SPARSE NON-LOCAL SIMILARITY MODELING FOR AUDIO INPAINTING

Ichrak Toumi, Valentin Emiya

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
firstname.lastname@lis-lab.fr

ABSTRACT

Audio signals are highly structured from a low, signal level to high cognitive aspects. We investigate how to exploit the common sparse structure between similar audio frames in order to reconstruct missing data in audio signals. While joint sparse models and related algorithms have been widely studied, one important challenge is to locate such similar frames : the search must be adapted to the joint-sparse model and should be fast and one must deal with missing data in the frames. We propose, compare and discuss several similarity measures dedicated to this task. We then show how this strategy can lead to better reconstruction of missing data in audio signals.

Index Terms— Sparse approximation, joint sparsity, inpainting, matching pursuit, audio.

1. INTRODUCTION

Audio signals are highly structured at many levels. The lowest level of structuring is local in time and includes frequency components, transients and noise, which have been efficiently modeled by sparse representations [1]. Non-local similarities are another source of structure: speech, music and other sounds are indeed composed of patterns that occur several times in a signal: phonemes, musical notes, and so on. Such non-local similarities have proven useful to process images [2] and sounds [3, 4].

Exploiting both local structures and non-local ones in the context of sparse representations has been made possible using joint or simultaneous sparse models, especially for image inpainting [5]. Similar image patches or audio frames are simultaneously decomposed in order to find sparse representations sharing the same support, as represented in Figure 1. Provided that these similar areas have been adequately selected, the sparse representations are better estimated [6, 7, 8].

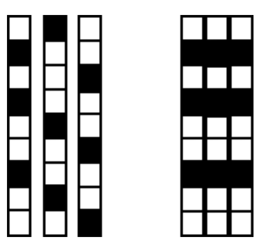

Fig. 1. Location of non-zero coefficients (black) in sparse (left) and joint-sparse (right) vectors.

We investigate the use of such joint-sparse models for audio inpainting, where selecting similar audio frames when data is missing is a non-trivial task. Audio inpainting is about restoration of localized missing parts in an audio signal. This could be due to impulsive noise or clicks, CD scratches, clipping or packet loss. The methods presented in the literature range from interpolation [9] and Bayesian approach [10] to sparse representation (SR) which proved to be fairly competitive with state-of-the-art methods [11]. In the latter, the local sparse structure of audio frames in a Gabor dictionary is exploited for recovering missing or distorted audio data with extensions of the Orthogonal Matching Pursuit (OMP) algorithm [12]. More recently, application to audio inpainting in the clipping case have been presented in [13] where the authors use dependencies between neighboring coefficients to improve the declipping algorithm. The results are promising and demonstrate that there is an interest in using such similarity information for predicting the missing samples. They show the benefits of joint sparsity for audio inpainting, but the approach is limited to the simultaneous decomposition of neighboring frames and the question of finding non-local similarities is not addressed. For non-local structures, recent works [3, 4] have introduced a method for restoring long duration gaps in audio signals using similarity graphs. The inpainting of a large hole is realized by extracting features from the surrounding data and by finding similar regions based on these border contents.

In this work, we propose a new framework based on a joint-sparsity model to overcome the audio inpainting problem in the time domain. The main issue and originality in this approach is to find the similar non-local regions in the audio signal, i.e., regions that share common features, and especially a common sparse support, while samples are missing in the target region as well as in the selected regions. This may be achieved by using similarity measures: the choice of an appropriate audio similarity measure is is crucial and non-trivial. We propose a comparative study between several measures of similarity. It shows that the correlation measure widely used to find similar patches in image processing is not appropriate for audio data. As a result, we select a more effective and robust similarity measure for audio inpainting.

The paper is organized as follows. In section 2, we present the simultaneous sparse approximation problem in the context of inpainting. Then in section 3, we introduce and compare several measures to select similar audio frames for a joint-sparse decomposition. Finally, we present inpainting results in section 4.

2. JOINT SPARSE REPRESENTATION

2.1. Notations

We adopt the following notations:

- $\underline{\mathbf{s}} \in \mathbb{R}^{L_s}$ is a full signal with length L_s and $\underline{\mathbf{m}} \in \{0,1\}^{L_s}$ is a related binary mask vector where the elements equal 1 for observed samples and 0 for missing samples.

This work was supported by ANR JCJC program MAD (ANR-14-CE27-0002).

