

SAGA: Sparse Adversarial Attack on EEG-Based Brain Computer Interface



Boyuan Feng, Yuke Wang, Yufei Ding
University of California, Santa Barbara

Introduction

We study sparse adversarial attack on EEG-based BCI.

Our assumption:

- Weak capability that only sparsely attacks a small portion of channels and time steps
- E.g., 3 out of 64 channels fail at a small portion of time steps

Our key contributions:

- Adaptive mask on the time dimension and the channel dimension to uniformly encode diverse sparsity constraints
- PGD-based iterative solver to effectively generate adversarial examples while satisfying sparsity constraints

Adaptive Mask

- Channel Mask: $\theta \in [0, 1]^C$
- Temporal Mask: $M \in [0, 1]^T$
- Sparsity cardinality constraints:
 - $\text{card}(\theta) \leq \epsilon_\theta$
 - $\text{card}(M) \leq \epsilon_M$
- Attack Formulation

$$\min -\ell(X + \text{diag}(\theta) \cdot X_\delta \cdot \text{diag}(M), Y)$$

$$\text{s.t. } \text{card}(\theta) \leq \epsilon_\theta, \text{card}(M) \leq \epsilon_M$$

Iterative Attack

- $S_1 = \{\theta \in [0, 1]^C \mid \text{card}(\theta) \leq \epsilon_\theta\}$
- $S_2 = \{M \in [0, 1]^T \mid \text{card}(M) \leq \epsilon_M\}$

Algorithm 1: Iterative Attack to solve Problem 6.

- 1 **Input:** Given X , fixed weight W , learning rate η_k , and iteration number K
- 2 Randomly initialize $\theta^{(1)}$, $M^{(1)}$, and $X_\delta^{(1)}$.
- 3 **for** $k = 1, 2, \dots, K$ **do**
- 4 **Channel Selection on θ :**
- 5 $\theta^{(k+1)} = \Pi_{S_1}[\theta^{(k)} + \eta_k g_\theta^{(k)}]$ with Eq 8.
- 6 **Frame Selection on M :**
- 7 $M^{(k+1)} = \Pi_{S_2}[M^{(k)} + \eta_k g_M^{(k)}]$ with Eq 9.
- 8 **Perturbation Generation on X_δ :**
- 9 $X_\delta^{(k+1)} = X_\delta^{(k)} + \eta_k g_X^{(k)}$ with Eq 10.
- 10 **end**
- 11 Return channel mask $\theta^{(K)}$, frame mask $M^{(K)}$, and feature perturbation $X_\delta^{(K)}$.

SAGA Performance

Table 1. Overall Performance under 5% Sparsity Constraints

Dataset	Method	EEGNet		DeepConv	
		Acc. (%)	Norm	Acc. (%)	Norm
SPM	Clean	94.12	-	85.29	-
	FGSM	70.59	112.9	61.76	112.9
	iFGSM	70.59	112.8	58.82	112.7
	SAGA	0	112.1	0	111.2
MI	Clean	99.4	-	66.67	-
	FGSM	88.89	164.3	66.67	164.3
	iFGSM	88.89	161.3	55.56	133.3
	SAGA	11.11	155.2	11.11	132.3
ERP	Clean	91.67	-	72.22	-
	FGSM	54.17	1.94	63.89	1.94
	iFGSM	54.17	1.89	63.89	1.89
	SAGA	16.67	1.86	8.33	1.84

