AUDITORY ATTENTION DECODING WITH EEG RECORDINGS USING NOISY ACOUSTIC REFERENCE SIGNALS

Ali Aroudi[1]*, Bojana Mirkovic*[2]*, Maarten De Vos*[3]*, Simon Doclo*[1]

[1] Department of Medical Physics and Acoustics, University of Oldenburg, Germany
[2] Department of Psychology, University of Oldenburg, Germany
[3]Institute of Biomedical Engineering, Oxford University, UK
ali.aroudi@uni-oldenburg.de

ABSTRACT

To decode auditory attention from electroencephalography (EEG) recordings in a cocktail-party scenario with two competing speakers a least-squares method has recently been proposed, showing a promising decoding accuracy. This method however requires the clean speech signals of both the attended and the unattended speaker to be available as reference signals, which is difficult to achieve from the noisy recorded microphone signals in practice. In addition, optimizing the parameters involved in the spatio-temporal filter design is of crucial importance in order to reach the largest possible decoding performance. In this paper, the influence of noisy acoustic reference signals and the spatio-temporal filter and regularization parameters on the decoding performance is investigated. The results show that to some extent the decoding performance is robust to noisy acoustic reference signals, depending on the noise type. Furthermore, we demonstrate the crucial influence of several parameters on the decoding performance, especially when the acoustic reference signals used for decoding have been corrupted by noise.

Index Terms— auditory attention decoding, noisy acoustic reference, speech envelope, brain computer interface

1. INTRODUCTION

The human auditory system has a remarkable ability to separate a speaker of interest from a mixture of speakers or to tune out interfering sounds in a noisy environment, known as the cocktail-party paradigm [1, 2]. Motivated by this observation, during the last decade a large research effort has focused on better understanding the neural activity of the auditory system [3, 4, 5, 6], especially regarding auditory attention in a cocktail-party situation with (two) competing speakers [7, 8]. It has been shown that the auditory cortical responses to speech are correlated with the envelope of attended (and unattended) speech signals [5, 8, 9, 10]. Based on this finding, it has been proposed to decode auditory attention from single-trial EEG recordings using a least-squares method that aims to reconstruct the envelope of the attended speech signal from the EEG recordings [8, 11]. It has been shown that it is possible to decode to which speaker a listener has attended with quite a large accuracy, implying the possibility of controlling the assistive hearing devices with auditory attention decoding in the future [8, 12].

The least-squares-based auditory attention decoding (AAD) method proposed in [8] uses a spatio-temporal filter to reconstruct the attended speech envelope from the EEG recordings. During a training step the filter coefficients are estimated based on a least-squares cost function, aiming to maximize the correlation between the reconstructed envelope and the attended speech envelope. In order to avoid over-fitting to the training data, the least-squares cost function is typically regularized, e.g., using the norm of the filter coefficients. Aiming to optimize the envelope reconstruction accuracy to increase the AAD performance, the spatio-temporal filter parameters (i.e. the number of coefficients and the latency) and the regularization parameter need to be tuned. Based on experimental results we show that tuning these parameters may result in a significant decoding performance increase up to 97%, especially when the acoustic reference signals are corrupted by noise.

The AAD method in [8] however requires the clean speech signals of both attended and unattended competing speakers to be available as acoustic reference signals. In practice, obtaining clean acoustic reference signals from the recorded microphone signals, containing a mixture of the speech signals and background noise, using acoustical signal processing algorithms is hard (if not impossible), since the output signals of blind source separation [13, 14] and speech enhancement algorithms [15, 16] typically contain cross-talk components and residual background noise. A crucial question then arises how robust the AAD is to noisy acoustic reference signals. To investigate the influence of noise on the AAD performance in this paper, we assume that the acoustic reference signals are not equal to the clean speech signals but have been corrupted by either cross-talk components or background noise, e.g., white or speech-shaped noise. Using simulation experiments we show that the least-squares-based AAD method is to some extent robust to noise where the influence of cross-talk (unattended speech signal) is considerably large than for white and speech-shaped noise.

