training and evaluation conditions, i.e. $tc = ec$, such that the influence of acoustical differences between training and evaluation conditions are excluded. Investigating the influence of such acoustical differences on AAD is beyond the scope of this paper.

In [5] it has been shown that tuning the parameters involved in the filter design (L, Δ, β) plays a key role in optimizing the decoding performance. In order not to favour one specific acoustic condition, in this paper the parameters have been tuned to optimize the average decoding performance over all considered acoustic conditions (per participant).

III. Acoustic and EEG Measurement Setup

Eighteen native German-speaking participants aged between 21 and 34 years with normal hearing took part in this study. Two stories in German, uttered by two different male speakers, were simultaneously presented to the participants using earphones at a sampling frequency of 48 kHz. Among all participants, 8 participants were instructed to attend to the left speaker, while 10 participants were instructed to attend to the right speaker. The stimuli were presented in 11 sessions, each of length 10 minutes, interrupted by short breaks. The participants were instructed to look ahead and minimize eye blinking. During the breaks, the participants were asked to fill out a questionnaire consisting of 10 multiple-choice questions related to each story. Two participants were excluded from the analysis, one participant due to poor attentional performance (as revealed by the questionnaire results) and the other one due to a technical hardware problem.

The presented stimuli at both ears were simulated by convolving the clean speech signals (stories) with (non-individualized) binaural acoustic impulse responses, either from [13] or [14], and adding diffuse noise, generated according to [15]. The left and the right speakers were simulated at $-45°$ and $45°$, respectively. Eight different acoustic conditions were considered (cf. Table I): anechoic, reverberant with a moderate and a large reverberation time ($T_{60} = 0.5$ s, $T_{60} = 1$ s), noisy with two different signal-to-noise ratios (SNR = 9.0 dB, SNR = 4.0 dB), and three combinations of reverberation and noise. For each participant the anechoic condition was assigned to the first session and subsequently to every other third session (i.e. session 4, 7, and 10). Aiming at minimizing the influence of the speech material on AAD, the acoustic conditions (except for the anechoic condition) were randomly assigned to the other sessions. For experimental analysis, the acoustic conditions

TABLE I

ACOUSTIC CONDITIONS USED FOR EXPERIMENTAL ANALYSIS AND STIMULI PRESENTATION.

Experimental Analysis	Acoustic Condition	SNR [dB]	T_{60} [s]
Anechoic	Anechoic [13]	∞	< 0.05
Reverberant	Reverberant I [13]	∞	0.5
	Reverberant II [14]	∞	1.0
Noisy	Noisy I [13]	9.0	< 0.05
	Noisy II [13]	4.0	< 0.05
Reverberant-Noisy	*Reverberant-Noisy* I [13]	9.0	0.5
	Reverberant-Noisy II [13]	4.0	0.5
	Reverberant-Noisy III [14]	9.0	1.0

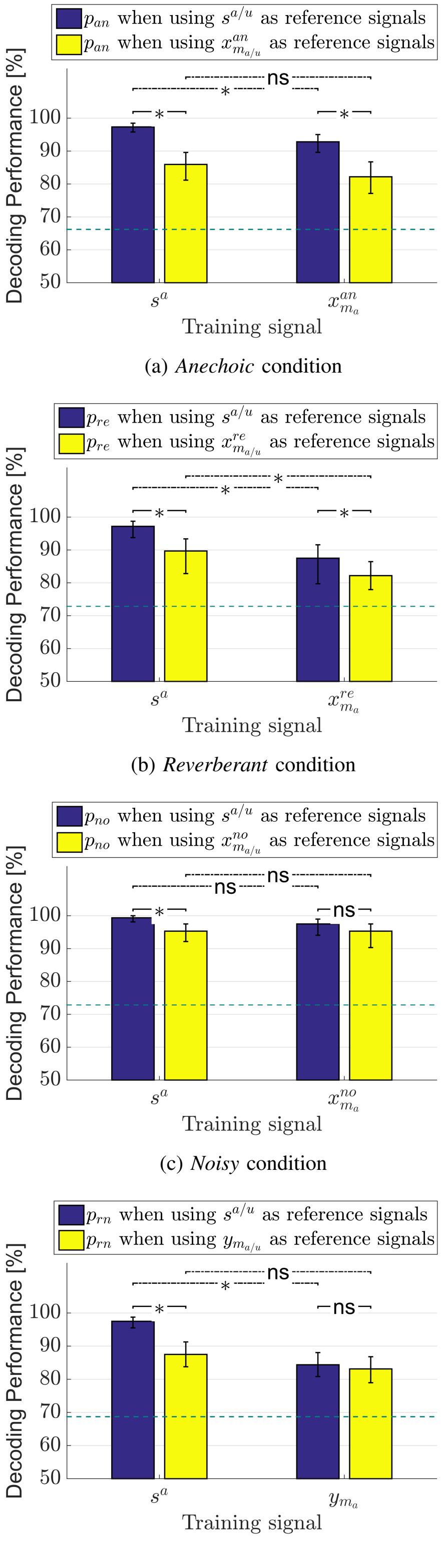
were grouped based on acoustic similarity as shown in Table I, resulting in four experimental analysis conditions, i.e. *anechoic*, *reverberant*, *noisy*, and *reverberant-noisy*.

