STRUCTURED SPARSITY FOR AUDIO SIGNALS

Kai Siedenburg[*†]*, Monika Dörfler*[†]

[*] Department of Mathematics, Humboldt University Berlin
[†]NuHAG, Faculty of Mathematics, University of Vienna, Austria
kai.siedenburg@gmx.de,monika.doerfler@univie.ac.at

ABSTRACT

Regression problems with mixed-norm priors on time-frequency coefficients lead to structured, sparse representations of audio signals. In this contribution, a systematic formulation of thresholding operators that allow for weighting in the time-frequency domain is presented. The related iterative algorithms are then evaluated on synthetic and real-life audio signals in the context of denoising and multi-layer decomposition. Further, initial results on the influence of the shape of the weighting masks are presented.

1. INTRODUCTION

Most audio signals of importance for humans, in particular speech and music, are highly structured in time and frequency. Typically, salient signal components are sparse in time (or frequency) and persistent in frequency (or time). Sparsity in time is connected to transient events, while sparsity in frequency is observed in harmonic components. Processing sound signals with time-frequency dictionaries is ubiquitous. The sparse structure usually seen can be further enhanced by procedures such as basis pursuit [1] or ℓ^1-regression [2]. In the context of time-frequency dictionaries, a natural step beyond classical sparsity approaches is the introduction of sparsity criteria which take into account the two-dimensionality of the time-frequency representations used. Mixed norms on the coefficient arrays make it possible to enforce sparsity in one domain and diversity and persistence in the other domain. Regression with mixed-norm priors was first proposed in [3]. In the current contribution, we consider a family of specific regression problems with ℓ^1 and ℓ^2 priors on the coefficients; the algorithms derived thereof are refined by using local neighborhood-weighting. The performance of the resulting different operators is systematically evaluated for classical signal processing tasks like de-noising and sparse multi-layer decomposition. Applications lead to quite satisfactory results in terms of measured (SNR) and listening.
The presented results reflect a first step in the exploitation of structured shrinkage in the sense of *informed analysis*, i.e., using some available prior knowledge about the signal under consideration. The main contribution is the generalization and application of structured shrinkage operators [3] to representations of audio signals by frames.

2. TECHNICAL TOOLS

We seek to expand a signal $s \in \mathbb{C}^L$ in the form

$$s(n) = \sum_{k,j} c_{k,j}\varphi_{k,j}(n) + r(n),\ n = 1, \dots, L \qquad (1)$$

This work was supported by the Austrian Science Fund (FWF) project LOCATIF(T384-N13) and the WWTF project Audio-Miner (MA09-024)

where the $\varphi_{k,j}$ denote the atoms of a time-frequency dictionary Φ, $c_{k,j}$ are the expansion coefficients and r is some residual. In order to guarantee perfect and stable reconstruction of a signal from its associated analysis coefficients $c_{k,j} = \langle s, \varphi_{k,j}\rangle$, we assume that the dictionary Φ forms a frame [4]. We consider Gabor frames, which are exhaustively used in music processing, be it under a different name: in their simplest instantiation they correspond to a sampled sliding window or short-time Fourier transform. Gabor frames consist of a set of atoms $\varphi_{k,j} = M_{bj}T_{ka}\varphi$, where T_x and M_ω denote the time- and frequency-shift-operator, resp., defined by $T_x\varphi(n) = \varphi(n-x)$, $M_\omega\varphi(n) = \varphi(n)e^{\frac{2\pi i n\omega}{L}}$, and φ is a standard window function. a and b are the time- and frequency sampling constants, and $j = 0, \dots, J-1, k = 0, \dots K-1$, with $Ka = Jb = L$.
We will even assume more, namely tightness of the frames in use, which means that, up to a constant which may be set to 1, we have $s = \sum_{k,j}\langle s, \varphi_{k,j}\rangle\varphi_{k,j}$, i.e., synthesis is done with the analysis window. Tight frames are easily calculated, see [5]. In the finite discrete case, the frame's atoms constitute the columns of a matrix Φ which is of dimension $L\times p$; for tight frames, we have $\Phi\cdot\Phi^*\cdot s = s$. Since we are especially interested in the redundant case $L < p$, the additional degrees of freedom are used to promote sparsity of the coefficients.