2. AUDITORY ATTENTION DECODING

Let us consider a cocktail-party scenario with two competing speakers, where the ongoing EEG responses of a subject listening to the mixture of these speakers have been recorded. The clean speech signals of the attended and the unattended speaker (clean acoustic reference signals) are denoted as $x_a[t]$ and $x_u[t]$, respectively, where $t = 1 \dots T$ denotes the time index. The envelopes of these speech signals are denoted as $s_a[k]$ and $s_u[k]$, where $k = 1 \dots K$ denotes the sub-sampled time index (cf. section 4). To decode auditory attention from N-channel EEG recordings $r_n[k]$, $n = 1 \dots N$, using a spatio-temporal filter with filter coefficients $w_{n,l}$, an estimate of the attended speech envelope $\hat{s}_a[k]$ is reconstructed using the EEG recordings as

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).

$$\hat{s}_a[k] = \sum_{n=1}^{N}\sum_{l=0}^{L-1} w_{n,l}\, r_n[k+\Delta+l], \quad (1)$$

where Δ models the latency of the EEG responses to the speech stimuli and L denotes the number of filter coefficients. In vector notation, (1) can be rewritten as

$$\hat{s}_a[k] = \mathbf{w}^T \mathbf{r}[k], \quad (2)$$

with

$$\mathbf{w} = \left[\mathbf{w}_1^T\ \mathbf{w}_2^T \dots \mathbf{w}_N^T\right]^T, \quad (3)$$

$$\mathbf{w}_n = [w_{n,0}\, w_{n,1} \dots w_{n,L-1}]^T, \quad (4)$$

$$\mathbf{r}[k] = \left[\mathbf{r}_1^T[k]\ \mathbf{r}_2^T[k] \dots \mathbf{r}_N^T[k]\right]^T, \quad (5)$$

$$\mathbf{r}_n[k] = [r_n[k+\Delta]\ r_n[k+\Delta+1] \dots r_n[k+\Delta+L-1]]^T. \quad (6)$$

During the training step, the filter $\mathbf{w}$ (also referred to as the decoder) is computed by minimizing the least-squares error between $s_a[k]$ (which is known during the training step) and the estimated envelope $\hat{s}_a[k]$, i.e.,

$$J_{LS}(\mathbf{w}) = E\left\{\left|s_a[k] - \mathbf{w}^T\mathbf{r}[k]\right|^2\right\}, \quad (7)$$

where $E\{.\}$ denotes the expected value operator. In order to avoid over-fitting to the training data, the cost function in (7) is typically regularized [3], e.g., using the norm of the filter coefficients, i.e.,

$$J_{RLS}(\mathbf{w}) = E\left\{\left|s_a[k] - \mathbf{w}^T\mathbf{r}[k]\right|^2\right\} + \beta\mathbf{w}^T\mathbf{w}, \quad (8)$$

with β the regularization parameter. The filter minimizing the regularized cost function in (8) is equal to

$$\mathbf{w}_{RLS} = (\mathbf{R} + \beta\mathbf{I})^{-1}\mathbf{r}, \quad (9)$$

with $\mathbf{R} = E\{\mathbf{r}[k]\,\mathbf{r}^T[k]\}$ the auto-correlation matrix, $\mathbf{I}$ the identity matrix, and $\mathbf{r} = E\{\mathbf{r}[k]\,s_a[k]\}$ the cross-correlation vector. For small values of the regularization parameter β, the reconstruction accuracy is increased for training data at the expense of decreasing the reconstruction accuracy for unseen data, referred to as over-fitting. However, for too large values of the regularization parameter, the reconstruction accuracy is severely decreased both for training and evaluation data. Aiming at improving the reconstruction accuracy in order to increase the decoding performance, determining appropriate values for L, Δ and β is hence of crucial importance.

Based on the correlation coefficients between the reconstructed speech envelope $\hat{s}_a[k]$ and the clean attended and unattended speech envelopes, i.e.,

$$\rho_a = \rho(s_a[k], \hat{s}_a[k]), \quad (10)$$

$$\rho_u = \rho(s_u[k], \hat{s}_a[k]), \quad (11)$$

with

$$\rho(z[k], y[k]) = \frac{E\left\{\left(z[k] - \mu_{z[k]}\right)\left(y[k] - \mu_{y[k]}\right)\right\}}{\sigma_{z[k]}\sigma_{y[k]}}, \quad (12)$$

where $\mu_{z[k]}$ and $\sigma_{z[k]}$ denote the mean and the variance of $z[k]$, it is then decided that auditory attention has been correctly decoded when $\rho_a > \rho_u$.

