Real-Time Decoding of Auditory Attention from EEG via Bayesian Filtering

Sina Miran, Sahar Akram, Alireza Sheikhattar, Jonathan Z. Simon, Tao Zhang, and Behtash Babadi

***Abstract*— In a complex auditory scene comprising multiple sound sources, humans are able to target and track a single speaker. Recent studies have provided promising algorithms to decode the attentional state of a listener in a competing-speaker environment from non-invasive brain recordings such as electroencephalography (EEG). These algorithms require substantial training datasets and often exhibit poor performance at temporal resolutions suitable for real-time implementation, which hinders their utilization in emerging applications such as smart hearing aids. In this work, we propose a real-time attention decoding framework by integrating techniques from Bayesian filtering, ℓ_1-regularization, state-space modeling, and Expectation Maximization, which is capable of producing robust and statistically interpretable measures of auditory attention at high temporal resolution. Application of our proposed algorithm to synthetic and real EEG data yields a performance close to the state-of-the-art offline methods, while operating in near real-time with a minimal amount of training data.**

I. INTRODUCTION

A distinctive feature of the human brain is the ability to select and attend to a target speaker in an auditory scene with multiple competing speakers. This phenomenon is known as the cocktail party effect and has been studied for decades [1][2]. Since the exact neural mechanisms underlying this phenomenon are mostly unknown, various studies have looked at the problem from a computational perspective, where they try to determine the attentional state of a listener using neuroimaging data in controlled experiments [3][4][5][6][7][8]. To simplify the task, these studies have considered experiments with two competing speakers and have assumed access to clean speech data of the speakers, which are the considered settings in our work as well. Other studies have discussed separation of individual speakers from the audio mixture as a supplementary step [9][10].

Most existing algorithms for attention decoding from EEG [3][7][8][11][12] use a decoder model to relate the observed neural activity to the acoustic features of the speech streams and utilize correlation-based measures of the reconstruction quality to decode the attentional state of the listener. To obtain robust results, they often require large training datasets for offline decoder estimation, e.g., 29 min of training data for 1 min of test data in [3]. Also, when operating at high temporal resolutions, their decoding accuracy degrades significantly [12]. However, in emerging real-time technologies such as brain-computer interface (BCI) systems and smart hearing aid devices, substantial training datasets might not be available, and temporal resolutions in the order of $\sim$1s (comparable to attention switching dynamics) are required. In addition, current methods do not produce statistically interpretable measures of the auditory attention for soft decision-making.

In this work, we address the foregoing issues by estimating the decoder coefficients in real-time, extracting suitable attention-modulated features from the estimates, and defining a state-space model on the features that yields a dynamic, robust, and probabilistic measure of the attentional state. Real-time decoder estimation is carried out by sparse adaptive filtering [5]. The state-space model is motivated by [4] and corrects for the rapid stochastic fluctuations of the extracted features at high temporal resolutions. In Section II, we introduce our proposed framework including the real-time estimation of decoder coefficients and the dynamic state-space model. Section III summarizes the results of applying our framework to both simulated data and real EEG recordings. This is followed by our concluding remarks in Section IV.

II. METHODS

For a dual-speaker experiment, consider a trial of length T samples with sampling frequency of f_s. Let $s_t^{(1)}$ and $s_t^{(2)}$ respectively denote the speech envelopes of speakers 1 and 2, for $t = 1, 2, \ldots, T$. Also, let e_t^c be the EEG response of the listener recorded at time t and channel c, for $c = 1, 2, \ldots, C$, and define $\mathbf{e}_t := [e_t^1; e_t^2; \ldots; e_t^C]$. In a decoding model of lag L_d samples, the role of the decoder for speaker i is to reconstruct $s_t^{(i)}$ from the vector $\boldsymbol{\mathcal{E}}_t := [1; \mathbf{e}_t; \mathbf{e}_{t+1}; \ldots; \mathbf{e}_{t+L_d}]$. We break the test trial into K consecutive, non-overlapping, and small windows of length W samples, i.e., $K := \lfloor \frac{T}{W} \rfloor$, and consider a piece-wise constant approximation to the decoder over these windows. Thus, W/f_s determines the temporal resolution of our proposed framework.