2.1. Regression with mixed norms

Sparsity of coefficients may be enforced by ℓ^1-regression, also known as the *Lasso* [2]. Given a noisy observation $y = s + e$ in $\mathbb{C}^L$ it finds

$$\hat{c} = \arg\min_{c\in\mathbb{C}^p} \frac{1}{2}\|y - \Phi c\|_2^2 + \lambda\Psi(c) \qquad (2)$$

with penalty term $\Psi(\cdot) = \|\cdot\|_1$ and $\lambda > 0$. Since the sequence $c_{k,j}$ is ordered along two dimensions for Gabor frames, the ℓ^1-prior Ψ in (2) may be replaced by a two-dimensional mixed norm $\ell^{p,q}$ which acts differently on groups (indexed by g in the sequel, may be either time or frequency) and their members (indexed by m):

$$\Psi(c) = \|c\|_{p,q} = \left(\sum_g \left(\sum_m |c_{g,m}|^p\right)^{q/p}\right)^{1/q} \qquad (3)$$

Subsequently, the notation (g, m) will be used in reference to the group-member structure, whereas (k, j) refers to the time-frequency indices of the Gabor-expansion. In terms of $\ell^{p,q}$, we consider the cases $p = 2, q = 1$ and $p = 1, q = 2$. The former is known as *Group-Lasso (GL)* [6] (promoting sparsity in groups and diversity in members) and the latter was termed *Elitist-Lasso (EL)* in [3]: the

$\ell^{1,2}$ constraint promotes sparsity in members, only the "strongest" members (relative to an average) of each group are retained.
Landweber iterations, which solve (2) in the ℓ^1-case, [4], also yield a solution to the generalized minimization problem induced by (3), if standard soft thresholding is replaced by a generalized thresholding operator $\mathbb{S}_{\lambda,\xi}(z_{g,m}) = z_{g,m}(1-\xi(z))^+$. Here, $\xi = \xi_{(g,m),\lambda}$ is a non-negative function dependent on the index (g,m) and λ. The solution to (2) is then given by the iterative Landweber algorithm: choosing arbitrary c^0, set

$$c^{n+1} = \mathbb{S}_{\lambda,\xi}(c^n - \Phi^*(y - \Phi c^n)). \tag{4}$$

It was shown in [7], that the use of the thresholding operators $\mathbb{S}_{\lambda,\xi}$, defined via ξ, leads to convergence of the iterative sequence (4) to the minimizer of (2):

$$p=1, q=1 : \xi^L(c_{g,m}) = \frac{\lambda}{|c_{g,m}|} \quad (\textit{Lasso}) \tag{5}$$

$$p=2, q=1 : \xi^{GL}(c_{g,m}) = \frac{\lambda}{(\sum_m |c_{g,m}|^2)^{\frac{1}{2}}} \quad (\textit{GL}) \tag{6}$$

$$p=1, q=2 : \xi^{EL}(c_{g,m}) = \frac{\lambda}{1+M_g\lambda}\frac{\|\boldsymbol{c}_g\|_1}{|c_{g,m}|} \quad (\textit{EL}) \tag{7}$$

where $\boldsymbol{c}_g = (c'_{g,1}, \ldots, c'_{g,M_g})$ and $\{c'_{g,m'}\}_{m'}$ denotes for each group g the sequence of scalars $|c_{g,m}|$ in descendant order. M_g denotes some natural number depending on the magnitudes of coefficients in the group $(c_{g,1}, \ldots, c_{g,M})$ [1].