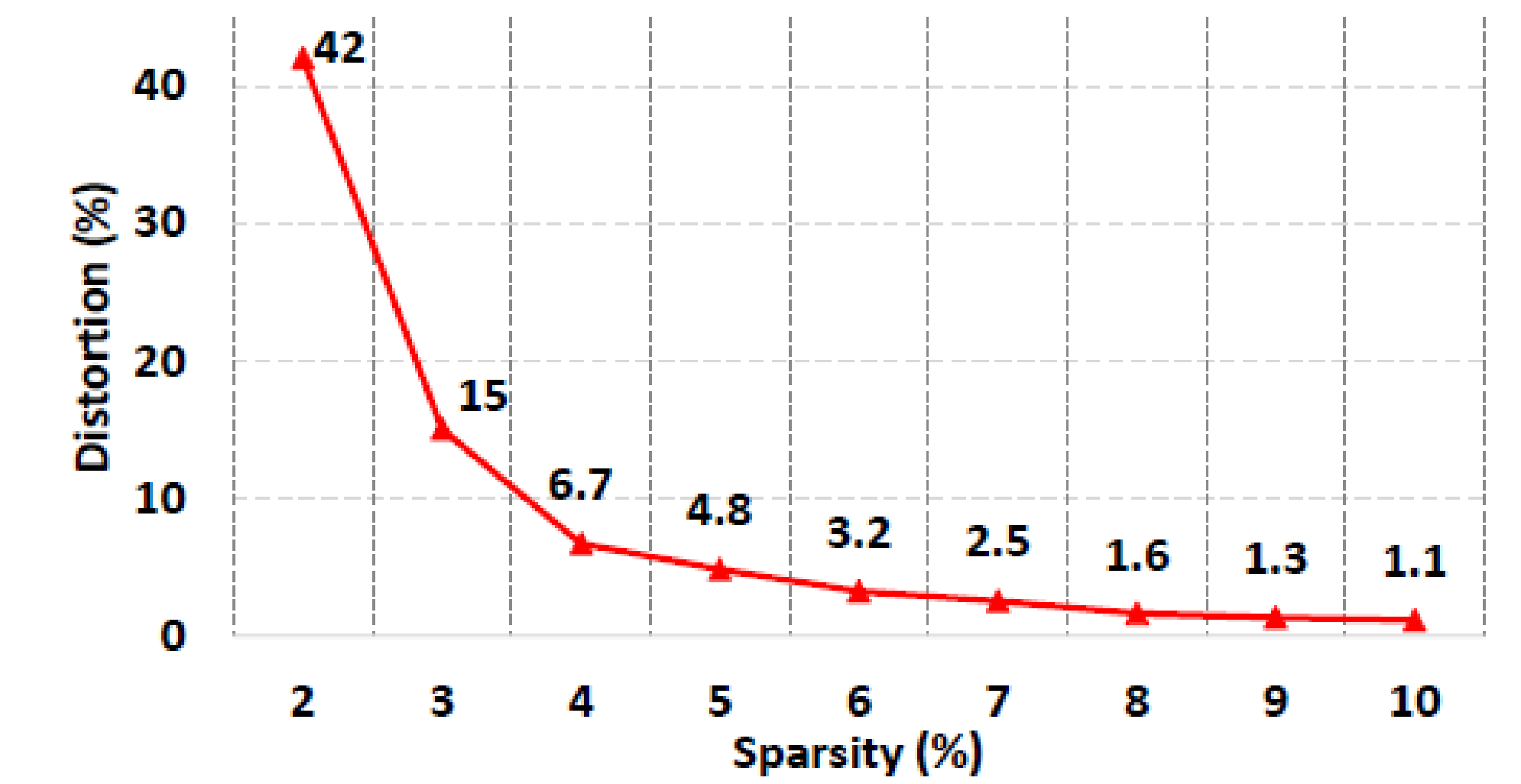


Figure 2. Minimum distortion under diverse sparsity to fool EEGNet on SPM with 0% accuracy.

