TIME-FREQUENCY REGULARIZATION OVERLAPPING GROUP SHRINKAGE

Alex Epstein, Nimish Magre, Jared Miller

Electrical and Computer Engineering Department,
Northeastern University, Boston, MA, 02115

ABSTRACT

In this work, we denoise speech samples using predetermined structural knowledge of decomposable convex optimization problems. To this end, we exploit the grouping/clustering property observed with speech spectrograms to iteratively obtain a sparse clean speech signal using a mixed norm penalty term. We build upon the Overlapping Group Shrinkage (OGS) algorithm [1] and introduce time-frequency weights to the cost function to rid the sparse clean signal of the residual noise. These time-frequency weighting extensions are also empirically shown to effectively handle impulsive noise types. The time-frequency weights may be targeted to suppress specific noise types at desired time slices to further improve the performance of the algorithm.

Index Terms— Time-Frequency Weighted Regularization, L1 optimization, Group sparsity, Denoising, Speech enhancement, Convex optimization, Translation-invariant denoising

1. INTRODUCTION

Unsupervised speech denoising involves decomposing a noisy speech sample Y into its clean speech component X and the noise component N. The noise typically could arise from white, colored, musical, impulsive, or background processes. Prior work in this domain includes representing the speech component X as a sparse combination of signal coefficients (e.g. wavelet) with l_1-norm regularization [2, 3]. However, the l_1-*norm* penalty term implicitly assumes that the signal coefficients are independent, and therefore fails to effectively eliminate the residual noise from the sparse clean speech samples [4].

Speech signals also tend to display a clustering property showing inter/intra-scale clustering (i.e significant values of X tend to not be isolated), in a similar manner to wavelet decompositions of physically-arising signals. This pre-known structural property is exploited using Overlapping Group Shrinkage (OGS) [1] to further improve the SNR results of the clean speech by using a group-sparsity promoting penalty term in the cost function. While the introduction of the $l_{1,2}$-*norm* penalty term helped obtain higher SNR scores in experiments, the SNR metric was not an accurate gauge for auditory intelligibility (especially in the white gaussian noise setting).

We extended OGS by adding a set of time-frequency weights to the cost function. The time weights are based on the average energy for all frequencies in every time-slice, and the frequency weights implement a lowpass behavior that limit the extracted signal to a typical speech frequency range. The introduction of these weights to the cost function helped produce improved impulsive noise suppression, and also allowed for the ability to handle multiple prior known noise types.

The contributions of this paper are:

- Energy based time weighted regularization to handle non-stationary noise types with different energy levels at different time steps.
- Signal based spectral weighted regularization to extract specific target signals or suppress specific noise types.
- Experiments for evaluating combinations of the above extensions on both flat full-spectrum noise (WGN) and impulsive noise (keyboard clicks).

2. PRELIMINARIES

2.1. Notation

- $*$ matrix multiplication
- $\odot$ element-wise multiplication (Hadamard)
- $H(z)$ filter coefficients in z domain
- $\mathcal{F}$ Fourier transform
- $||\cdot||_F^2$ Squared Frobenius norm
- $1..N$ Natural numbers between 1 and N

2.2. Prior Work

For applications involving real-world signals, such as classification, compression and synthesis, the wavelet domain has been a natural setting [5]. Often, each significant coefficient (e.g. wavelet, chirp, Short Term Fourier Transform (STFT) or a large signal amplitude) is treated as being independent of all others and these coefficients are modelled either as jointly Gaussian or as non-Gaussian but independent [6]. Another approach that utilizes these integral transformations is a denoising method that decomposes the noisy speech spectogram

into a low rank noise structure, a sparse clean signal, and a low-intensity residual noise (musical noise) [7, 8, 9]. The random residual noise structure assumed to follow independent and identically distributed zero-mean Gaussian distributions. However with real natural signals, neighbouring coefficients tend to have statistical dependency. This clustering property observed within typical speech spectrograms indicates that if a particular coefficient is large/small in amplitude, then the adjacent coefficients also tend follow similar behavior [10, 11]. Algorithms that utilize hard/soft thresholding functions to obtain the sparse solutions typically tend to utilize the l_1-*norm* to induce sparsity and overlook the grouping property [2, 3]. Instances of group-aware sparse decomposition and recovery methods methods include [1, 12, 13].