- The signal is segmented into overlapping frames with length L: for $i \in [1, L_s - L + 1]$, $\mathbf{s}_i = [\underline{\mathbf{s}}(i+n)]_{n=0}^{L-1} \in \mathbb{R}^L$ denote the frame starting at sample i and $\mathbf{m}_i = [\underline{\mathbf{m}}(i+n)]_{n=0}^{L-1} \in \{0,1\}^L$ denotes the related mask vector.
- supp $\mathbf{x}$ denotes the indexes of non-zero entries of a vector $\mathbf{x}$.
- $\|\mathbf{X}\|_{p,q} \triangleq \sum_r \|\mathbf{x}^r\|_q^p$ is the mixed norm $\ell_{p,q}$ of matrix $\mathbf{X}$ for any $p, q \in \mathbb{R}^+$. $\mathbf{x}^r$ denotes the r^{th} row of $\mathbf{X}$.
- $\mathbf{x}^\nu = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ is the normalized version of any vector $\mathbf{x} \neq 0$.
- $|.|$ is the cardinal of a set and $\odot$ the Hadamard product.

2.2. Standard sparse approximation problem

In a Sparse Representations (SR) modeling framework [14], each audio frame can be written as:

$$\mathbf{s} = \mathbf{D}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{L \times \Omega}$ is the so-called dictionary, $\mathbf{x} \in \mathbb{R}^\Omega$ is the sparse representation and $\mathbf{n} \in \mathbb{R}^L$ is the noise.
The SR model was adapted to the audio inpainting case in [11]. Given a frame $\mathbf{s}$ with binary mask $\mathbf{m} \in \{0,1\}^L$, the observation is

$$\mathbf{y} = \mathbf{m} \odot \mathbf{s}. \tag{2}$$

and the inpainting optimization problem writes

$$\arg\min_{\mathbf{x}} \ \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{m} \odot (\mathbf{D}\mathbf{x})\|_2^2 < \epsilon \tag{3}$$

where $\epsilon > 0$ is a tolerance on the residual energy. Since the ℓ_0 norm leads to an NP-hard problem, an approximated sparse solution may be obtained using a variant of the OMP algorithm [12] where all the dictionary columns are internally re-normalized to unit norm due to the partial observations. The unknown samples are then recovered from the given sparse solution.

2.3. Joint sparse problem

Let us assume that we are provided a similarity measure γ to select audio frames with similar sparse support. Given a target frame $\mathbf{s}$ and a set of candidate frames $\mathbf{s}_i, i \in [1, L_s - L + 1]$, we define the selected frames by the index set:

$$\mathcal{S} \triangleq \{i \in [1, L_s - L + 1] \mid \gamma(\mathbf{s}, \mathbf{s}_i) \geq \gamma_{\mathcal{S}}\} \tag{4}$$

where parameter $\gamma_{\mathcal{S}}$ is adjusted to control the number of selected frames $|\mathcal{S}|$. Then, the joint sparsity problem is formulated as

$$\arg\min_{\mathbf{X}_{\mathcal{S}}} \|\mathbf{X}_{\mathcal{S}}\|_{p,q} \quad \text{s.t.} \quad \|\mathbf{S}_{\mathcal{S}} - \mathbf{D}\mathbf{X}_{\mathcal{S}}\|_F^2 < \epsilon_{\mathcal{S}} \tag{5}$$

where $\epsilon_{\mathcal{S}} > 0$ and $\mathbf{S}_{\mathcal{S}} = [\mathbf{s}_i]_{i \in \mathcal{S}}$ (resp. $\mathbf{X}_{\mathcal{S}} = [\mathbf{x}_i]_{i \in \mathcal{S}}$) is a matrix in which each column is a selected frame (resp. a related sparse vector). In the literature, many algorithms have been designed to solve this optimization problem where the pair (p, q) takes the values $(0, \infty)$ to count the number of non-zero rows or (1, 2), for a convex relaxation, so that joint-sparsity is enforced in $\mathbf{X}_{\mathcal{S}}$.