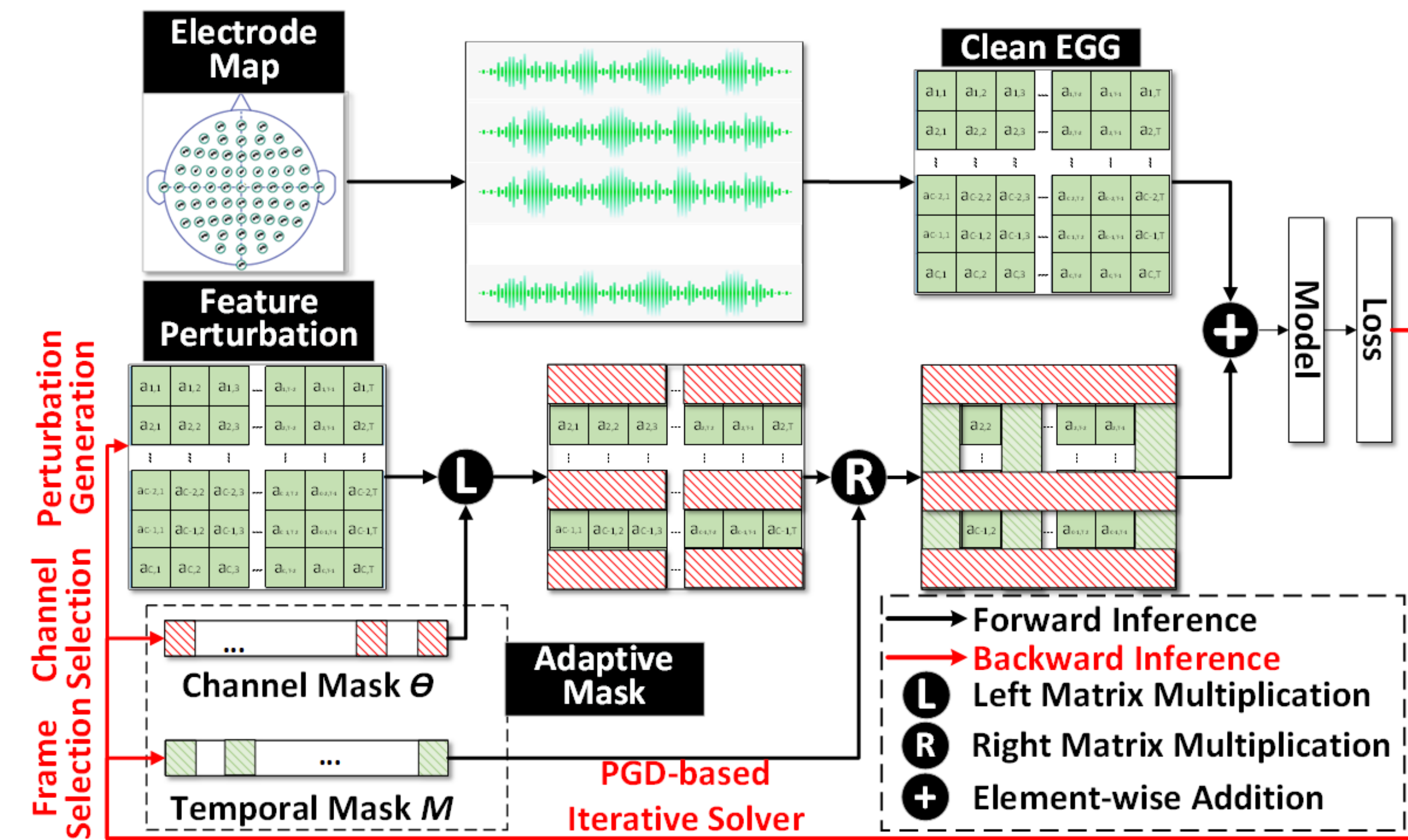


Figure 1. Overview of SAGA Design

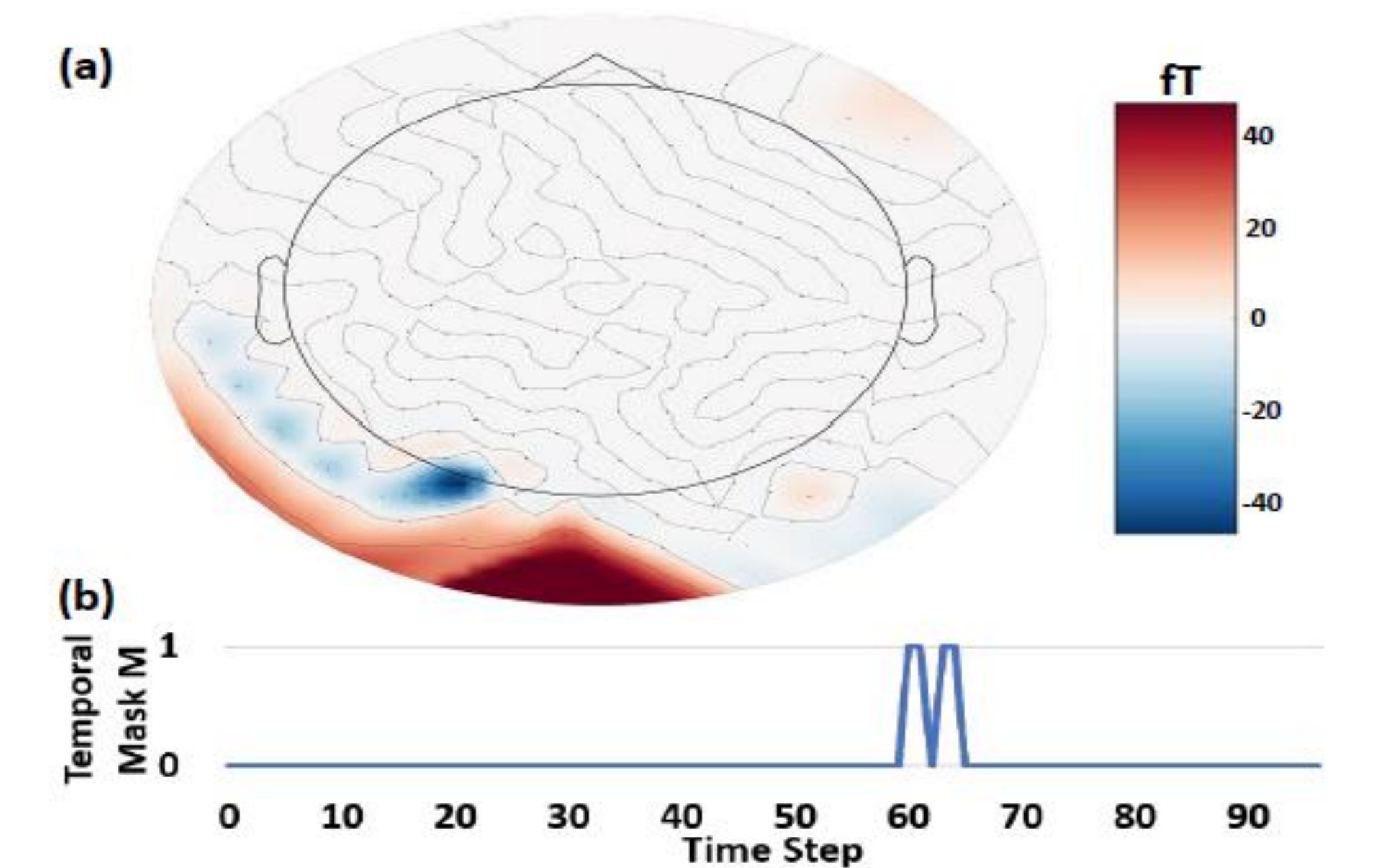


Figure 3. Visualization of selected masks
(a) Channel Mask; (b) Temporal Mask