3. NOISY ACOUSTIC REFERENCE SIGNALS

In [8] it has been assumed that the clean attended and unattended speech signals are available for decoding, i.e., for computing the correlation coefficients in (10) and (11). However, in practice these clean speech signals are typically not available. To investigate the influence of noisy acoustic reference signals on the AAD performance, it is assumed here that the reference signals are not equal to the clean speech signals of the speakers but have been corrupted either by each other (simulating residual cross-talk at the output of a source separation algorithm) or by stationary noise (simulating residual noise at the output of a speech enhancement algorithm).

In the first case, the noisy acoustic reference signals are equal to

$$\tilde{x}_a[t] = x_a[t] + \beta_a x_u[t], \quad (13)$$

$$\tilde{x}_u[t] = x_u[t] + \beta_u x_a[t], \quad (14)$$

where the scalars β_a and β_u determine the amount of cross-talk. The amount of cross-talk can be characterized by the signal-to-noise ratio (SNR), i.e.,

$$\mathrm{SNR}_{\tilde{x}_{a/\tilde{x}_u}} = 10\log_{10}\left(\frac{P_{x_{a/x_u}}}{\beta_{a/u}^2 P_{x_{u/x_a}}}\right). \quad (15)$$

where $P_{x_{a/x_u}}$ denote the power of $x_a[t]$ or $x_u[t]$.

In the second case, the noisy acoustic reference signals are equal to

$$\tilde{x}_a[t] = x_a[t] + \alpha_a n_a[t], \quad (16)$$

$$\tilde{x}_u[t] = x_u[t] + \alpha_u n_u[t], \quad (17)$$

where the scalars α_a and α_u determine the amount of stationary noise, and $n_a[t]$ and $n_u[t]$ denote different realizations of the same noise type. The amount of noise can again be characterized by the SNR, i.e.,

$$\mathrm{SNR}_{\tilde{x}_{a/\tilde{x}_u}} = 10\log_{10}\left(\frac{P_{x_{a/x_u}}}{\alpha_{a/u}^2 P_{n_{a/n_u}}}\right). \quad (18)$$

where $P_{n_{a/n_u}}$ denotes the power of n_a or n_u.

For both cases, auditory attention is then correctly decoded when $\tilde{\rho}_a > \tilde{\rho}_u$, with

$$\tilde{\rho}_a = \rho(\tilde{s}_a[k], \hat{s}_a[k]), \quad (19)$$

$$\tilde{\rho}_u = \rho(\tilde{s}_u[k], \hat{s}_a[k]), \quad (20)$$

where $\tilde{s}_a[k]$ and $\tilde{s}_u[k]$ denote the envelopes of $\tilde{x}_a[t]$ and $\tilde{x}_u[t]$, respectively.

4. ACOUSTIC AND EEG MEASUREMENT SETUP

Eight native German-speaking participants (aged between 21 and 29) took part in this study. The participants reported no present or past neurological or psychiatric conditions and normal hearing. Two German stories, uttered by male and female speakers, were presented to the participants using earphones at a sampling frequency of

48kHz. The only difference between the left and the right earphone signals was an interaural level difference of 3dB (i.e., no interaural time difference), so that the participants had the feeling that one speaker was located at the left side and the other speaker at the right side. The participants were asked to attend to either the left or the right speaker during the whole experiment (4 participants to the right speaker and 4 participants to the left speaker). The participants were instructed to keep looking ahead and minimize eye blinking. The presentation of the stories lasted 48 minutes, and was interrupted by four breaks, during which the participants were asked to fill out a questionnaire consisting of eight multiple-choice questions (see [12] for more detailed information about the measurement procedure).