We will review the OGS [1] algorithm in the context of 2d data (STFT time-frequency content). The noisy reference y and optimization variable x are matrices of size $N_t \times N_f$. Elements of y (and x) are indexed by $y(i)$ for $i \in \mathcal{I} = (1..N_t) \times (1..N_f)$. A convolutional filter of size $K_t \times K_f$ is used to define overlapping groups $\mathcal{J}$. For each $j \in \mathcal{I}$ such that $j_1 \leq N_t - K_f$ and $j_2 \leq N_f - K_f$, we define the $K_t \times K_f$ matrix $[j] = (j_t + 1..j_t + K_t) \times (j_f..j_f + K_f)$ as a group in $\mathcal{J}$. Th $K_t \times K_f$ submatrix of y indexed at $[j]$ is denoted by $y[j]$.

A regularized decomposition problem to obtain the sparse clean signal $x(i)$ from the noisy observations $y(i)$ is,

$$x^* = \arg\min_x F(x) = \frac{1}{2}||y - x||_F^2 + \lambda R(x) \tag{1}$$

where the penalty function $R(x)$ is chosen to promote the group sparsity behaviour of signal x[4]. The OGS algorithm minimizes the cost function represented by Equation 1 with the non-separable sum-of-norms penalty term [14, 4]:

$$R(x) = \sum_{j \in \mathcal{J}} ||x(j)||_F \tag{2}$$

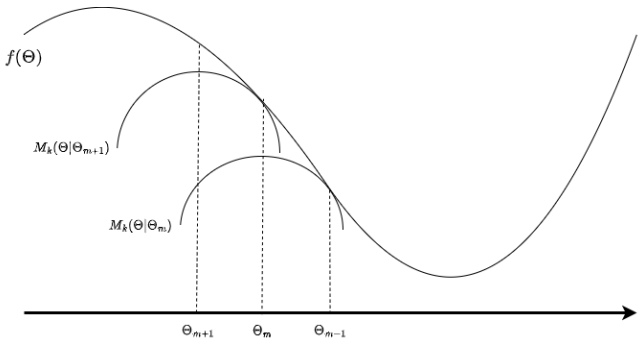

Fig. 1. Minorize-Maximize method to local method to maximize the objective function [15, 16] using a quadratic surrogate function.

The OGS algorithm makes use of the Majorization-Minimization (MM) [17] method wherein Equation 1 is solved via the introduction of separable surrogate function. Given a value x, the index set $\mathcal{I}' \subseteq \mathcal{I}$ is the set of indices i such that $x(i)$ is nonzero. The regularization function in Equation 2 can be decomposed into terms with per-index penalties $r(\cdot)$,

$$r(x(i)) = \sum_{[j] \in \mathcal{J} \ | \ i \in [j]} (||x\{j\}||_F)^{-1}, \quad R(x) = \sum_{i \in \mathcal{I}'} r(x(i)). \tag{3}$$

OGS performs the recurrent relation at each iteration k,

$$x^{(k+1)}(i) = \arg\min_{(x \in \mathcal{C})} (y(i) - x)^2 + \lambda r(x^{(k)}(i))|x|^2. \tag{4}$$

The majorization occurs because the $x^{(k)}$ inside r is fixed in each iteration. The solution at each iteration k for (4) is,

$$x^{(k+1)}(i) = \begin{cases} \frac{y(i)}{1+\lambda r(x^{(k)}(i))}, & i \in \mathcal{I}', \\ 0 & otherwise. \end{cases} \tag{5}$$

Procedure 5 will proceed until convergence, which is usually specified as a maximal number of iterations. An important note in the OGS algorithm is that if x is initialized to zero, then all the consecutive updates will lead to an optimized x equal to zero due to the penalty term $r(i; x)$ being undefined. Sparsified values will asymptotically approach zero while following the rule (5).

