# Sparse Reconstruction of Heard Speech Spectrograms from EEG

Vinay Raghavan*
*Department of Electrical Engineering*
*Mortimer B. Zuckerman Mind Brain Behavior Institute*
*Columbia University*
New York, NY, USA
vsr2119@columbia.edu

Aditya Sinha*
*Department of Electrical Engineering*
*Columbia University*
New York, NY, USA
as5624@columbia.edu

***Abstract*—Reconstruction of heard speech spectrograms from neural data is important in areas like auditory attention decoding and communicating with locked-in patients. For this problem, we receive the neural responses from multiple channels of electroencephalogram (EEG) recordings over time as measurements and use pre-learned spectro-temporal receptive fields (STRFs) as the sensing matrix. From this data, we perform a sparse reconstruction of the original heard speech power mel-spectrogram, which outperforms dense reconstruction due to the inherent sparsity in speech spectrograms. We show that the projected subgradient method produces the most accurate reconstructions, as measured by the error in the objective function for a single time frame, the l2-error of a single time frame, and the Frobenius norm of the error in the final reconstruction.**

***Index Terms*—stimulus reconstruction, EEG, spectrogram, STRF, sparsity**

## I. Introduction

The early auditory system is responsible for basic processing of speech and other sounds, including a decomposition of the heard sound into basic time-frequency representations, before higher level phonetic and lexical processing occurs [1]. This initial decomposition yields a faithful representation of the spectro-temporal properties of the perceived audio expressed in auditory cortex [2]. It has also been shown that it is possible to reconstruct a time-frequency representation of heard speech from neural responses, especially ones that are recorded from areas in and around auditory cortex, with improved results from invasive electrocorticography (ECoG) recordings compared to non-invasive electroencephalography (EEG) recordings due to the improved signal-to-noise ratio [3]. Reconstructing speech from auditory cortex is a fundamental problem in the field of auditory attention decoding and has further clinical applications in patients with locked-in syndrome who are otherwise unable to communicate with the outside world. Typically, heard speech reconstruction is performed with the objective of minimizing the mean-square error between the original spectrogram and the recovered spectrogram [4]. However, based on our knowledge about the basic structural properties of speech spectrograms, we can expect that each time frame of a speech spectrogram will be sparse over the frequency bins [5] [6]. Therefore, we can take advantage of this inherent structure of speech spectrograms by utilizing methods that perform sparse reconstruction of speech spectrograms to improve the quality of our results.

## II. Technical Approach

### A. Data Collection

Speech from two co-located talkers, one male and one female, was presented to the subjects in a quiet, electrically-shielded, audiometric booth. The audio was presented from a single loudspeaker directly in front of the subject, with each trial lasting approximately 1 minute. The stimuli consisted of passages from the Connected Speech Test (CST) database, used for its ability to provide an objective measure of intelligibility. Each passage was heard only once by the subject.

The data used here were collected from a single human subject (age 23, male, right-handed). The subject gave written, informed consent to participate in the experiment, in a protocol approved by the Columbia University Human Research Protection Office and Institutional Review Board. The instrumentation used for collection included a wet-electrode, 64-channel g.tec EEG cap with a sampling rate of 2400 Hz, and the data were recorded using the g.Recorder data acquisition software. Reference electrodes were placed on both ears. Prior to analysis, all EEG data were down-sampled to 100 Hz, including the application of an anti-aliasing low-pass filter with a cutoff frequency of 50 Hz. The EEG data were further low-passed filtered with a cutoff frequency of 8 Hz before reconstruction was attempted because of prior work showing that frequencies below 8 Hz are linearly related to the stimulus power [2]. The auditory stimuli were converted to power mel-spectrograms using the LibROSA package for Python, which performed a short-time Fourier transform with a window size of 2048 samples and a hop length of 240 samples (10 ms) to line up with the EEG data down-sampled to 100 Hz [7].