In the k^{th} window, define the covariate matrix as $\mathbf{X}_k := \left[\boldsymbol{\mathcal{E}}_{(k-1)W+1}, \boldsymbol{\mathcal{E}}_{(k-1)W+1}, \ldots, \boldsymbol{\mathcal{E}}_{kW}\right]^\top$ and the target vector for speaker i as $\mathbf{y}_k^{(i)} := \left[s_{(k-1)W+1}^{(i)}; s_{(k-1)W+2}^{(i)}; \ldots; s_{kW}^{(i)}\right]$, for $i = 1, 2$ and $k = 1, 2, \ldots, K$. Motivated by the Recursive Least Squares (RLS) algorithm and Lasso [5][13], the decoder coefficients for speaker i in the k^{th} window, i.e., $\hat{\boldsymbol{\theta}}_k^{(i)} \in \mathbb{R}^{C(L_d+1)+1}$, are updated as

$$\hat{\boldsymbol{\theta}}_k^{(i)} = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{k} \lambda^{k-j} \left\| \mathbf{y}_j^{(i)} - \mathbf{X}_j \boldsymbol{\theta} \right\|_2^2 + \gamma \| \boldsymbol{\theta} \|_1 \quad (1)$$

where hyperparameters $\lambda \in (0, 1]$ and γ are called the forgetting factor and ℓ_1-regularization parameter, respectively.

*This work was supported by National Science Foundation Awards No. 1552946 and 1734892, and a research gift from Starkey Hearing Technologies.

S. Miran, A. Sheikhattar, J. Z. Simon, and B. Babadi are with the Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mails: {smiran,arsha89,jzsimon,behtash}@umd.edu)

S. Akram is with with Facebook, Menlo Park, CA 94025, USA (e-mail: saharakram@fb.com)

T. Zhang is with Starkey Hearing Technologies, Eden Prairie, MN 55344, USA (e-mail: tao_zhang@starkey.com)

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

The parameter λ controls the tradeoff between adaptivity and robustness of the estimates. As a rule of thumb, $\frac{W}{1-\lambda}$ gives the *effective* number of samples used for estimating $\hat{\boldsymbol{\theta}}_k^{(i)}$ in (1) [5]. The parameter γ is determined by cross-validation and enforces sparsity of the estimates $\hat{\boldsymbol{\theta}}_k^{(i)}$. In each window, we solve the modified Lasso problem in (1) using the Forward-Backward Splitting (FBS) algorithm [14], which results in low-complexity updates suitable for real-time applications.

We next compute an attention-modulated feature for *each speaker* in *each window* using the estimated decoder coefficients $\hat{\boldsymbol{\theta}}_k^{(i)}$ and the data $\{\mathbf{y}_k^{(i)}, \mathbf{X}_k\}$. We denote these features by $m_k^{(i)}$, for $i = 1, 2$ and $k = 1, \ldots, K$. For instance, a correlation-based feature is defined as $m_k^{(i)} := \left|\mathrm{corr}\left(\mathbf{y}_k^{(i)}, \mathbf{X}_k \hat{\boldsymbol{\theta}}_k^{(i)}\right)\right|$. The rationale behind this feature is that the decoder corresponding to the attended speaker (with stronger presence in the neural response) is expected to result in more accurate stimulus reconstruction [6]. Another example of an attention-modulated feature is $m_k^{(i)} := \| \hat{\boldsymbol{\theta}}_k^{(i)} \|_1$ with the intercept coefficient of $\hat{\boldsymbol{\theta}}_k^{(i)}$ discarded in the ℓ_1 norm. Earlier studies have argued that the decoder/encoder corresponding to the attended stimulus has more significant peaks around specific time lags in the auditory response [6][15]. Thus, the ℓ_1-based feature is expected to capture such significant coefficient peaks in the decoder of the attended speaker.

We next consider a fixed-lag sliding window framework as shown in Fig. 1. At window $k = k_0$, our goal is to exploit the latest K_A features calculated for each speaker to obtain a probabilistic measure of the attentional state at window $k = k_0 - K_F$. The parameter K_F is the forward-lag hyperparameter and controls the tradeoff between real-time and robust estimation of the attentional state.

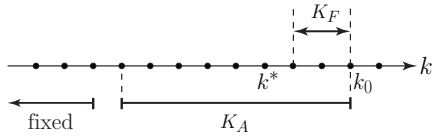

Fig. 1: Fixed-lag sliding window framework.