2.2. Refining the algorithms

To exploit structures in audio signals, like persistence in time or frequency, we refine the shrinkage operators introduced above for application in audio analysis. The coefficient $c_{g,m}$ (or groups of them) undergo shrinkage according to the energy of a *time-frequency neighborhood*. In contrast to the *groups* of GL and EL, the neighborhoods can be modeled flexibly, e.g., using weighting and overlap. Hence, we compose ξ with some neighborhood weighting functional η_N:
To an index $\gamma = (g,m)$ in a structured index set $\mathcal{I}$, we associate a *(weighted) neighborhood* $N(\gamma) = \{\gamma' \in \mathcal{I} : w_\gamma(\gamma') \neq 0\}$ with weights w_γ defined on $\mathcal{I}$ such that $w_\gamma(\gamma) > 0$, $w_\gamma(\gamma') \geq 0$ for all $\gamma' \in \mathcal{I}$ and $\sum_{\gamma' \in N(\gamma)} w_\gamma(\gamma')^2 = 1$. Then, with $\eta_N(c_\gamma) = \left(\sum_{\gamma' \in N(\gamma)} w_\gamma(\gamma')^2 |c_{\gamma'}|^2\right)^{1/2}$, we obtain the generalized shrinkage operators by setting[2]

$$\begin{aligned}
\xi^{WGL} &= \xi^L \circ \eta_N \quad (\textit{windowed GL (WGL)}),\\
\xi^{PEL} &= \xi^{EL} \circ \eta_N \quad (\textit{persistent EL (PEL)}),\\
\xi^{PGL} &= \xi^{GL} \circ \eta_N \quad (\textit{persistent GL (PGL)})
\end{aligned}$$

in (5)-(7). These generalized shrinkage operators are not associated to a simple convex penalty functional, cp. [3]. Convergence properties of their Landweber-iterations are currently under study, and numerical experiments suggest convergence.

[1] Cp. [7] for a more involved, but exact definition of the M_g in EL.

[2] [3] introduces WGL as generalization of GL while from a formal point of view it would be more appropriate to call it windowed *Lasso*. Nonetheless, we stick to the former nomenclature.

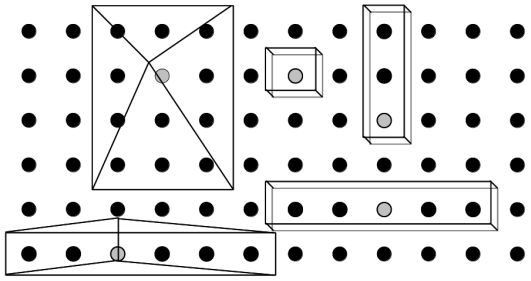

Figure 1: *Sketch of different shapes of the parameterization of the neighborhoods in a schematic time-frequency plane. Rectangular and triangular ("tent"-like) windows were implemented.*

3. SIMULATIONS

The generalized shrinkage operators were implemented in `MATLAB` with the following parameterization of the neighborhoods: For each time-frequency-index (k,j) and a neighborhood size vector $\sigma = (\sigma_1, \ldots, \sigma_4)$, the neighborhood N_σ is defined as the set of indices $N_\sigma(k,j) = \{(k',j') : k' \in \{k-\sigma_4, k+\sigma_2\}, j' \in \{j-\sigma_3, j+\sigma_1\}\}$. Neighborhoods of indices close to a border of the time-frequency plane are obtained by mirroring the index set at the respective border. Rectangular and triangular weighting of the neighborhoods was implemented, with rectangular weighting only in section 3.1 and 3.2. In the plots, an index after the operator's abbreviation specifies the group-label as time or frequency (not needed for Lasso and WGL), e.g., PEL-t signifies that the group in the respective elitist lasso is time. For the neighborhood-smoothed operators WGL, PEL, and PGL the neighborhood-size vector σ is given. To test the obtained variety of shrinkage operators, we used a simulated "toy"-signal consisting of a stationary, a transient and a noise part.[3] The stationary part consists of four harmonics with fundamental frequency 440Hz and decreasing amplitudes. The obtained harmonics were shaped by a linear envelope in attack and decay. The transient part was simulated by 4 equidistant impulses with similarly decaying amplitudes. Finally, Gaussian white noise with SNR about 15 and 3dB was added.
Landweber iterations are known to converge very slowly and various methods of acceleration have been proposed [8]. As it was out of the scope of this paper to elaborate on these ideas, we used the basic iteration scheme (4). The iterations presented in the following were stopped after 100 steps. Then almost all of the final relative iteration errors were below 0.3%.