### B. Data Properties and Assumptions

Based on our knowledge of the cortical tracking of the time-frequency representation of heard audio, we can utilize the specific properties of the time-frequency representation of speech to improve upon existing methods of heard speech

*Both authors contributed equally to this manuscript reconstruction. Most importantly, the time-frequency representation of speech is sparse over frequency bands due to the harmonic properties of the vocal chord and is also sparse over time, but only for specific frequency bands which can further depend on the speaker [5] [6]. For this reason, we focus on the inherent and reliable sparsity over frequency. Furthermore, an important assumption we make is that the heard speech spectrogram is linearly related to the neural responses. This is certainly not a reliable assumption if the objective is to perfectly reconstruct the heard speech spectrogram; however, linear approximations of the heard speech spectrogram from neural recordings have been shown to be sufficient for a variety of tasks for which stimulus reconstruction is useful, including auditory attention decoding [8] [9]. Therefore, we hypothesize that our results will be a more faithful reconstruction of the heard speech spectrogram; however, we do not expect our results to produce intelligible speech.

### *C. Problem Setup*

For this problem, we receive the neural responses from $N = 64$ electrode channels of EEG recordings over time, $T$, as measurements, $Y \in \mathbb{R}^{N \times T}$. From these measurements, we aim to reconstruct the power mel-spectrogram, $S_0 \in \mathbb{R}^{F \times T}$, which consists of $T$ column vectors $s_t \in \mathbb{R}^{F \times 1}$ where $F = 64$ for the frequency bands in the spectrogram. The spectrogram is mapped to each neural response channel, $n$, through a 1D-convolution over time with a spectro-temporal receptive field (STRF), $A_n \in \mathbb{R}^{F \times (\tau+1)}$ where $\tau$ is the number of time-lags over which the convolution is performed [10]. There are $\tau + 1$ columns in the STRF to account for the zero-lag column. The time-lags, $\tau$, are a chosen parameter of the model and typically go up to 170-250 ms for reconstructions from basic stimulus encoding in primary auditory cortex [4] [8]. We chose $\tau = 350$ ms to make sure we capture more than enough information. Therefore, to map the spectrogram to all $N$ channels of the EEG responses, $N$ STRFs are needed.

$$y_n = A_n * S_0,\ n = 1, 2, ..., N \tag{1}$$

In order to fit this problem into sparse recovery frameworks, we reformulate the convolution of each 2D STRF, $A_n$ with the spectrogram, $S_0$, as a matrix multiplication by flattening each STRF into row vectors, $a_n \in \mathbb{R}^{1 \times F(\tau+1)}$, and augmenting the spectrogram with lagged versions of itself, $S_t \in \mathbb{R}^{F \times T}$, to form a lagged-spectrogram matrix, $X \in \mathbb{R}^{F(\tau+1) \times T}$.

$$S_t = \begin{bmatrix} | & | & & | & | & & | \\ 0 & 0 & \dots & s_1 & s_2 & \dots & s_{T-t} \\ | & | & & | & | & & | \end{bmatrix}$$

$$A = \begin{bmatrix} — & a_1 & — \\ — & a_2 & — \\ & \vdots & \\ — & a_n & — \end{bmatrix},\ X = \begin{bmatrix} — & S_0 & — \\ — & S_1 & — \\ & \vdots & \\ — & S_\tau & — \end{bmatrix}$$

The problem formulation can subsequently be treated as a simple matrix multiplication:

$$Y = AX \tag{2}$$

Following this reformulation, we treat each time frame of the neural responses, $y_t \in \mathbb{R}^{N \times 1}$, and each time frame of the lagged-spectrogram matrix, $x_t \in \mathbb{R}^{F(\tau+1) \times 1}$ as measurements and reconstructions respectively, independent from all other time frames, such that

$$y_t = Ax_t,\ t = 1, 2, ..., T \tag{3}$$

This permits the straightforward use of the following efficient sparse recovery algorithms.

### *D. Algorithms*

We utilized four algorithms for performing sparse recovery of the heard speech spectrogram: the projected subgradient method, the accelerated proximal gradient method, the augmented Lagrangian method, and the Frank-Wolfe/conditional gradient method. In this section, the time frame information of each measurement and reconstruction is omitted in favor of iteration information. In general, these algorithms are applied in the same way to each time frame.

*1) Projected Subgradient Method:* The projected subgradient method is a constrained version of the subgradient method. It involves alternating between taking subgradient steps, which move in the direction $-\text{sign}(\hat{x})$ for an $\ell^1$ norm objective, and orthogonal projections onto the feasible set $\{\hat{x} | A\hat{x} = y\}$ defined by the constraint. Therefore, each projected subgradient step is defined as

$$\hat{x}_{k+1} = \mathcal{P}_C[\hat{x}_k - t_k g_k],\ g_k \in \partial \| \cdot \|_1(\hat{x}_k) \tag{4}$$

where $t_k$ is the step size at iteration $k$ and $g_k$ is the subgradient of the $\ell^1$ norm of $\hat{x}$ at iteration $k$.