3. METHOD

3.1. Cost Function Modifications

We introduce a weighting function $W(\cdot)$ that provides a set of penalties on the regularization term $R(x)$ from (2). The weights are computed based on the input (noisy) data y, resulting in the new cost expression,

$$x^* = \arg\min_x F(x) = \frac{1}{2}||y - x||_F^2 + \lambda \sum_{i \in \mathcal{I}'} W(i) r(x(i)), \tag{6}$$

and update rule,

$$x^{(k+1)}(i) = \begin{cases} \frac{y(i)}{1+\lambda W(i) r(x^{(k)}(i))}, & i \in \mathcal{I}', \\ 0 & otherwise. \end{cases} . \tag{7}$$

We form a set of time-weights $W_t(y)$ in each time window and frequency weights $W_f(y)$ in each frequency band of the STFT. The accumulated weight matrix $W(y)$ is a rank-1 matrix formed by the positive time-frequency weights,

$$W(y) = W_f W_t \tag{8}$$

The OGS algorithm in (5) is equivalent to the uniform weighting $W(y) = 1$. We now explain how the weights W_t and W_f are chosen.

3.1.1. Time Weighting

The time weights are computed as a function of the signal energy in each time window. The intuition here is that if the speech signal remains at a constant energy level, then shifts in the energy are attributable to the noise content. The time-weights are based on the energy content of the signal y.

The average power for each time slice across all the frequency bins is,

$$E_t(y) = \sum_{k=1}^{f_l} |y_{k*}|^2 \tag{9}$$

The time-slice energy normalized against the average energy in any one bin is,

$$E_r(y) = \frac{E_t(y)}{Avg(E_t(y))} \tag{10}$$

The resultant weights are,

$$W_t(y) = \frac{E_r(y)}{Max(E_r(y))} \tag{11}$$

These time weights can be precomputed given the signal y. The computational burden of calculating the time-weights scales linearly with the time horizon.

3.1.2. Frequency Weighting

We add frequency weights to implement a lowpass filter $H(z)$ with a cutoff frequency of 4 kHz, thus penalizing high-frequency components of the recovered solution x. The frequency weights W_f may be defined using the FFT of $H(z)$ as,

$$W_f = |\mathcal{F}[H(z)]| \tag{12}$$

The implemented filter's magnitude response can be seen in Figure 2 which was derived using Parks-McClellan's algorithm[18] for linear phase filter design. Knowledge about the noise process' frequency content can be used to design the filter $H(z)$, such as adapting for White, Colored, or Impulsive noise. The frequency weights are independent of the time horizon.

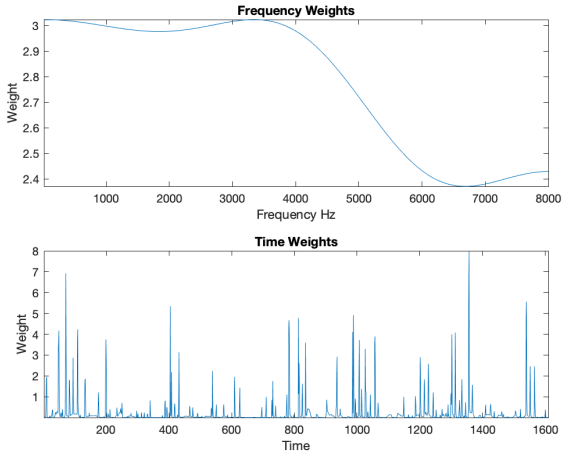

Fig. 2. Example frequency weighting with gain and smoothing applied along with time weighting.

3.2. Algorithm

The TFROGS scheme is described in Algorithm 1. All weights $W(y)$ in Equation 8 are positive, so the monotonic nonincreasing objective property of OGS is retained with the time-frequency weights. The termination criteria is based on the maximum number of iterations N_{it}, and follows prior work in OGS. Other convergence conditions could include absolute error, relative error, or small residuals in x. Brackets are used to indicate the dimensions of matrices and vectors.