3.1. Structured denoising

As a first experiment the standard de-noising problem with additive Gaussian white noise was considered. We use a tight Gabor-frame with Hann window of length 1024 and hop size 256 (at sampling rate 44100Hz). We measure the operator's performance in SNR: with the estimation's approximate Landweber-limit c^* of (4) and $\hat{s} = \Phi c^*$, the SNR is $\mathit{SNR}(\hat{s}, s) = 20\log_{10}(\frac{s}{\hat{s}-s})$.
For comparison, the SNR is then plotted against the number of positive coefficients. Of the variety of possible operators, Fig-

[3] Corresponding sound files and more detailed visualizations are presented at the conference and on the webpage `http://homepage.univie.ac.at/monika.doerfler/StrucAudio`.

ure 2 presents the SNR curves of the best basic operators and their neighborhood smoothed counterparts (of which again the best of each type were chosen for the figure) at two different noise levels. It is obvious that for the lower noise level Lasso and WGL (with neighborhoods in frequency) perform best, where the WGL still outperforms the Lasso. Also, WGL with neighborhoods in time outperforms Lasso for the higher noise level. Yielding high SNR over a broad range of sparsity values, WGL thus seems to be a good choice for the de-noising task (and we made the same observations for "real life"-audio signals). Compared to WGL, the other operators GL, PGL, EL and PEL perform quite badly for the lower noise level. While PGL is constantly worse than GL, PEL seems to have some advantages over EL for higher sparsity levels. However, GL is surprisingly the second best operator for de-noising the toy example at the high noise level. Experiments with longer and more complex audio excerpts do not replicate this result, which is not surprising, since the structure of GL naturally promotes simple signal structures (which can be advantageous in some cases, see 3.2). Conclusively, the neighborhood smoothing seems to pay off in the de-noising task with Gaussian white noise, where the persistent operators WGL and PEL outperform their respective counterparts Lasso and EL. An exception is made by PGL, which performs in our experiments constantly worse than GL. Concerning the perceptual quality of the de-noised audio-material, the neighborhood smoothing of the modified operators promotes continuity in the coefficients and thus reduces the probability of isolated high energy coefficients and we observed less musical noise and higher perceptual audio quality, especially under WGL.