EEG responses were recorded using 96 channels ($N = 96$), referenced to the nose electrode and recorded with a sampling rate of 500 Hz (analog filter settings $0.0153 - 250$Hz). The EEG data were offline re-referenced to a common average reference, band-pass filtered between 2 and 8Hz using a third-order Butterworth band-pass filter, and subsequently downsampled to $f_s = 64$Hz. The speech envelopes of the (noisy) speech signals were obtained using a Hilbert transform of the signals, followed by low-pass filtering at 8Hz and downsampling to $f_s = 64$Hz.

For each participant the 48-minute EEG responses were split into 24 trials, each of length 2 minutes. The leave-one-out cross-validation approach was used for training and testing, i.e., for each trial $j = 1 \ldots 24$ different decoders $\mathbf{w}_{RLS}$ were computed for all other trials, and using the averaged decoder the decoding accuracy for trial j was computed. Each participant's own data were used for decoder training and testing. The average decoding performance was defined as the percentage of correctly decoded trials over all trials and participants.

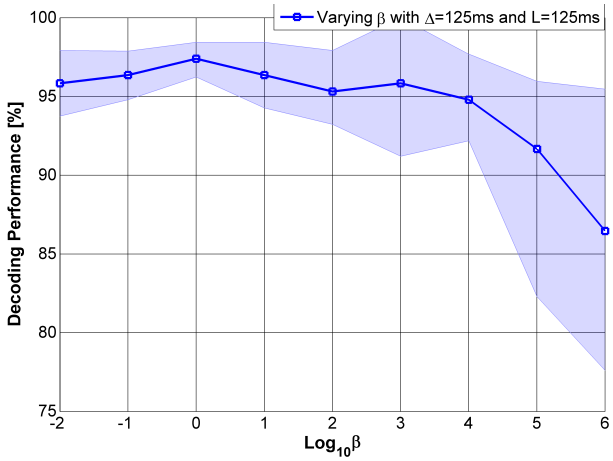
5. EXPERIMENTAL RESULTS

In this section the decoding performance of the least-squares method will be presented using the experimental setup discussed in the previous section. In Section 5.1 the spatio-temporal filter parameters and the regularization parameter will be optimized using clean acoustic reference signals. In Section 5.2 the influence of noisy acoustic reference signals on the decoding performance will be investigated.

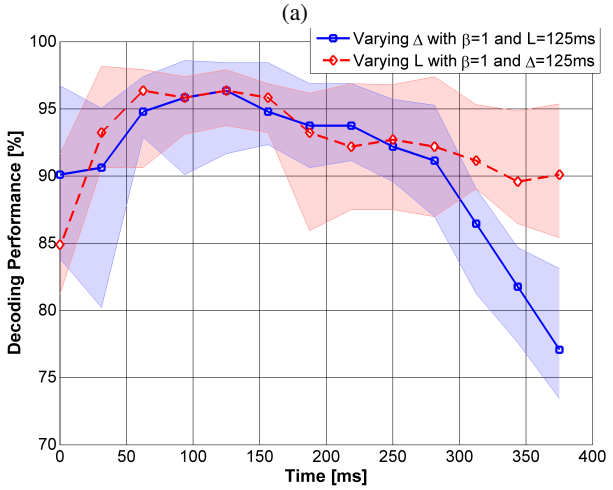
5.1. Clean Acoustic Reference Signals

In order to optimize the AAD performance, we have considered the following parameter values: latency Δ ranging from 0ms to 375ms in steps of 31.25ms (covering the EEG response latencies reported in [3]), filter length L ranging from 0ms to 375ms in steps of 31.25ms, and regularization parameter β ranging from 10^{-2} to 10^{6}. For all possible combinations of parameter values, the average decoding performance was computed over all trials and participants. The largest AAD performance (97%) was obtained for $\Delta = 125$ms, $L = 125$ms and β=1, implying that the most contributing EEG recordings to reconstruct the attended speech envelope are those with latencies between 125ms and 250ms and that regularization is required in order to avoid over-fitting.