In this paper, we propose to use the greedy algorithm called Simultaneous Orthogonal Matching Pursuit (S-OMP) which generalizes the OMP algorithm to the joint-sparsity case [6]. This approach is efficient when all the input signals are well approximated by the same set of atoms which is our case. The inpainting problem with (S-OMP) for a frame $\mathbf{s}$ is then

$$\arg\min_{\mathbf{X}_{\mathcal{S}}} \|\mathbf{X}_{\mathcal{S}}\|_{0,\infty} \text{ s.t. } \|\mathbf{M}_{\mathcal{S}} \odot (\mathbf{S}_{\mathcal{S}} - \mathbf{D}\mathbf{X}_{\mathcal{S}})\|_F^2 < \epsilon_{\mathcal{S}} \tag{6}$$

where $\mathbf{M}_{\mathcal{S}} = [\mathbf{m}_i]_{i \in \mathcal{S}}$ is the binary measurement matrix of the selected frames and $\|\mathbf{X}_{\mathcal{S}}\|_{0,\infty} = \left|\bigcup_{k \in \mathcal{S}} \text{supp}(\mathbf{x}_k)\right|$ counts the number of non-zero rows in matrix $\mathbf{X}_{\mathcal{S}}$.

The description of the S-OMP inpainting algorithm is given in Table 1. When $|\mathcal{S}| = 1$, it is equivalent to the OMP inpainting algorithm in [11]. The main idea in the algorithm, is that all dictionary columns are re-normalized to unit norm for only the reliable samples for each selected frame k in the similar set $\mathcal{S}$. Once the algorithm is executed for a frame $\mathbf{s}$, reconstruction is done using only coefficients corresponding to that frame index.

Input: $\mathbf{S}_{\mathcal{S}}, \mathbf{M}_{\mathcal{S}}, \mathbf{D} = \{\mathbf{d}_j\}_{j \in \Omega}, T, \epsilon_{\mathcal{S}}$
Initialize: • Iteration counter $iter = 0$ • Support Set $\mathcal{J}_0 = \emptyset$ • Residual $\mathbf{R}_0 = \mathbf{M}_{\mathcal{S}} \odot \mathbf{S}_{\mathcal{S}} = \mathbf{Y}_{\mathcal{S}}$
Repeat: until $iter = T$ or $\|\mathbf{R}_{iter}\|_F < \epsilon_{\mathcal{S}}$ • $iter = iter + 1$ • **For** $k \in \mathcal{S}$: – Dictionary normalization $\tilde{\mathbf{D}}_k = \text{diag}(\mathbf{m}_k) \times \mathbf{D} \times \mathbf{W}_k$ where $\mathbf{W}_k$ is a diagonal matrix such that $\mathbf{W}_k(j,j) = \|\text{diag}(\mathbf{m}_k) \times \mathbf{d}_j\|_2^{-1}$. – $\mathbf{proj}_k = \langle \tilde{\mathbf{d}}_{k_j}, \mathbf{R}_{iter-1}\mathbf{e}_k \rangle$, $\forall j \in \Omega$ and where e_k denotes the k^{th} canonical base vector in $\mathbb{R}^{\lvert\mathcal{S}\rvert}$. • Select atom $\hat{j} = \arg\max_{j \in \Omega} \sum_{k \in \mathcal{S}} \lvert\mathbf{proj}_k\rvert$ • Update support $\mathcal{J}_{iter} = \mathcal{J}_{iter-1} \cup \{\hat{j}\}$ • Update current solution for each column k $\mathbf{x}_{k,iter} = \arg\min_u \left\|\mathbf{y}_k - \tilde{\mathbf{D}}_{k,\mathcal{J}_{iter}} u\right\|_2$ • Update residual for each collumn k $\mathbf{r}_{k,iter} = \mathbf{y}_k - \tilde{\mathbf{D}}_{k,\mathcal{J}_{iter}} \mathbf{x}_{k,iter}$
Output: $\mathbf{X}_{\mathcal{S}}$ with: $\mathbf{x}_k = \mathbf{W}_k \mathbf{x}_{k,iter}, \forall k \in \mathcal{S}$

Table 1. S-OMP Inpainting algorithm.

3. SELECTION OF SIMILAR FRAMES

The effectiveness of the joint-sparsity approach depends on selecting similar frames. For a target frame $\mathbf{s}$ with approximated sparse vector $\mathbf{x}$, we would like to select a set of frames that have sparse approximations that share the same support as $\mathbf{x}$ without having to make a sparse decomposition at the frame selection step. To do so, we are looking for similarity measures that can mimic the comparison between the sparse supports. Such similarity measures have been developed in image processing [2] and are generally based on the correlation between images patches, which would correspond to correlating audio frames in our case. We study here how appropriate this choice is, showing that alternate similarity measures should be preferred.