*2) Accelerated Proximal Gradient Method:* The accelerated proximal gradient method is an unconstrained method that involves the same objective as the proximal gradient method with a modification to accelerate convergence. Specifically, for this problem, we seek to find

$$\underset{\hat{x}}{\text{argmin}}\ \frac{1}{2}\|A\hat{x} - y\|_2^2 + \lambda \|\hat{x}\|_1$$

which normally involves taking a gradient step to minimize the $\ell^2$ norm term followed by a soft-threshold as the proximal operator of the $\ell^1$ norm of $\hat{x}$. We can speed up convergence to its theoretical maximum by using the following updates until $\hat{x}$ converges:

$$w_{k+1} = p_k - \frac{1}{L}A^*(Ap_k - y) \tag{5}$$

$$\hat{x}_{k+1} = \text{soft}(w_{k+1}, \lambda/L) \tag{6}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2},\ \beta_{k+1} = \frac{t_k - 1}{t_{k+1}} \tag{7}$$

$$p_{k+1} = \hat{x}_{k+1} + \beta_{k+1}(\hat{x}_{k+1} - \hat{x}_k) \tag{8}$$

where $L$ is the Lipschitz constant which is computed as the largest eigenvalue of $A^*A$.

*3) Augmented Lagrangian Method:* The augmented Lagrangian method is an unconstrained optimization method that is similar to penalty methods but adds an additional penalty term designed to mimic a Lagrange multiplier. Here we seek to minimize the $\ell^1$ norm of $\hat{x}$ subject to the constraint $A\hat{x} = y$. Rather than enforcing the constraint, we add the penalty term $\frac{\mu}{2}\|A\hat{x} - y\|_2^2$ and the Lagrangian penalty $\langle \lambda, A\hat{x} - y\rangle$ so our overall objective is

$$\mathcal{L}_\mu(\hat{x}, \lambda) = \|\hat{x}\|_1 + \frac{\mu}{2}\|A\hat{x} - y\|_2^2 + \langle \lambda, A\hat{x} - y\rangle. \tag{9}$$

Therefore, the optimization strategy involves first taking a proximal gradient step with respect to $\mathcal{L}_\mu(\hat{x}, \lambda)$ followed by a subgradient step of the dual function, yielding

$$\hat{x}_{k+1} \in \underset{\hat{x}}{\operatorname{argmin}}\ \mathcal{L}_\mu(\hat{x}, \lambda) \tag{10}$$

$$\lambda_{k+1} = \lambda_k + \mu(A\hat{x}_{k+1} - y) \tag{11}$$

*4) Frank-Wolfe Method:* The Frank-Wolfe method, also known as the conditional gradient algorithm, is a constrained optimization method that is scalable enough to solve extremely large sparse problems due to the property that each iteration solves a subproblem that is simpler and easier to compute than the proximal operator. Here we seek to do the following

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|A\hat{x} - y\|_2^2 \\ \text{subject to} \quad & \|\hat{x}\|_1 \leq \tau \end{aligned} \tag{12}$$

Where $\tau$ is the maximum size of $\ell^1$ ball on which $\hat{x}$ exists. We can find the $\hat{x}$ that satisfies this by performing the following updates until $\hat{x}$ converges:

$$r_k = A\hat{x}_k - y \tag{13}$$

$$i_k = \arg\max_i |a_i^* r_k| \tag{14}$$

$$\sigma = \operatorname{sign}(a_{i_k}^* r_k) \tag{15}$$

$$v_k = -\tau\sigma e_{i_k} \tag{16}$$

$$\hat{x}_{k+1} = \frac{k}{k+2}\hat{x}_k + \frac{2}{2+k}v_k \tag{17}$$

Where $a_i$ is the $i^{\text{th}}$ column of $A$ and $e_i \in \mathbb{R}^F$ is the standard basis vector.

## III. Experiments

### A. Methods

We have modeled our problem formulation as that of sparse reconstruction. We reconstruct $X$ by performing a column-wise reconstruction using the columns of $Y$. This boils down the problem into $T$ independent subproblems, where each subproblem reconstructs a single time step.