The extracted features typically exhibit stochastic fluctuations due to calculation within small windows of length W and the presence of background neural activity. We model the feature fluctuations using two different distributions for the attended and unattended speakers, namely the attended and unattended distributions. To this end, we employ Log-Normal distributions, which are unimodal and admit closed-from analytical update solutions, to capture the concentrations in feature values for attended and unattended speakers. For $k = 1, 2, \ldots, K_A$, let n_k be a binary random variable taking value 1 (resp. 2) if speaker 1 (resp. 2) is attended in the k^{th} window. Thus, we have the following observation equations:

$$\begin{cases} \begin{cases} m_k^{(i)} \,\big|\, n_k = i \sim \text{Log-Normal}\left(\rho^{(a)}, \mu^{(a)}\right) \\ m_k^{(i)} \,\big|\, n_k \neq i \sim \text{Log-Normal}\left(\rho^{(u)}, \mu^{(u)}\right) \end{cases}, \quad i = 1, 2 \\ \rho^{(a)} \sim \text{Gamma}\left(\alpha_0^{(a)}, \beta_0^{(a)}\right), \quad \mu^{(a)} \,\big|\, \rho^{(a)} \sim \mathcal{N}\left(\mu_0^{(a)}, \rho^{(a)}\right) \\ \rho^{(u)} \sim \text{Gamma}\left(\alpha_0^{(u)}, \beta_0^{(u)}\right), \quad \mu^{(u)} \,\big|\, \rho^{(u)} \sim \mathcal{N}\left(\mu_0^{(u)}, \rho^{(u)}\right) \end{cases} \tag{2}$$

where $\left\{\rho^{(a)}, \mu^{(a)}, \rho^{(u)}, \mu^{(u)}\right\}$ are the parameters of the Log-Normal distributions, which in turn have conjugate priors with hyperparameters $\left\{\alpha_0^{(a)}, \beta_0^{(a)}, \mu_0^{(a)}, \alpha_0^{(u)}, \beta_0^{(u)}, \mu_0^{(u)}\right\}$.

Our goal is to estimate $p_k := \mathrm{P}(n_k = 1)$ for $k = 1, 2, \ldots, K_A$ with their respective confidence intervals, so that they are robust to the stochastic fluctuations of $m_k^{(i)}$'s. To this end, we consider the following state-space model with parameters $z_{1:K_A}$ and $\eta_{1:K_A}$ on the logit transform of p_k's, to enforce temporal continuity:

$$\begin{cases} p_k = \mathrm{P}(n_k = 1) = 1 - \mathrm{P}(n_k = 2) = \frac{1}{1+\exp(-z_k)} \\ z_k = z_{k-1} + w_k \\ w_k \sim \mathcal{N}(0, \eta_k) \\ \eta_k \sim \text{Inverse-Gamma}(a_0, b_0) \end{cases} \tag{3}$$

where $\{a_0, b_0\}$ are the hyperparameters of the prior on η_k.

The parameters of the state-space model in each instance of the sliding window are $\boldsymbol{\Omega} = \left\{z_{1:K_A}, \eta_{1:K_A}, \rho^{(a)}, \mu^{(a)}, \rho^{(u)}, \mu^{(u)}\right\}$, which have to be inferred from the observed features $m_{1:K_A}^{(1)}$ and $m_{1:K_A}^{(2)}$. We employ the Expectation Maximization (EM) framework of [4] to infer these parameters in each sliding window. To make our models suitable for real-time application, we initialize the parameters in each window by those estimated in the previous one. In addition, the closed-form update rules for the Log-Normal distribution parameters result in considerable computation savings. At each sliding window, $\hat{p}_{K_A - K_F}$ and its confidence intervals are reported as the final measure of attentional state with a delay of $K_F W$ samples.

III. RESULTS

In this section, we apply our real-time attention decoding framework to synthetic data and real EEG. We choose γ using cross-validation and set λ just large enough to ensure stable decoder estimates. We consider a sliding window of length 15 s (K_A) with a 1.5 s forward-lag (K_F). For initialization and tuning the conjugate prior hyperparameters in (2), we treat the first 15 s of each $\backsim$60 s trial as training data. The hyperparameters are tuned based on the fitted attended and unattended Log-Normal distributions in this phase, with large variances to make the conjugate priors non-informative. We choose a mean of 0.2 and variance of 5 for the Inverse-Gamma prior in (3) to prevent rapid fluctuations of the p_k's. To evaluate the performance and robustness of our proposed real-time estimator, we use the batch-mode estimator of [4] as a performance benchmark, in which the entire trial duration is used to decode the attentional state in an offline fashion.

