Estimation of Heard Speech Spectrograms Using Neural Responses and Trained STRFs

E6886 Course Project Proposal

Vinay Raghavan and Aditya Sinha

We will estimate the input spectrograms of heard speech from the neural responses recorded while a subject listened to the speech. Reconstruction of the heard speech spectrogram from the neural data is relevant in areas like speech decoding and attention decoding. For this problem, we receive as measurements (y) the neural responses across time (t) for the (n) electrodes used for recording. The input (x) is a spectrogram with (f) frequency bins across time t. Note that this is sparse both across time and frequency, especially if we use basic thresholding. The sensing matrix (A) for each electrode is a Spectro-Temporal Receptive Field (STRF), which models the response of auditory neurons in the forward encoding path for a given speech segment. The neural response for an electrode is obtained as a 2D convolution of the STRF kernel (f x t_l, where t_l is the width of the kernel in time) with the speech spectrogram. This convolution takes strides only across the time axis, and not the frequency axis, making it's modeling as a matrix multiplication problem readily tractable.
We have,

$$y_i = A_i * x, \quad i = 1, 2, ..., n$$

where $*$ represents convolution

We model this as,

$$y = Ax$$

in the manner depicted on the next page.

We will employ an L1 minimization approach, possibly using the subgradient method to find the sparse input. We want the reconstructed spectrogram to retain the slow and intermediate temporal fluctuations of the true heard speech spectrogram, such as those corresponding to syllable rate. We also hope to retain other features, such as harmonics and formant frequencies.

References:

[1] Aertsen, A. M. H. J., & Johannesma, P. I. M. (1981). The spectro-temporal receptive field. *Biological cybernetics*, 42(2), 133-143.

[2] Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., ... & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1), e1001251.

$y = Ax$

$y \in \mathbb{R}^{n \times t}$

$A \in \mathbb{R}^{n \times (f t_\ell)}$

$x \in \mathbb{R}^{(f t_\ell) \times t}$

$n \rightarrow$ #electrodes

$t \rightarrow$ time support

$t_\ell \rightarrow$ support of STRF

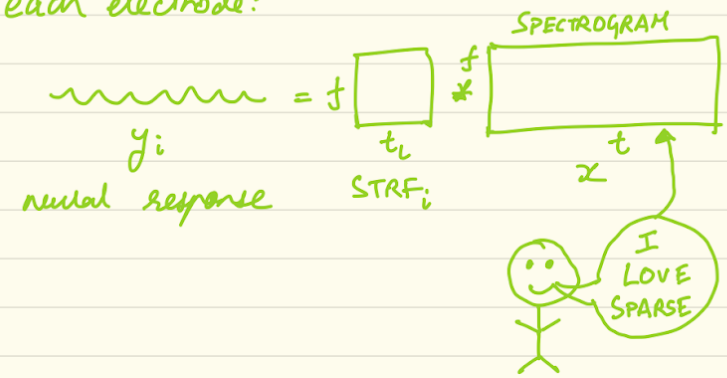
$f \rightarrow$ frequency bins

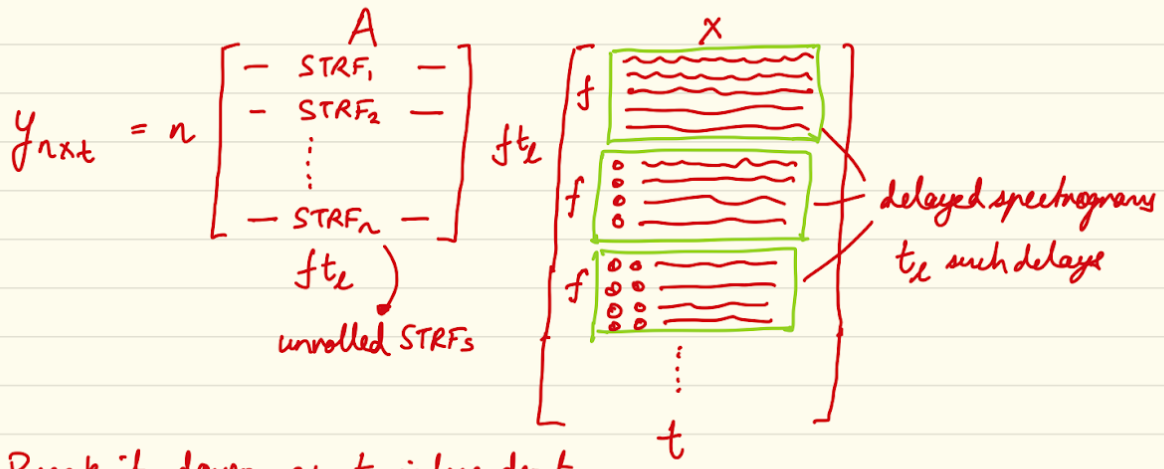
$y \rightarrow$ neural responses of n electrodes

$x \rightarrow$ intelligently arranged input.

$A \rightarrow$ intelligently arranged STRFs

each electrode:





Break it down as t independent sparse problems:

$P_1, P_2 \ldots\ldots P_t \in P$

know how to solve!!

$P_i : \quad y_{n \times 1} = A_{n \times (f t_\ell)} \; x_{f \times 1}$