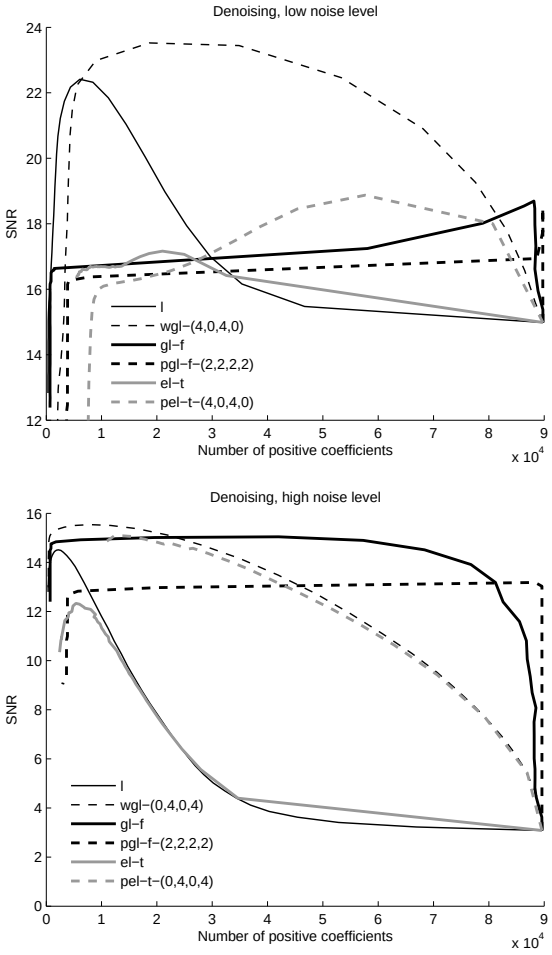

Figure 2: *Overview of de-noising behavior of (modified) shrinkage operators (*l *is for Lasso). Performance measured in SNR against the number of positive coefficients at two different noise levels (*15*dB and* 3*dB).*

3.2. Multilayer decomposition

We continue processing the toy-example by aiming to extract the signal's tonal and transient parts at the lower noise level (15dB SNR). For estimating the tonal layer, we use a tight Gabor frame with Hann window, window-length 4096 and hop size 1024. The transients are estimated starting from the transient layer + noise (which corresponds to the unrealistic but complexity reducing assumption of perfect tonal estimation) using short windows (256 samples, hop size 64). Table 1 and 2 present the performance of a sample of operators, again of each type a "basic" and neighborhood-smoothed one. The tables show the maximum SNR value of the estimation measured w.r.t. the "true" respective layer and the corresponding percentage of retained coefficients.

Concerning the tonal estimation displayed in Table 1, WGL with neighborhoods expanding in time perform best for the extraction of the tonal part in terms of SNR, while retaining relatively few coefficients. GL, PGL and PEL exhibit comparable SNR, but only with far more coefficients. As in the situation of de-noising above, the neighborhood-smoothing is useful for the Lasso and the Elitist-Lasso, but not for Group-Lasso.

As Table 2 shows clearly, GL with time as group-index performs best in terms of SNR for the estimation of the transient layer. This result is not very surprising since the example's transient layer has a simple structure which supports the performance of GL. However, GL-t should be a good choice for transient extraction in more complex signals, since it yields broadband transients without extracting many horizontal (tonal) signal parts.

The next experiment addressed the decomposition of musical audio without the presence of quasi-ground truth. For this "real-life" application, the choice of sparsity level λ is always a difficult task. We chose the SNR-maximizing candidates from the simulations. This yielded WGL (with rectangularly weighted neighborhoods extending 4 elements in each direction of time) as estimator of the tonal layer with sparsity level $\lambda = 0.080$, and GL (with the time-index as group label) for the transient layer with $\lambda = 0.072$. We used a 5 seconds excerpt of a Jazz-record containing piano, double-bass and drums. In the decomposition, the drums (and some percussive elements of the bass) are well separated from the harmonics of piano and bass. Using GL as transient estimator works well in this example, it captures all of the soft 16th notes drum-patterns. We observed a trade-off in the choice of the sparsity level: increasing sparsity in the tonal estimation improves the separation of both layers but leads to increased damping of higher, low-energy partials of the tonal part.

3.3. Shapes

As described at the beginning of this section, the neighborhoods' shapes (constituted by size and weighting) were implemented and parametrized in a straight-forward fashion, so far allowing for rectangular domains with either uniform (i.e. rectangular) or triangular

Table 1: *Comparison of the performance of different operators in tonal estimation: maximum SNR values of estimation and "true" layer and corresponding number of retained coefficients in percent.* * *refers to neighborhoods* $(4, 0, 4, 0)$ *while* $^+$ *to* $(0, 4, 0, 4)$.

Operator	Lasso	WGL$^+$	GL-f	PGL-f*	EL-t	PEL-t$^+$
max. SNR	28.7	31.2	30.5	30.7	26.2	30.6
%Coeffs	0.4	1.1	3.3	17.0	2.8	4.1

Table 2: *Transient estimation: maximum SNR values of estimation and "true" layer and corresponding number of retained coefficients in percent. As above:* * *means* $(4, 0, 4, 0)$ *and* $^+$ *means* $(0, 4, 0, 4)$.