A. Application to Simulated Data

Consider the following generative model:

$$e_t = w_t^{(1)}\left(s_t^{(1)} * h_t\right) + w_t^{(2)}\left(s_t^{(2)} * h_t\right) + u_t, \tag{4}$$

where e_t is the neural response, e.g., an EEG channel output, $s_t^{(1)}$ and $s_t^{(2)}$ are the speech envelopes at time t for speakers 1 and 2, respectively, h_t is the temporal response function, and

u_t is observation noise. As shown in Fig. 2 and motivated by existing work [5], h_t is modeled as a set of sparse lag components smoothed by a Gaussian kernel. The weight signals $w_t^{(1)}$ and $w_t^{(2)}$ respectively determine the contribution of speakers 1 and 2 to e_t and are modulated by the attentional state. We assume that speaker 1 (resp. 2) is attended at time t iff $w_t^{(1)} > w_t^{(2)}$ (resp. $w_t^{(1)} < w_t^{(2)}$). We consider two speech signals of length 60 s with $f_s = 200$ Hz, and the weight signals, shown in Fig. 3-A, are such that speaker 1 is attended in $[0, 30)$ s and speaker 2 is attended in $(30, 60]$ s. We consider $u_t \overset{\text{iid}}{\sim} \mathcal{N}(0.02, 2.5\times10^{-5})$ to simulate the neural response.

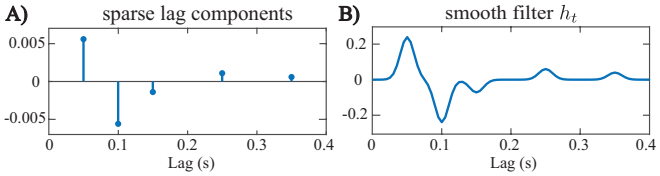

Fig. 2: Impulse response h_t used in Eq. (4): A) sparse lag components, B) the smoothed impulse response.

To estimate the decoder coefficients from $e_t, \ldots, e_{t+L_d}$ to $s_t^{(1)}$ or $s_t^{(2)}$, we set $W = 50$ (temporal resolution of 0.25 s), $K = 240$, $\lambda = 0.95$ (*effective* data length of 5 s), $\gamma = 0.001$, and $L_d = 80$ (decoder lag of 0.4 s). Thus, the overall delay in attention decoding is $\frac{K_F W + L_d}{f_s} = 1.9$ s. The upper panels in Fig. 3-B and 3-C show the two features discussed in Section II computed from the decoder estimates. These features are both attention-modulated since *on average*, the features of speaker 1 are larger in $[0, 30)$ s, and those of speaker 2 are larger in $(30, 60]$ s. However, they exhibit stochastic fluctuations which require further processing by the state-space model.

The lower panels in Fig. 3-B and 3-C show the estimated probabilities of attending to speaker 1 ($\hat{p}_k$) and their 90% confidence intervals for the correlation-based and ℓ_1-based features, respectively. When the entire confidence interval of $\hat{p}_k$ is higher (resp. lower) than the 0.5 level, we classify the k^{th} window as attending to speaker 1 (resp. 2), and when the confidence interval includes 0.5, we mark the window as undetermined. The state-space model translates the highly variable features into robust and statistically interpretable measures of the attentional state. As expected, the real-time estimator is more susceptible to the feature fluctuations than the batch-mode estimator as it only observes 1.5 s into the future features. This might lead to misclassifications as shown by red arrows in Fig. 3-B and 3-C. We have archived a MATLAB implementation of our method in Github which reproduces the results in Fig. 3 [16].

B. Application to Experimentally Recorded EEG

Experiment Details: We performed a constant-attention experiment with two male speakers, where the subjects were instructed to maintain their attention on speaker 1 in each trial. 64-channel EEG was recorded and digitized at 10 KHz. The study includes 3 normal-hearing subjects (mean age of 49.5 years with standard deviation of 7.18 years) and 24 trials of length $\sim$ 60 s each per subject. The stimuli included