After obtaining the data, we run the sparse reconstruction algorithms mentioned in Section II.D for a single time step, tune the respective parameters, and compare the results. Following this, we run the best algorithm (w.r.t. recovery error) on a range of time steps and further fine-tune to improve this error. Finally, we run this tuned algorithm over all time steps to obtain the reconstructed spectrogram, and compare it visually and quantitatively with the original spectrogram. Note that in this sparse reconstruction, due to the intractability of enforcing non-negativity of reconstructed values (power spectrogram amplitudes are positive), we have simply thresholded the reconstructions.

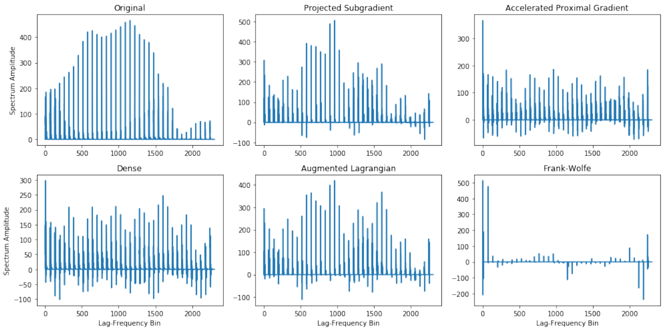

Fig. 1. Spectrum amplitudes for the original, dense reconstruction, and reconstructions from all sparse methods at t=4.00 sec.

### B. Results

Table I shows the error in objective and the recovery error for a single time step (t=4.00 sec) for all the algorithms mentioned. We also have dense reconstruction as a benchmark for comparison. The reconstructed columns are shown in Fig. 1. We see that overall, the projected subgradient method outperforms other sparse methods, and only the projected subgradient and augmented Lagrangian methods outperform dense reconstruction.

Table II shows the recovery error for the projected subgradient method and dense method over all time steps both before and after the reconstructed spectrogram has been realigned

TABLE I
Errors for each sparse recovery algorithm (@ t=400)

| | Sparse Recovery Algorithms | | | | |
|---|---|---|---|---|---|
| | **Proj Subgrad** | **APG** | **ALM** | **FW** | **Dense** |
| Objective Error | **2.32e-13** | 2.99e-2 | 2.01e-2 | 3.34e-1 | **1.55e-14** |
| Lagged Recovery Error | **1281.8** | 1649.8 | 1282.2 | 2267.7 | 1582.3 |

TABLE II
Final reconstruction errors (Frobenius Norm)

| | **Projected Subgradient** | **Dense** |
|---|---|---|
| Lagged Recovery Error | **74401.9** | 74511.6 |
| Delagged Recovery Error (mean) | **9748.0** | 10880.2 |
| Delagged Recovery Error (median) | 10456.5 | 11036.5 |
| Delagged Recovery Error (sparse) | 13793.9 | 10939.6 |

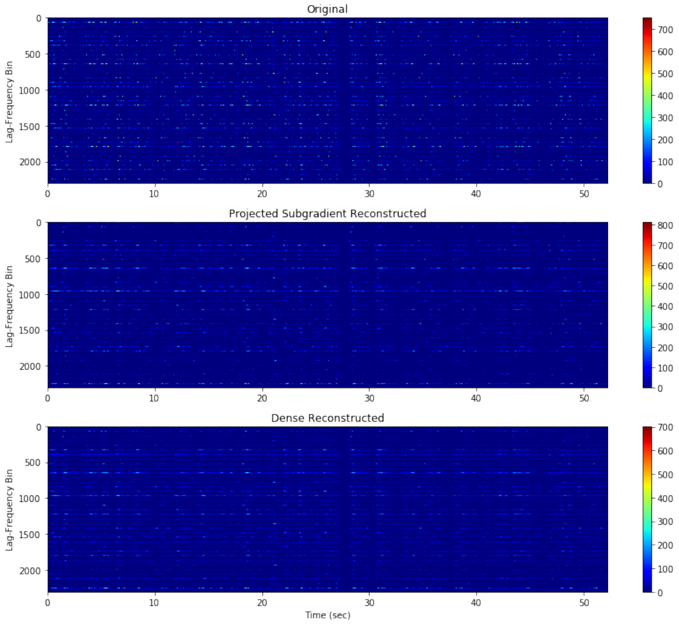

Fig. 2. Full lagged spectrograms from the original, the projected subgradient reconstruction, and the dense reconstruction.