Operator	Lasso	WGL*	GL-t	PGL-t$^+$	EL-f	PEL-f*
max. SNR	10.4	13.2	14.4	9.5	10.4	13.3
%Coeffs	1.0	2.9	2.2	38.9	1.4	3.7

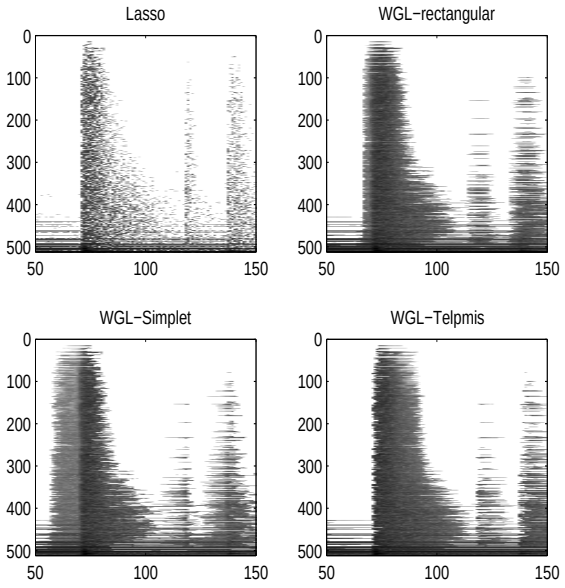

Figure 3: *Iterated WGL shrinkage results for different shapes (i.e. weightings) of the neighborhood on an snare drum hit excerpt. From left to right: Lasso, Rectangular, Simplet (= simple tent, starting at 1 and then linearly decaying to zero), Telpmis (= time-reversed simple tent).*

(i.e. "tent"-like) weightings. These shapes do not necessarily have to be symmetric at the origin, as the energy of most audio signals is not symmetrically distributed around its peaks either. This fact can be exploited to feature different parts of a signal under observation. Consider Figure 3, where the iterated WGL-shrinkage results with four different neighborhood-shapes, each solely extending in time, are compared (based on a Gabor-frame with window length 1024 and overlap of 4). It is obvious that the shapes yield different (sparse) perspectives on the signal content. Whereas the symmetric neighborhoods naturally captures parts before and after the attacks (or rather time-points of maximum energy), the asymmetric ones rather retain components before (resp. after) the attacks. The orientation of the neighborhood therefore systematically promotes the preservation of different temporal segments of the signal.

4. SUMMARY AND PERSPECTIVES

We presented first results on structured sparsity approaches for Gabor frames to audio signals. Future work will focus on the convergence of the algorithms, both in a theoretical and computational setting. By taking into account methods as [9] the proposed algorithms should be accelerated significantly. On the contrary, evaluations of the algorithms' perceptual qualities will be considered. Further, using various shapes for the weight, we aim at the extraction of more specific structures, in the sense of *sound objects* [10].

5. ACKNOWLEDGMENTS

We thank Matthieu Kowalski and the anonymous reviewers for their valuable advice.

6. REFERENCES

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

[2] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[3] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing, doi:10.1007/s11760-008-0076-1*, 2009.

[4] Stéphane Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, 2009.

[5] M. Dörfler, "Time-frequency Analysis for Music Signals. A Mathematical Approach," *Journal of New Music Research*, vol. 30, no. 1, pp. 3–12, 2001.

[6] Ming Yuan and Yi Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.

[7] Matthieu Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 303–324, 2009.

[8] Ignace Loris, "On the performance of algorithms for the minimization of l1-penalized functionals," *Inverse Problems*, vol. 25, 2009.

[9] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[10] G. Cornuz, L. Daudet, P. Leveau, and E. Ravelli, "Object coding of harmonic sound using sparse and structured representations.," in *Proc. of DAFx-07, Bordeaux, France*, 2007.