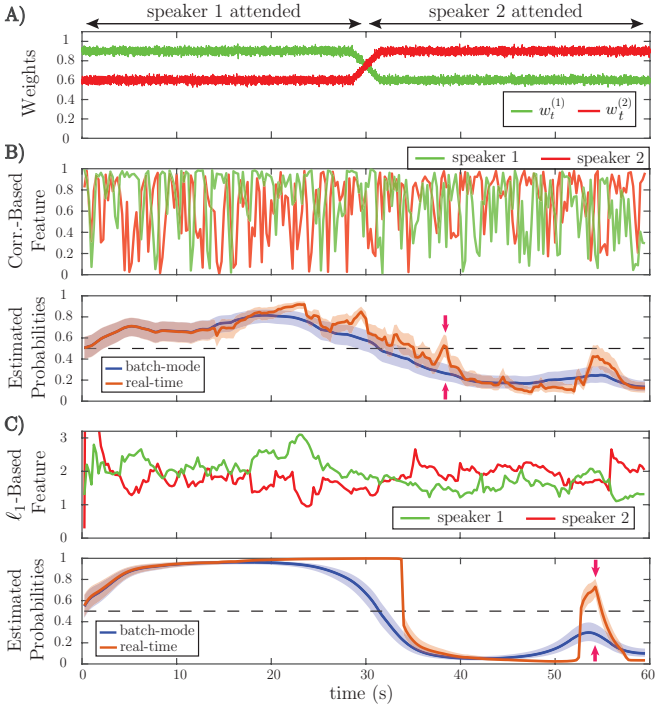

Fig. 3: Results for simulated EEG: A) Input weights $w_t^{(1)}$ and $w_t^{(2)}$ in Eq. (4). B) Outputs of the correlation-based feature and state-space model estimators. C) Outputs of the ℓ_1-based feature and state-space model estimators.

story segments recorded at Hafter Lab at UC Berkeley and were delivered to subjects in a dichotic fashion with adjusted and matched volumes at both ears. At the end of each trial, subjects answered two multiple-choice semantic questions about the attended story to maintain alertness. The procedures were approved by the Institutional Review Board.

Analysis Results: Speech envelopes and EEG were down-sampled to $f_s = 64$ Hz, and we considered the first 53 s of each trial as they had variable durations. To reduce the dimension of decoders for real-time implementation, we selected 28 frontal EEG channels, namely Fz, F1-8, FCz, FC1-6, FT7-10, C1-6, and T7-8. Based on [7], such a channel selection can yield a performance close to that of using all channels. For decoder estimation, we set $W = 16$ (temporal resolution of 0.25 s), $K = 212$, $L_d = 16$ (decoder lag of 0.25 s), $\gamma = 0.4$, and $\lambda = 0.975$ (*effective* data length of 10 s). Thus, the overall delay in attention decoding is 1.75 s. We used the ℓ_1-based feature which was considerably more attention-modulated in our dataset than the correlation-based feature.

Fig 4 shows the performance of the real-time and batch-mode estimators for three representative trials. Going from panel A to C, the attention modulation level of the ℓ_1-based features decreases, which results in more misclassifications. The attention modulation effect is expected to vary among trials depending on the performance of the subject, background neural activity, and effectiveness of our feature extraction. As in Fig. 3, we observe that the real-time estimator exhibits performance close to that of the batch-mode, while having sharper transitions due to the small 1.5 s forward-lag and less robustness to feature fluctuations.

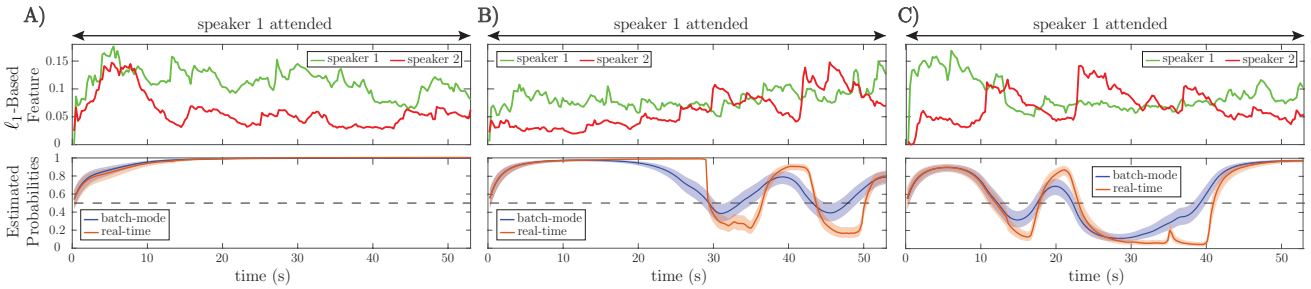

Fig. 4: The ℓ_1-based features and state-space estimators for three representative real EEG trials: A) The feature reliably separates the attended and unattended speakers. B) The feature separates the attended and unattended speakers *on average* over the trial. C) The feature either fails to separate the two speakers or results in a larger output for the unattended speaker on average, which is unexpected.