3.1. Proposed measures

We propose to use different similarity criteria in order to find a useful measure for joint-sparse approximation in a redundant Fourier dictionary. Several measures between a target frames $\mathbf{s}$ and a candidate frame $\mathbf{s}_i$ are compared:

- **Normalized correlation:** the normalized correlation is defined in the time domain and can be computed in the Fourier domain by considering the angle between the DFT $\hat{\mathbf{s}}$ and $\hat{\mathbf{s}}_i$ of both frames:

$$\gamma_{corr}(\mathbf{s}, \mathbf{s}_i) \triangleq \langle \mathbf{s}^\nu, \mathbf{s}_i^\nu \rangle = \langle \hat{\mathbf{s}}, \hat{\mathbf{s}}_i \rangle = \cos\left(\angle\left(\hat{\mathbf{s}}^\nu, \hat{\mathbf{s}}_i^\nu\right)\right) \quad (7)$$

- **Spectral cosine similarity:** in order to discard phase effects, one may compute the cosine similarity between the normalized modulus of the DFT vectors as

$$\gamma_m(\hat{\mathbf{s}}, \hat{\mathbf{s}}_i) = \cos\left(\angle\left(|\hat{\mathbf{s}}|^\nu, |\hat{\mathbf{s}}_i|^\nu\right)\right)$$

- **Itakura Saito (IS) divergence:** from the IS divergence widely used for audio processing

$$d_{IS}(\hat{\mathbf{s}}, \hat{\mathbf{s}}_i) = \frac{1}{L}\sum_k \left[\left|\frac{\hat{\mathbf{s}}^\nu(k)}{\hat{\mathbf{s}}_i^\nu(k)}\right| - \log\left(\left|\frac{\hat{\mathbf{s}}^\nu(k)}{\hat{\mathbf{s}}_i^\nu(k)}\right|\right) - 1\right],$$

  One can get the IS similarity by normalizing it with respect to the maximum value and subtract the result from 1.

3.2. Experiment

In order to analyze how appropriate the similarity measures are, we propose to compare their ability to select the same frames as a reference method that actually computes the sparse decompositions and find the most similar ones.

The reference method select frames based on the Hamming distance between the supports of the sparse representations of the target frame $\hat{\mathbf{s}}$ and the candidate frames $\hat{\mathbf{s}}_i$ for $i \in [1, L_s - L + 1]$, using the OMP algorithm. We also compute all the above similarity measures and select the $|\mathcal{S}|$ most similar frames in each case. A similarity map is generated for the reference method and for each measure showing the selected frames and how they are located.

In Figure 2, we give an example for a small region in a speech audio signal sampled at 8KHz. We can see that the structures in the Hamming map are approximately reproduced in the similarity maps of the other measures except the correlation measure in which aligned structures parallel to the diagonal appear. Indeed, when correlating two frames, interferences due small phase differences cause low correlation values even when frames have similar sparse supports. As a consequence, the correlation similarity is not a good proxy for the reference Hamming similarity between sparse representations : the correlation should be computed with a unit hop size in order to select similar frames that are perfectly aligned, which would be computationally demanding and not necessary for the joint sparsity algorithm using a Fourier dictionary. On the contrary, the other two similarity measures are not sensitive to phase differences.

In order to select the best similarity measure, we compare the sets of frames selected by the reference method and by the proposed measures, for various ratios of missing data, using the same experimental conditions as in Section 4. Results are averaged for a speech and a piano examples and are shown in Table 2. They confirm that the correlation measure is not appropriate and that the spectral cosine similarity seems to be more efficient than the IS similarity for all the considered ratios of missing data. In the rest of the paper, we have selected the similarity based on the spectral cosine similarity.

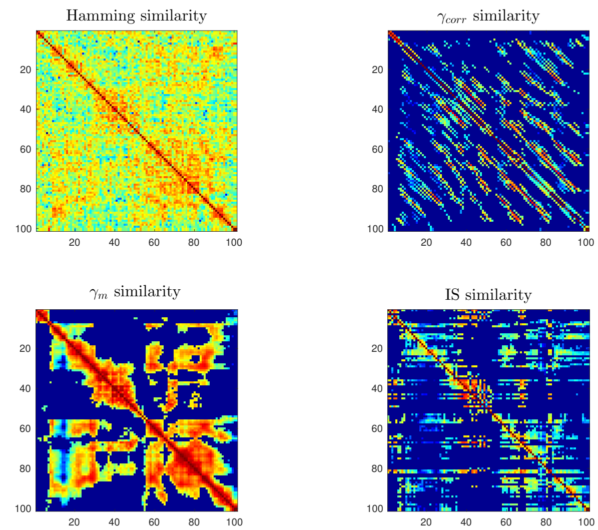

Fig. 2. Similarity maps for the reference frame selection method (Hamming similarity, top left) and the proposed measures.