in time and combined using three different methods: mean, median, sparse. The mean combination method involves taking the mean over all realigned spectrograms. The median combination method involves taking the median over all realigned spectrograms. The sparse combination method involves taking the mean over all realigned spectrograms and zeroing out any TF bins that were zero in any of the lagged spectrograms. Fig. 2 visualizes the recovered lagged spectrograms. Fig. 3 visualizes the recovered spectrograms delagged using the mean method. Fig. 4 visualizes the most relevant frequency bands recovered spectrograms delagged using the mean method. Finally, Fig. 5 visualizes a representative, five-second section of the recovered spectrograms delagged using the mean method for closer inspection and comparison.

## IV. Discussion

For a single time step, the projected subgradient method outperforms all the other methods. Although it does have a slightly larger objective error than dense reconstruction, it performs a much more accurate recovery, and the recovery error, which is a more important metric of evaluation, is smaller. In general, for our problem formulation, the objective error is more of a guide to monitor convergence of the algorithm, whereas the actual efficacy of each algorithm is judged by how closely the recovered spectrogram resembles the original one. This is most-clearly depicted by the fact that the dense reconstruction has a low objective error; however, the reconstructed column does not appear similar to the actual spectrogram column. In fact, this is one of our biggest motivations for exploring sparse methods in this problem setting.

The augmented Lagrangian Method comes as a close second, producing a similar recovery error, which is again much

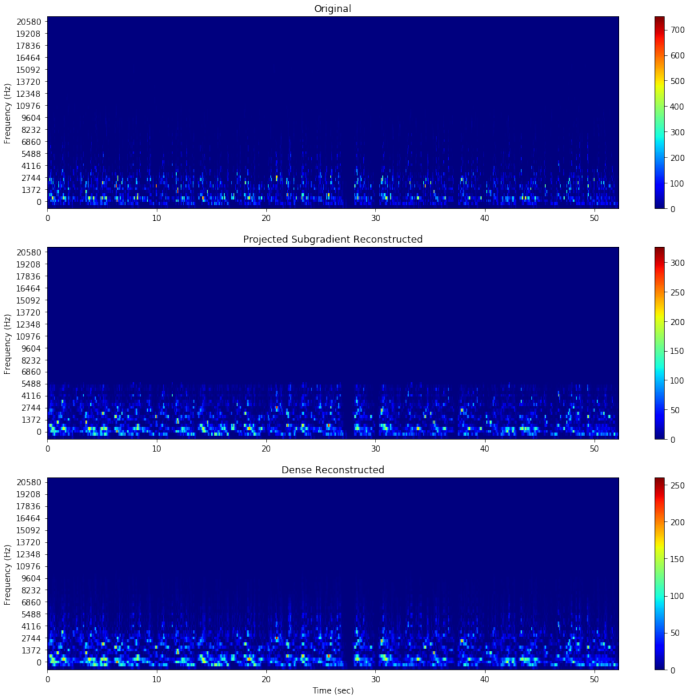

Fig. 3. Delagged spectrograms (mean method) from the original, the projected subgradient reconstruction, and the dense reconstruction.