In order to investigate the sensitivity of the decoding performance around the optimal parameter values, Fig. 1a depicts the average decoding performance for different values of the regularization parameter β (for $\Delta = 125$ms and $L = 125$ms), whereas Fig. 1b depicts the average decoding performance for different values of the spatio-temporal filter parameters L and Δ (for $\beta = 1$). The shaded area represents the 95% confidence interval, which was



(a)



(b)

Fig. 1. Average decoding performance using clean reference signals for different values of the (a) regularization parameter β, and (b) spatio-temporal filter parameters Δ and L. The shaded area represents the 95% confidence interval.

estimated using bootstrap sampling. The results show that the decoding performance is relatively insensitive to the regularization parameter (only for very large and very small values the decoding performance significantly decreases). In addition, when the latency and the filter length are chosen in the range 62.5ms to 156.25ms, the decoding performance remains relatively high, whereas for smaller and for larger values the decoding performance significantly decreases. Nevertheless, determining appropriate parameters is crucial to achieve a large and robust decoding performance.

5.2. Noisy Acoustic Reference Signals

In this section we investigate the influence of noisy acoustic reference signals (cf. Section 3) on the decoding performance for three types of acoustic noise: cross-talk, white noise, and speech-shaped noise. For all noise types, the mixing parameters β_a and β_u (in (13) and (14)) and α_a and α_u (in (16) and (17)) were computed such that $\text{SNR} = \text{SNR}_{\tilde{x}_a} = \text{SNR}_{\tilde{x}_u}$, with SNRs ranging from -30dB to 30dB in steps of 2dB. Note that for all conditions we have used the

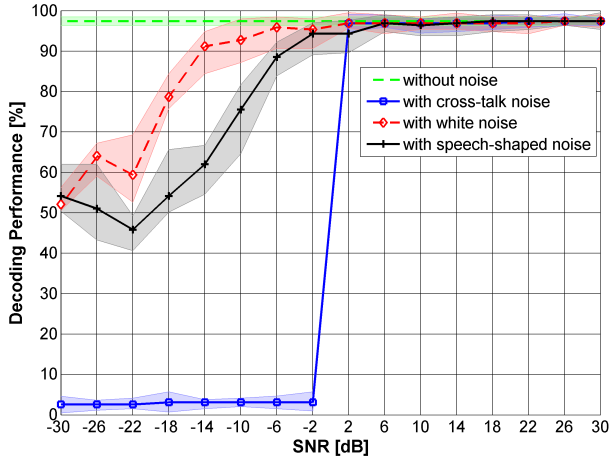


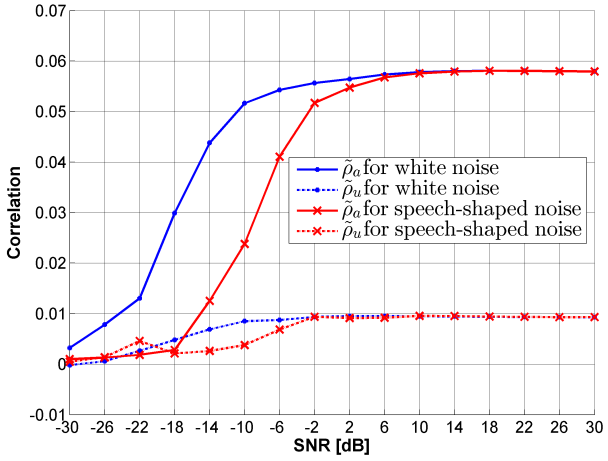
Fig. 2. Average decoding performance for different SNRs (cross-talk, white noise, and speech-shaped noise). The shaded area represents the 95% confidence interval.

decoders trained on clean acoustic reference signals using the optimal parameters determined in Section 5.1.