The EEG responses were recorded using $C = 64$ channels at a sampling frequency of 500 Hz, and referenced to the nose electrode. The EEG responses were offline re-referenced to a common average reference, band-pass filtered between 2 and 8 Hz using a third-order Butterworth band-pass filter, and subsequently downsampled to $f_s = 64$ Hz. The envelopes of the speech signals were obtained using a Hilbert transform, followed by low-pass filtering at 8 Hz and downsampling to $f_s = 64$ Hz. For the training and evaluation steps, the EEG recordings of each session were split into 10 trials, each of length 60 seconds. Each participant's own data were used for filter training and evaluation.

IV. RESULTS AND DISCUSSION

For all considered acoustic conditions (cf. Section II-A and Table I), Fig. 2 presents the decoding performance when using either the clean speech signals ($s^{a/u}$) or the unprocessed binaural signals ($x^{an}_{m_{a/u}}$, $x^{re}_{m_{a/u}}$, $x^{no}_{m_{a/u}}$, $y_{m_{a/u}}$) as *training* and as *reference* signals.

First, we investigate the case where the clean attended speech signal s^a is used as training signal (i.e. left part of the figure). It can be observed that when using the clean speech signals both as training and as reference signals a very good decoding performance (larger than 97%) can be achieved for the anechoic condition, as has been previously shown in [4], [5], [12], as well as for the other considered acoustic conditions. When using the binaural signals as reference signals for decoding, the decoding performance is still significantly larger than chance level (dashed line) for all acoustic conditions, although -as expected- the decoding performance is significantly lower than when using the clean speech signals as reference signals. In [5] it was shown that when using a mixture of the attended and the unattended clean speech signals with a positive signal-to-interference ratio (SIR) as *reference* signals[1], a comparable AAD performance can be achieved as when using clean speech signals as reference signals. This can explain the feasibility of decoding auditory attention using the binaural signals as reference



(a) *Anechoic* condition

(b) *Reverberant* condition

(c) *Noisy* condition

(d) *Reverberant-Noisy* condition

Fig. 2. Comparison of decoding performance (average across all trials and participants) between using the clean speech signals and using unprocessed binaural signals as *training* or *reference* signals in (a) the *anechoic* condition, (b) the *reverberant* condition, (c) the *noisy* condition, (d) the *reverberant-noisy* condition. The asterisks represent the significant decoding performance difference ($p < 0.05$) using a paired Wilcoxon signed rank test, the dashed lines represent the confidence interval's upper boundary of chance level based on a binomial test at the 5% significance level, and the error bars represent the bootstrap confidence interval at the 5% significance level.

[1] Note that in [5] the attended and the unattended speech signals were mixed to simulate the residual cross-talk at the output of a source separation algorithm, hence did not correspond to the signals presented to the listeners while doing EEG recordings.

signals since the broadband energy ratio between the attended speech component and the unattended speech component (SIR) is larger at the m_a-th than at the m_u-th ear due to the head filtering effect.

Secondly, we investigate the impact of using the binaural signal at the m_a-th as training signal instead of the clean attended speech signal (i.e. right part of the figure). On the one hand, when using the clean speech signals as reference signals, for most acoustic conditions (except noisy) there is a significant difference between using the clean attended speech signal or the binaural signal at the m_a-th as training signal. On the other hand, when using the binaural signals as reference signals, for most acoustic conditions (except reverberant) there is no significant difference between using the clean attended speech signal or the binaural signal at the m_a-th as training signal.

In conclusion, these results show that using the binaural signals as training and as reference signals for decoding auditory attention is feasible (with a relatively large decoding performance) for all considered acoustic conditions.

V. Conclusion

In this paper, we have explored the potential of using unprocessed binaural signals for AAD, both in the training step as well as in the decoding step, for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy). The experimental results show that for all conditions it is possible to decode auditory attention using the binaural signals as reference signals with a relatively large decoding performance. In addition, when using the binaural signals as reference signals for AAD, for most conditions there is no significant difference between using the clean speech signals and using the binaural signals for training the filter coefficients.

VI. Acknowledgment

The authors would like to thank Mareike Engelberts and Bojana Mirkovic for the EEG data collection and Kai Siedenburg for his assistance in proofreading the paper.

References

[1] J. Blauert, *Spatial hearing : the psychophysics of human sound localization*. Cambridge, Mass. MIT Press, 1997.

[2] 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, May 2012.

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

[4] 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*, 2014.

[5] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, “Auditory attention decoding with EEG recordings using noisy acoustic reference signals,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 694–698.

[6] S. Van Eyndhoven, T. Francart, and A. Bertrand, “EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses,” *IEEE Transactions on Biomedical Engineering*, no. 99, pp. 1–1, 2016.

[7] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand, “Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses,” in *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, Florida (USA), Aug. 2016, pp. 77–80.

[8] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, “Target speaker detection with concealed EEG around the ear,” *Frontiers in Neuroscience*, vol. 10, p. 349, 2016.

[9] R. Zink, A. Baptist, A. Bertrand, S. Van Huffel, and M. De Vos, “Online detection of auditory attention in a neurofeedback application,” in *Proc. 8th International Workshop on Biosignal Interpretation*, Osaka, Japan, Nov. 2016, pp. 1–4.

[10] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[11] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multi-microphone speech enhancement and source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, 2017.

[12] 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. 46007, 2015.

[13] H. Kayser, S. Ewert, J. Annemüller, 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, p. 10, 2009.

[14] M. Jeub, M. Schäfer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. of International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, Jul. 2009, pp. 1–5.

[15] E. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.