Fig. 5 summarizes the classification results for our EEG analysis. Fig. 5-A shows an illustration of the classification process for a trial using the state-space model output. In Fig. 5-B, each circle corresponds to a trial and shows the percentage of correctly classified vs. incorrectly classified windows (out of $K = 212$ windows). Fig. 5-C shows the classification results for each subject averaged over all the trials. Nearly 80% of the windows in each trial are classified correctly per subject. This is comparable to the state-of-the-art results of $\backsim 90\%$ accuracy in [3] for offline attention decoding at low temporal resolutions using large training datasets.

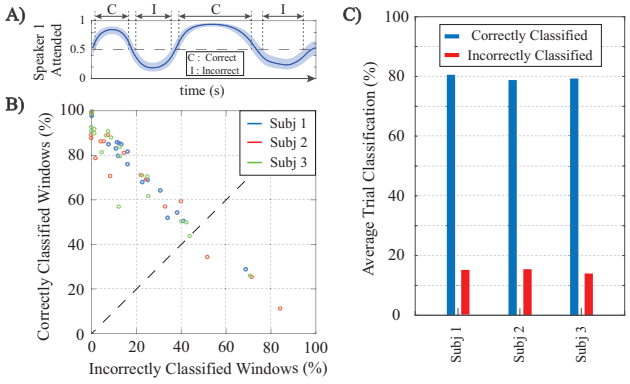

Fig. 5: Real-time classification results for real EEG data. A) generic illustration of the classification procedure. B) Classification results per trial for all subjects; each circle corresponds to a trial. C) Average classification performance over all trials for the three subjects.

IV. CONCLUSION

We proposed a new framework for real-time auditory attention decoding from EEG in a dual-speaker setting. We estimated the decoder coefficients in *real-time* at *high temporal resolution* via sparse adaptive filtering, which mitigates the need for large training datasets and offline estimation. We then extracted attention-modulated features using the estimated decoders. Finally, we employed a state-space model to correct for the stochastic fluctuations of the features and to obtain robust probabilistic measures of the attentional state at high temporal resolution. The required training data for parameter tuning in our model is minimal compared to existing methods. Application to synthetic and real EEG data revealed that our framework offers performance comparable to the state-of-the-art offline attention decoding algorithms that operate at low temporal resolution and use large training datasets to pre-estimate the decoder coefficients.

REFERENCES

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

[2] J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay, *The Auditory System at the Cocktail Party*. in the Springer Handbook of Auditory Research series, 2017.

[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," *Cereb. Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

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

[5] S. Akram, J. Z. Simon, and B. Babadi, "Dynamic estimation of the auditory temporal response function from meg in competing-speaker environments," *IEEE Trans. on Biomed. Eng.*, 2016.

[6] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Pro. Natl. Acad. Sci.*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

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

[8] 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*, no. accepted, 2016.

[9] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," *J. Neur. Eng.*, vol. 14, no. 5, 2017.

[10] 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," *IEEE Trans. on Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, 2017.

[11] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *J. Neur. Eng.*, vol. 13, no. 5, p. 056014, 2016.

[12] R. Zink, S. Proesmans, A. Bertrand, S. Van Huffel, and M. De Vos, "Online detection of auditory attention with mobile eeg: closing the loop with neurofeedback," *bioRxiv*, p. 218727, 2017.

[13] A. Sheikhattar, J. B. Fritz, S. A. Shamma, and B. Babadi, "Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis," *IEEE Trans. on Sig. Proc.*, vol. 64, no. 8, pp. 2026–2039, 2015.

[14] T. Goldstein, C. Studer, and R. Baraniuk, "A field guide to forward-backward splitting with a FASTA implementation," *arXiv eprint*, vol. abs/1411.3406, 2014. [Online]. Available: http://arxiv.org/abs/1411.3406

[15] 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.

[16] S. Miran, *MATLAB Implementation*. https://github.com/sinamiran/Real-Time-Decoding-of-Auditory-Attention-via-Bayesian-Filtering, 2018.