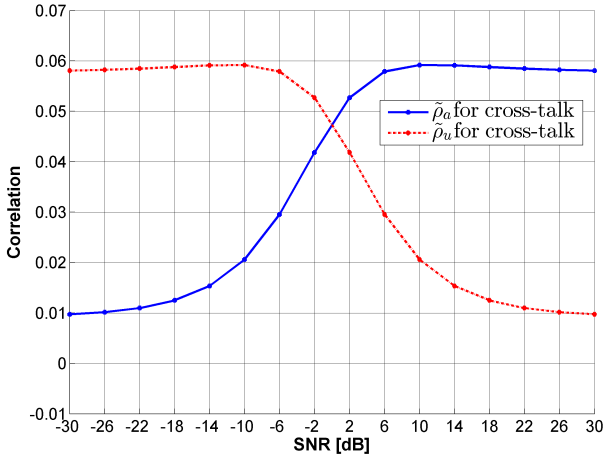
Fig. 2 depicts the average decoding performance for all considered noise types for different SNRs. In general, it can be observed that the influence of white noise is smaller than speech-shaped noise and cross-talk. This can be explained by considering the influence of noise on the correlation coefficients $\tilde{\rho}_a$ and $\tilde{\rho}_u$ in (19) and (20), which are used to decode auditory attention. For different SNRs, Fig. 3 depicts the correlation coefficients between the reconstructed speech envelope and the attended/unattended speech envelope, averaged across participants (note that these average correlation coefficients are not directly used for decoding). On the one hand, from Fig. 3a it can be observed that for white noise and speech-shaped noise both the (average) attended and unattended correlation coefficients decrease, such that below a certain SNR both correlation coefficients become similar and auditory attention can not be accurately decoded any more. On the other hand. from Fig. 3b it can be observed that for cross-talk the (average) attended correlation coefficient decreases, whereas the (average) unattended correlation coefficient increases, such that below a certain SNR (around 0dB) decoding auditory attention becomes impossible.

To evaluate the robustness of the decoding performance against noise, we defined two criteria: 1) the SNR for which the average decoding performance is larger than 90%, denoted as $\mathrm{SNR_L}$, and 2) the mean decoding performance between $\mathrm{SNR_L}$ and $\mathrm{SNR_U} = 30$dB. For the spatio-temporal filter and regularization parameters used in Fig. 2, $\mathrm{SNR_L}$ was equal to 0dB for cross-talk, -4dB for speech-shaped noise and -14dB for white noise, whereas the mean decoding performance was equal to 97% for cross-talk, 91% for speech-shaped noise and 96% for white noise. Similarly as in Section 5.1, we computed the mean decoding performance (i.e. averaged between $\mathrm{SNR_L}$ and $\mathrm{SNR_U}$) for all possible combinations of parameter values. Surprisingly, for all noise types the largest mean decoding performance was obtained using the same parameters as for the clean acoustic reference signals, i.e. $\Delta = 125$ms, $L = 125$ms and β=1.

In summary, the experimental results have shown that the least-squares-based AAD method is to some extent robust to noisy acoustic reference signals, depending on the noise type. In addition, tuning the spatio-temporal filter and the regularization parameters plays a crucial role to increase the performance and the robustness.



(a)



(b)

Fig. 3. Correlation coefficients between the reconstructed speech envelope and the attended/unattended speech envelope, averaged across participants for different SNRs for (a) white noise and speech-shaped noise, and (b) cross-talk.

6. CONCLUSION

In this paper, we have investigated the influence of noisy acoustic reference signals on decoding auditory attention in a scenario with two competing speakers. The experimental results have shown that the least-squares-based AAD method is to some extent robust to noisy acoustic reference signals, where the influence of cross-talk (i.e., unattended speech signal) is considerably larger than for white and speech-shaped noise. In addition, tuning the spatio-temporal filter parameters is of crucial importance, especially when the reference signals used for decoding have been corrupted by noise.

7. REFERENCES

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

[2] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts: The MIT Press, 1990.

[3] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[4] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, 2011.

[5] B. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS Biology*, vol. 10, no. 1, p. 1001251, 2012.

[6] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, 2014.

[7] A. J. Power, J. J. Foxe, E. J. Forde, R. B. Reilly, and E. C. Lalor, "At what time is the cocktail party? A late locus of selective attention to natural speech," *European Journal of Neuroscience*, vol. 35, no. 9, pp. 1497–1503, 2012.

[8] 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.

[9] 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.

[10] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a cocktail party," *Neural Engineering*, vol. 11, no. 4, p. 46015, 2014.

[11] J. A. O'Sullivan, R. B. Reilly, and E. C. Lalor, "Improved decoding of attentional selection in a cocktail party environment with EEG via automatic selection of relevant independent components," in *Proc. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[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] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[14] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[15] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, *Handbook on Array Processing and Sensor Networks (S. Haykin, K. J. Ray Liu, eds.)*. Wiley, 2010, ch. Acoustic beamforming for hearing aid applications, pp. 269–302.

[16] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.