Cosparse Denoising: The Importance of Being Social

Clément Gaultier
Inria, Centre Inria Rennes
Rennes, France

Srđan Kitić
Technicolor,
Cesson-Sevigné, France

Nancy Bertin
IRISA - CNRS UMR 6074,
Rennes, France

Rémi Gribonval
Inria, Centre Inria Rennes
Rennes, France

***Abstract*—This work investigates the performance of cosparse *vs.* social cosparse regularizations in addressing the audio denoising problem. Beyond the cosparse (also known as sparse analysis) model, results show that exploiting structures in the time-frequency domain is beneficial to audio signal restoration for high degradation levels.**

I. Introduction

The last decades popularized time-frequency (TF) sparse models [1] for audio reconstruction. Some work [2], [3] showed that the sparse analysis (or cosparse) way of modeling signals could exhibit low computational costs and are worth considering, particularly if the target application requires real-time processing. The cosparse model considers $\mathbf{z} \stackrel{\text{def}}{=} \mathbf{A}\mathbf{x}$ approximately sparse with $\mathbf{A}$ the analysis operator (forward frequency transform like DCT or DFT) and $\mathbf{x}$ the time-domain signal of interest. Besides, structured sparsity (from the synthesis point of view) [1] and especially the social sparsity framework [4] were recognized to be able to better capture and take into account some typical TF patterns in audio signals.

Motivated by the success of the sparse analysis version of the SPADE algorithm [3] for declipping, as well as the potential of the social sparsity framework [4] for denoising, we postulate that coupling these two concepts could be beneficial to audio restoration. The following work compares the performance of regular cosparse and social cosparse models to state-of-the-art time-frequency block-thresholding method (BT) [5] on the audio denoising problem.

II. Cosparse and Social Cosparse Algorithms

We consider the following degradation model for the case of time-domain signals corrupted with additive noise: $\mathbf{y} = \mathbf{x} + \mathbf{e}$, with $\mathbf{x}$ an audio signal and $\mathbf{e}$ modeled as white Gaussian noise of variance σ^2. We process blocks of overlapping frames $\mathbf{Y}_\mathsf{n} \in \mathbb{R}^{\mathsf{L}\times(2\mathsf{b}+1)}$ from the signal $\mathbf{y}$, on which a Hamming analysis window is applied ($\mathbf{y}_\mathsf{n}$ is the center frame of the block and L the frame size in samples).

Cosparse regularized approaches to inverse problems can be cast as an optimization problem, where the cost function to minimize is a sum of a data-fidelity term (here $\|\hat{\mathbf{X}}_\mathsf{n} - \mathbf{Y}_\mathsf{n}\|_\mathrm{F}^2$) and a regularization term enforcing sparsity. Depending on the choice of these two terms, non-convexity and/or non-smoothness can prevent from using conventional optimization algorithm to solve it. The ADMM framework [6], which we use here, allows to alleviate this problem. We define an iterative ADMM procedure in which, notably, a well-chosen *shrinkage* or *thresholding* is applied at each iteration and acts as a proxy for the cosparse regularization term.

The choice of this proxy is the key difference between regular cosparse and social cosparse algorithms. In the regular cosparse case, the well-known *hard-thresholding* operator $\mathcal{H}_\mu(\cdot)$ is applied at each iteration on the current estimate $\mathbf{A}\hat{\mathbf{x}}_\mathsf{n}$. In the social cosparse case, as we wish to promote some time-frequency particular structures or patterns, hard-thresholding is replaced by *Persistent Empirical Wiener (PEW) shrinkage* defined in [7]. This shrinkage explicitly includes a time-frequency *neighborhood* Γ which promotes local time-frequency structures around each time-frequency point ij. Thus, the shrinkage on a given frame n involves not only the frame n itself but also $2\mathsf{b}$ adjacent frames symmetrically selected around $\mathbf{y}_\mathsf{n}$ or its frequency representation $\mathbf{z}_\mathsf{n}$. We can summarize this shrinkage as:

$$\mathcal{S}_\mu(\mathbf{Z}_\mathsf{n}|\Gamma)_{(\mathsf{ij})} = \mathbf{Z}_{\mathsf{n}(\mathsf{ij})} \cdot \max\left(\left(1 - \frac{\mu^2}{\|\mathbf{Z}_\mathsf{n}\mathbf{P}_\mathsf{ij} \circ \Gamma\|_2^2}\right), 0\right), \quad (1)$$

where we recall that $\mathbf{Z}_\mathsf{n} = \mathbf{A}\mathbf{X}_\mathsf{n}$, and (ij) is a time-frequency index. One TF neighborhood is chosen for all TF points in a given $\mathbf{Z}_\mathsf{n}$ among a subset $\left\{\Gamma_{(\mathsf{k})}\right\}_{\mathsf{k}=1..6}$ of possible predefined patterns (examples are given in Figure 1). $\mathbf{P}_\mathsf{ij}$ are the indexes corresponding to the binary TF patch associated with Γ centered in (ij). $\mathcal{S}_\mu(\cdot)_{(\mathsf{ij})}$ is applied component-wise, therefore it can be computed through multidimensional convolution in the Fourier domain. The shrinkage parameter μ (and thresholding parameter μ in the regular sparse case respectively) is adapted at each ADMM iteration following a specific decreasing scheme. The exact procedures for the choice of Γ and μ are not described here due to lack of space. After the denoising step, we perform an overlap-add synthetization from the denoised estimates $\hat{\mathbf{x}}_\mathsf{n} \in \mathbb{R}^\mathsf{L}$ to yield the denoised signal $\hat{\mathbf{x}}$.

III. Experimental Study

We conducted numerical tests on items from the RWC Music Database [8]. We processed excerpts from the "Pop", "Jazz" and "Classical Orchestra" music subcategories. Around 1 hour of audio content in total from each genre was contaminated with additive white noise at five Signal-to-Noise Ratios (SNR) {0, 5, 10, 15, 20} dB. Each excerpt was then denoised using BT [5], the cosparse and the social-cosparse approaches. The algorithms parameters are listed in Table I. The local TF neighborhoods available for the social cosparse approach are those presented in Figure 1. Figure 2 displays the improvements in dB as a function of the input SNR for the BT, social cosparse and regular cosparse approaches. Results presented for popular, jazz and classic orchestral music show that either the social cosparse method (for low input SNR) or both the social and simple cosparse techniques (for high input SNR) numerically outperform BT. While performance is somewhat similar between social and regular approaches at high enough SNR (or even in favor of the latter, on one subset), we observe a clear superiority of the social approach in severe noisy conditions.

IV. Conclusion

This work shows that cosparse models are suitable to recover signals in the audio denoising context. While both cosparse and social cosparse approaches perform better than state-of-the-art at high SNR, the regular cosparse method shows its limitations from moderate to low SNR. By contrast, the joint use of cosparse and structured sparsity models is particularly efficient at low SNR, and numerically outperforms state-of-the-art block-thresholding in this case.

Acknowledgment

This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906).

Figure 1. Example set of predefined time-frequency neighborhoods for PEW shinkage. On each $\Gamma_{(k)}$, columns correspond to time frames and rows to frequency bins. For instance, we expect that $\Gamma_{(1)}$ will emphasize tonal content, while $\Gamma_{(2)}$ should be more suitable for transients and attacks.

Table I
EXPERIMENTAL PARAMETERS

Parameters	Frame size [samples]	Overlap [%]	Overlapping segments	Accuracy	Analysis operator	Set of neighborhood
Value	$\mathsf{L} = 1024$	75	$\mathsf{b} = 5$	$\beta = 10^{-3}$	$\mathbf{A} = \mathrm{DFT}$	$\{\Gamma_{(1)}, \Gamma_{(2)}, \ldots \Gamma_{(6)}\}$.