Missing data (%)	0	20	40	60	80
γ_{corr}	32.6	30.8	29.3	28.8	27.6
γ_m	**39.4**	**36.1**	**33.7**	**31.2**	**29.3**
γ_{IS}	34.2	32.4	31.1	29.5	28.1

Table 2. Mean intersection between the sets of frames selected by the reference method and the proposed similarity measures (as a ratio of the total number of selected frames.), as a function of the ratio of missing data.

4. INPAINTING RESULTS

In this section, we show how the frame selection strategy combined with the joint-sparse estimation can solve an audio inpainting problem[1]. The performance is evaluated by computing the average signal-to-noise-ratio on all the reconstructed frames as defined in [11]:

- either on all the samples "$\mathrm{SNR}_{\text{full}}$".
- or only on the recovered samples "SNR_m".

The results are obtained for two different audio signals sampled at 8KHz: a melody piano and a male speech composed of one sentence. The duration of each of them is 4 seconds and they are segmented with an 8ms hop size in order to obtain about 500 frames

[1]The code and data to reproduce the experiments of this paper are available on `https://mad.lis-lab.fr/`.

with length 32ms (256 samples) in each sound. Performance results are average over those 500 frames. The inpainting S-OMP algorithm with a complex Fourier dictionary (S-OMP F) is compared to the inpainting OMP algorithm with two types of dictionaries: the same complex Fourier dictionary (OMP F) and the real Gabor dictionary (OMP G) used in [11]. All dictionaries are with size 256×512. The sparse estimation algorithms stop when 64 atoms are selected.

In Figure 3, we take a particular case where 50% of the samples are missing, by blocks of duration 0.25 ms. We plot the SNR_m values as a function of the number of selected frames $|\mathcal{S}|$. For this case, we can see that the good SNR values are obtained for a size $|\mathcal{S}|$ equal to or greater than 4, depending on the signal. For $|\mathcal{S}| = 1$, the inpainting S-OMP F is equivalent to the inpainting OMP F algorithm. On those examples, one can see that only a few frames need to be selected to obtain a performance improvement. Selecting too many frames should be avoided since it would cause a significant increase in the computational time and since the performance may slightly decrease, probably due to the selection of unsuitable frames. For the rest of the experiments, we set $|\mathcal{S}|$ to 4.

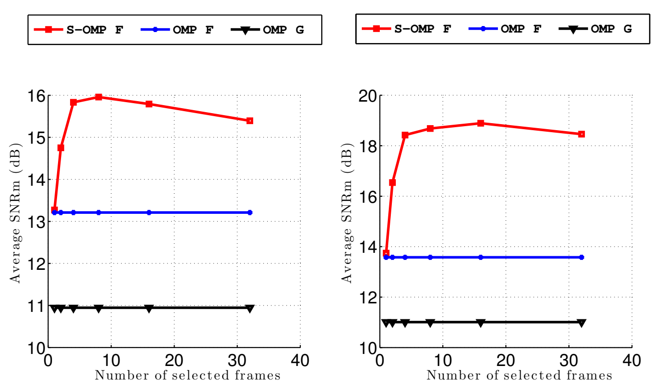

Fig. 3. Average SNR_m vs. number of selected frames $|\mathcal{S}|$ for the speech (left) and piano (right) examples.

We present inpainting results as a function of the ratio of missing data in Figure 4. One can see that the inpainting S-OMP based on the joint sparsity outperforms the two versions of the inpainting OMP when 30% to 70% of the data is missing but it is less efficient for extremely low or high ratios. The flat slope of the S-OMP curve in this $30-70\%$ area would suggest that the similar frames are useful to reach better performance in a larger range of problems, being more robust in adversarial conditions with many missing data, until a limit of 70% where the performance starts to drop here, probably due to more errors in the frame selection or the sparse estimation. For a very low ratio of missing data, the proposed method does not outperform the OMP approaches which can be explained as follows. First, the number of observation is high enough so that OMP estimates a good sparse decomposition. Second, S-OMP constrains the support of the selected frames to be equal, which may degrade the results if those supports are not exactly the same, as with real data. We also report that in those examples, 80 % of the selected similar frames are neighboring frames located at 4 frames or less from $\mathbf{s}_i$, while, depending on the signal, from 10 % to 15 % of the selected frames are at a distance larger than 20 frames.