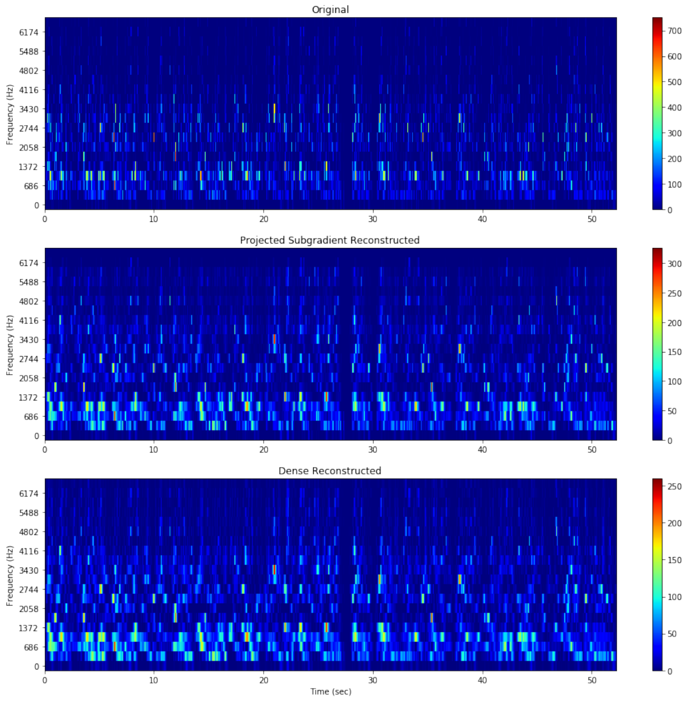

Fig. 4. Delagged spectrograms (mean method) from the original, the projected subgradient reconstruction, and the dense reconstruction in the frequency bands containing most of the power.

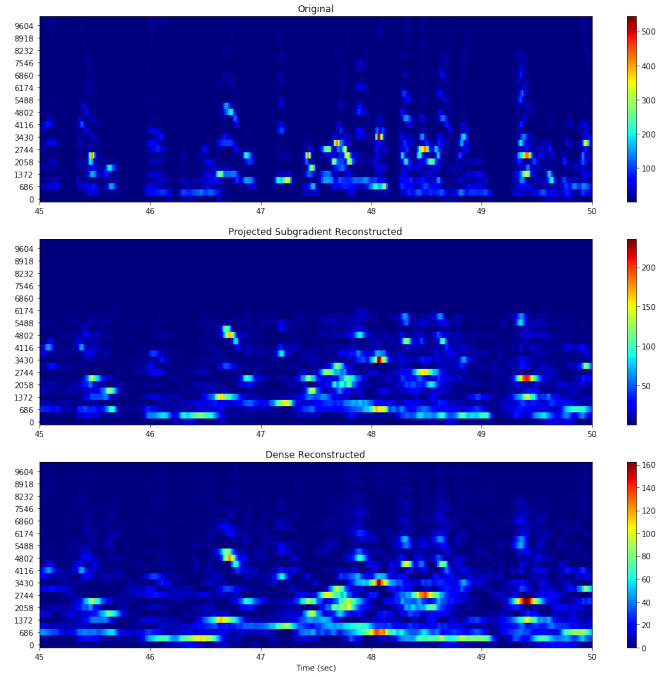

Fig. 5. A representative, five-second (t=45-50 sec) section of the original, projected subgradient reconstruction, and dense reconstruction spectrograms delagged using the mean method.

better than that of the dense reconstruction. However, it has a much worse objective value and takes longer to converge. The accelerated proximal gradient and Frank-Wolfe/conditional gradient methods perform relatively poorly; we do get a reconstruction somewhat similar to the dense one using APG, but the reconstruction using Frank-Wolfe is too sparse for it to be of any use. Even though the algorithms produce a decent value of the objective error, we reject these two methods.

After running the projected subgradient algorithm for the entire time duration, we see on comparison that a lot of the features of the original lagged spectrogram have been retained in the sparse reconstruction. We perform delagging and combination of the constituent spectrogram blocks using various methods (mean, median, sparse) and find out that combination using mean performs the best. The reasoning behind doing this is that data redundancy has an averaging effect on combination and yields better results than considering only the top block of rows of the lagged reconstruction. We see the reconstruction to be much closer to the original than the dense one. It does lose out on the higher frequency features however, since they have relatively less power and are zeroed out in a sparse reconstruction. However, that is a relatively benign drawback, since most of the information content in speech is present in the lower frequencies.

Through analysis of the five-second section of the spectrograms, seen in Fig. 5, we can interpret particular reasons why the projected subgradient reconstruction performs better than the dense reconstruction. First, we see that sparse reconstruction is better at modeling very short periods of silence, which are common in between words and in plosive phonemes. This can be seen in Fig. 5 towards the end of the 45th second. Furthermore, we see that sparse reconstruction is better at modeling fast changes in regions of the spectrogram with high power, as seen in Fig. 5 towards the end of the 47th second. In contrast, the dense reconstruction appears to smear these sections more, making them last longer than they should. Finally, we find that sparse reconstruction is somewhat better at matching the amplitudes of the spectrogram, as can be discerned from the range of values present in each spectrogram.