Algorithm 1 The TFROGS Algorithm

input: $y \in \mathbb{R}, \lambda, H(z), J, N_{it}$

$x = y \qquad (initialization)$

$I' = \{i \in I, x(i) \neq 0\}$

$W_f = [|\mathcal{F}[H(z)]|]_{f_l \times 1}$

$E_t(y) = \left[\sum_{i=1}^{f_l} |y_{i*}|^2\right]_{1 \times t_l}$

$E_r(y) = [\frac{E_t(y)}{Avg(E_t(y))}]_{1 \times t_l}$

$W_t(y) = [\frac{E_r(y)}{Max(E_r(y))}]_{1 \times t_l}$

$W(y) = [W_f W_t]_{f_l \times t_l}$

repeat

$$r(x) = \begin{cases} \sum_{[j] \in \mathcal{J} \ | i \in [j]} (||x\{j\}||_F)^{-1} & i \in \mathcal{I}' \\ \infty & \text{else} \end{cases}$$

$x(i) = \frac{y(i)}{1 + \lambda W(i) r(x(i))}$

until N_{it} *iterations*

return: x

4. EXAMPLES AND RESULTS

We implemented Algorithm 1 using MATLAB code and audio samples publicly available[1]. We compare performance of OGS[1] and Algorithm 1 applied to denoising speech data with White Gaussian ($\mu = 0$ and $\sigma = 0.05$) or Impulsive (keyboard typing) noise. The clean speech sample involved an individual reading a set of Phonetically Balanced Sentences (Harvard sentences [19]). The appropriate noise processes were added to the clean speech. We utilized the SNR calculation function from [20], with the note that the SNR does not entirely capture speech intelligibility (this perceptive intelligibility is greatly degraded when impulsive noise is applied). [21, 22].

Table 1. Comparing different noise types for the OGS vs the TFROGS algorithm.

Noise Type	SNR (dB)	OGS SNR (dB)	TFROGS SNR (dB)
White Gaussian	-0.87	8.21	7.87
Impulsive	-7.42	-3.8	4.43

4.1. White Gaussian Noise

The WGN case has W_f equal to a matrix of ones, due to its flat frequency spectrum. While the TFROGS algorithm performs competitively to OGS in terms of the objective SNR metric for additive WGN, it does seem to provide a slight improvement in performance for intelligibility. With the generated WGN sample, since the noise and speech components typically would have equal power across time (noisy signal has SNR of $\sim$0 dB), the TFROGS algorithm more aggressively attenuates the speech signal in the absence of larger noise energy and vice versa as conveyed through Equation 11. Due to this energy similarity, time weighting does not provide a significant improvement in the objective SNR metric.

4.2. Impulsive Noise

In the case of impulsive noise the proposed time-frequency weights provide a significant boost of $\sim$10 dB to the objective SNR metric.

Figure 3 and Figure 4 provide a comparison between the last iterations of the original OGS and the modified TFROGS algorithms at the final iteration $k = N_{it}$. The top left panel plots values of the following intermediate quantity $||x^{(k)}[j]||_F : j \in \mathcal{J}$ (group norms). The top right panel is the group-based penalties $r(x(i))$ from (2). The bottom left panel displays the time-frequency weighted regularizer $W(i)r(x(i))$. The bottom right panel plots the spectrogram of the new iteration $\frac{y(i)}{1+\lambda W(i)r(x(i))}$.

The bottom-right corner of Figure 3 shows dark-blue vertical bands, corresponding to entire time slices being attenuated across all frequencies. In contrast, the time-frequency weightings in Figure 4 preserve the low frequency content at these stripes while reducing the high frequency noise in order to improve speech intelligibility.

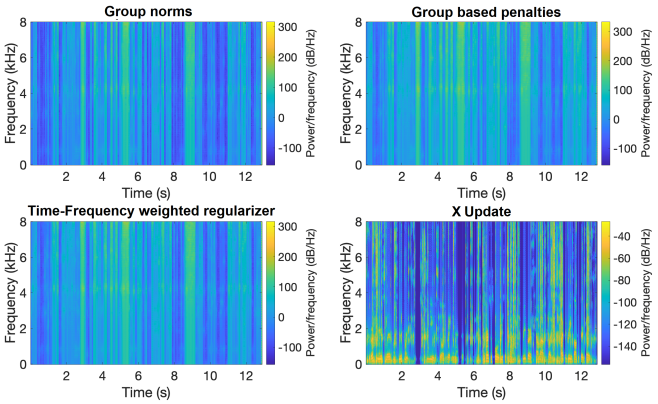

Fig. 3. Spectrograms of intermediate values from the last iteration of OGS for impulsive noise.