To get a better idea of the performance of the algorithm, we increased the duration of the missing intervals, with a fixed ratio of 50% missing data. This means that for a 4 seconds signal, 2

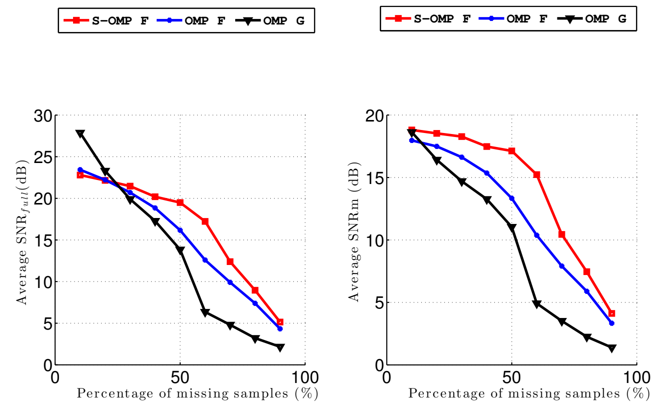

Fig. 4. Average SNR_{full} (left) and SNR_m (right) as a function of the percentage of missing samples.

seconds are missing which is considerably high. Results are given in Figure 5. The missing duration varies between 0.25 ms and 4 ms. The S-OMP inpainting algorithm outperforms both the state-of-the-art OMP G and OMP F between 0.25 and 2 ms and start to fail for very large intervals.

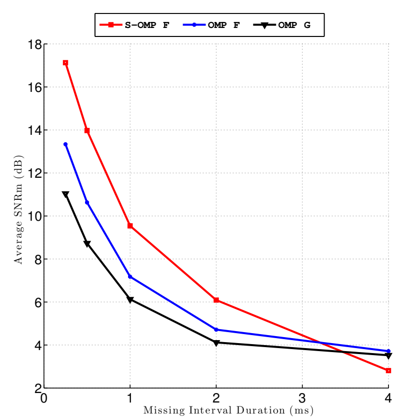

Fig. 5. Average SNR_m for varying missing intervals durations

5. CONCLUSION

In this work, we have proposed a non-local similarity joint-sparse modeling framework for audio inpainting. A focus has been dedicated to the choice of an appropriate similarity measure, showing that the correlation measure used in image processing should be avoided. Proof-of-concept experiments on a restricted set of signals have illustrated the behavior of the proposed method, showing that it can outperform purely-local sparse inpainting strategies in a larger range of difficulty in audio inpainting problems. They also suggest that only a few number of similar frames are necessary to improve performance.

Those results would now require a larger exploration of the use of non-local structures in sparse models for audio inpainting: comparing S-OMP and other joint-sparse algorithms [7, 8], on various problems including, e.g., declipping, and on more data.

6. REFERENCES

[1] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1 –11, 2009.

[2] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, June 2005, vol. 2, pp. 60–65.

[3] Y. Bahat, Y. Y. Schechner, and M. Elad, "Self-content-based audio inpainting," *Signal Processing*, vol. 111, no. 0, pp. 61 – 72, 2015.

[4] N. Perraudin, N. Holighaus, P. Majdak, and P. Balázs, "Similarity graphs for the concealment of long duration data loss in music," *arXiv:1607.06667*, 2016.

[5] J. Mairal, F. R. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *IEEE12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, 2009*, 2009, pp. 2272–2279.

[6] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation: Part i: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[7] S. Foucart, "Recovering jointly sparse vectors via hard thresholding pursuit," in *in Proc. of SAMPTA*. 2011, Online.

[8] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jing, "Greedy algorithms for joint sparse recovery," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1694–1704, April 2014.

[9] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, 1986.

[10] S. J. Godsill and P. J. W. Rayner, "A bayesian approach to the restoration of degraded audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 267–278, 1995.

[11] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M.D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 922 – 932, 2012.

[12] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Nov. 1993, pp. 40 –44 vol.1.

[13] K. Siedenburg, M. Kowalski, and M. Dorfler, "Audio declipping with social sparsity," in *ICASSP*. 2014, pp. 1577–1578, IEEE.

[14] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer Publishing Company, Incorporated, 1st edition, 2010.