### *A. Future Directions*

We noted in Section III.A that none of our sparse reconstruction algorithms enforce non-negativity of the elements of reconstruction, and thus, we manually enforce it by thresholding the output. This might deal a severe blow to the objective and recovery errors, since our algorithms did no such effort to enforce non-negativity and sought to find an optimal sparse solution while allowing elements to be negative. This is the most significant shortcoming of the aforementioned methods, and the intractability of enforcing an extra constraint in these motivates us to look towards better methods.

Future work upon this would build on the idea of enforcing multiple constraints on the same variable of optimization. This can be done by using the Alternating Direction Method of Multipliers (ADMM) with the splitting trick, where we enforce the constraints in a sum form on different variables, and externally enforce the constraint for all these variables to be equal to each other and to $\hat{x}$. In fact, the ADMM framework can even be used to reconstruct the whole spectrogram at once, instead of dividing into subproblems for each time step. This can be done by considering each column to be an independent separable variable in ADMM, with similar constraints and updates for all of them. This would complicate the updates and each iteration would have number of updates $\sim 2t$ and might lower the efficiency and convergence rate, but sequential updates of the columns has the potential to give us a better recovery.

Apart from sparse reconstruction, we can consider a couple of alternative approaches to model this problem. For speech signals in general, the variation across frequencies is gradual for quite a few sections, and we also have a lot of zero rows, with little to no power in that band. With this information in mind, we can think of applying rank minimization to the same problem. This might give a better approximation of the high frequency features, since the sparsity constraint no longer drives them to zero. On the flip side, it might smear out the features across time, since it forces rows to be linear combinations of each other.

Looking at the problem as a direct linear convolutional map from the spectrogram to the neural response is another avenue worth exploring. The advantage it has over matrix mutiplication is that the spectrogram doesn't need to be lagged and has much fewer rows (by a factor of $\tau$), potentially resulting in a better reconstruction. This is however limited by the mathematical intractability of calculating the adjoint of the matrix convolution operator, which is required in all of the updates.

## Acknowledgment

We would like to thank Prof. John Wright for teaching us about the useful properties of sparsity and how to take advantage of those properties for problems like this. We would also like to thank Sam Buchanan for helping us fill all the gaps in our understanding, both throughout the semester and for this final project. Finally, V.R. would like to thank Prof. Nima Mesgarani for the support and resources needed to collect the data used for this project.

## References

[1] E. D. Young, Neural representation of spectral and temporal information in speech, Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 363, no. 1493, pp. 923945, Mar. 2008.

[2] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder, Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a Cocktail Party, Neuron, vol. 77, no. 5, pp. 980991, Mar. 2013.

[3] B. N. 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. e1001251, Jan. 2012.

[4] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex, Journal of Neurophysiology, vol. 102, no. 6, pp. 33293339, Dec. 2009.

[5] T. J. Gardner and M. O. Magnasco, Sparse time-frequency representations, Proceedings of the National Academy of Sciences, vol. 103, no. 16, pp. 60946099, Apr. 2006.

[6] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, Speech Enhancement Under Low SNR Conditions Via Noise Estimation Using Sparse and Low-Rank NMF with KullbackLeibler Divergence, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 7, pp. 12331242, Jul. 2015.

[7] B. McFee, M. McVicar, S. Balke, V. Lostanlen, C. Thom, C. Raffel, D. Lee, Kyungyun Lee, O. Nieto, F. Zalkow, D. Ellis, E. Battenberg, R. Yamamoto, J. Moore, Ziyao Wei, R. Bittner, Keunwoo Choi, Nullmightybofo, P. Friesch, Fabian-Robert Stter, , Thassilo, M. Vollrath, Siddhartha Kumar Golu, Nehz, S. Waloschek, , Seth, R. Naktinis, D. Repetto, C. Hawthorne, and CJ Carr, librosa/librosa: 0.6.3. Zenodo, 13-Feb-2019.

[8] J. A. OSullivan, 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, vol. 25, no. 7, pp. 16971706, Jan. 2014.

[9] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, A Tutorial on Auditory Attention Identification Methods, Frontiers in Neuroscience, vol. 13, Mar. 2019.

[10] A. M. H. J. Aertsen and P. I. M. Johannesma, The Spectro-Temporal Receptive Field, Biological Cybernetics, vol. 42, no. 2, pp. 133143, 1981.