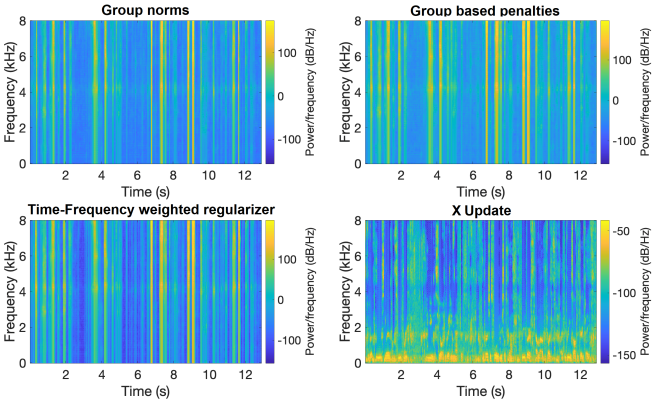

Fig. 4. Spectrograms of intermediate values from the last iteration of TFROGS for impulsive noise.

5. CONCLUSION

In this work we proposed an extension of OGS that adds time-frequency weights. The prior knowledge of the noise characteristics can be used to choose the weights and enhance speech intelligibility in speech processing. In experiments, adding time-frequency weights was competitive with standard (uniformly-weighted) OGS in the Gaussian case, and the intelligibility was greatly improved in the impulsive noise case. Future work includes extending the algorithm to the streaming setting with small time windows and low latency. Another aspect is generating adaptive parameter selections to no longer require hand-tuning and supervision of weights.

[1] https://github.com/alexanderepstein/tfrogs

6. REFERENCES

[1] Po-Yu Chen and Ivan W Selesnick, "Translation-invariant shrinkage/thresholding of group sparse signals," *Signal Processing*, vol. 94, pp. 476–489, 2014.

[2] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[3] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[4] Ilker Bayram, "Mixed norms with overlapping groups as signal priors," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4036–4039.

[5] Ingrid Daubechies, *Ten Lectures on Wavelets*, SIAM, 1992.

[6] Eero P Simoncelli and Bruno A Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[7] Jianjun Huang, Xiongwei Zhang, Yafei Zhang, Xia Zou, and Li Zeng, "Speech denoising via low-rank and sparse matrix decomposition," *ETRI Journal*, vol. 36, no. 1, pp. 167–170, 2014.

[8] Tianyi Zhou and Dacheng Tao, "Godec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.

[9] Yipeng Li and DeLiang Wang, "Musical sound separation based on binary time-frequency masking," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–10, 2009.

[10] Juan Liu and Pierre Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Transactions on Image processing*, vol. 10, no. 11, pp. 1647–1658, 2001.

[11] Eero P Simoncelli and Bruno A Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[12] Lukas Meier, Sara Van De Geer, and Peter Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[13] Xin Liu, Guoying Zhao, Jiawen Yao, and Chun Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2502–2514, 2015.

[14] Fredrik Lindsten, Henrik Ohlsson, and Lennart Ljung, "Clustering using sum-of-norms regularization: With application to particle filter output computation," in *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 2011, pp. 201–204.

[15] Kenneth Lange, "The mm algorithm," April 2007, `https://www.stat.berkeley.edu/~aldous/Colloq/lange-talk.pdf`.

[16] David R. Hunter and Kenneth Lange, "Quantile regression via an mm algorithm," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 60–77, 2000.

[17] Mário AT Figueiredo, José M Bioucas-Dias, and Robert D Nowak, "Majorization–minimization algorithms for wavelet-based image restoration," *IEEE Transactions on Image processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[18] T. Parks and J. McClellan, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Transactions on Circuit Theory*, vol. 19, no. 2, pp. 189–194, 1972.

[19] EH Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[20] Guoshen Yu, StÉphane Mallat, and Emmanuel Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1830–1839, 2008.

[21] Jianfen Ma, Yi Hu, and Philipos C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387, May 2009.

[22] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.