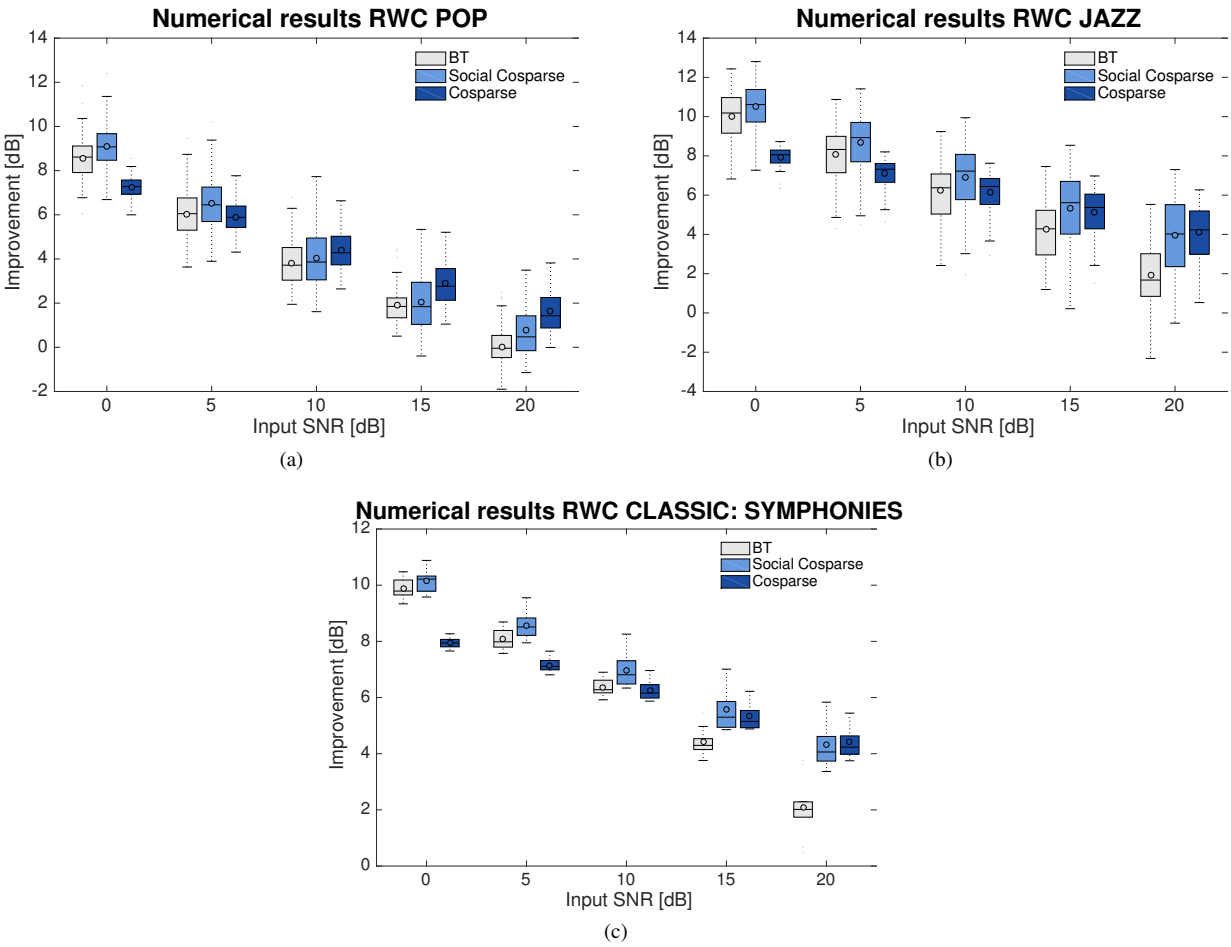

Figure 2. Experimental Comparison for SNR improvements

REFERENCES

[1] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[2] S. Kitić, "Cosparse regularization of physics-driven inverse problems," PhD Thesis, IRISA, Inria Rennes, 2015. [Online]. Available: https://hal.archives-ouvertes.fr/tel-01237323

[3] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *Latent Variable Analysis and Signal Separation (LVA/ICA)*. Liberec, Czech Republic: Springer, 2015, pp. 243–250.

[4] M. Kowalski, K. Siedenburg, and M. Dorfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

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

[6] J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[7] M. Kowalski, "Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 4151–4155.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases." in *ISMIR*, vol. 2, 2002, pp. 287–288.