STRUCTURED SPARSITY FOR AUDIO INPAINTING

Václav Mach
Doctoral Degree Programme (3), FEEC BUT
E-mail: vaclav.mach@phd.feec.vutbr.cz

Supervised by: Pavel Rajmic
E-mail: rajmic@feec.vutbr.cz

Abstract: Completing missing or distorted samples is a typical task of audio restoration engineers. State-of-the-art interpolation techniques utilized for this purpose are compromised today by novel methods based on sparse representations of signals. Preliminary results of sparse solutions are promising. In this paper, the topic of sparse solutions of underdetermined linear systems is described in a theoretical way. Further, the comparison of interpolation and sparse solutions based on ℓ_1-minimization is presented.

Keywords: Audio, Inpainting, Interpolation, Structured, Sparsity, ℓ_1-minimization.

1 INTRODUCTION

In many cases signal transmission is affected by errors in both analogue or digital ways. Typical example in sound signal processing is a distortion or signal loss on archive audio recordings, e.g. vax cylinders, magnetic tapes, gramophone records [1]. Another example of such a problem is packet loss in internet telephony (VoIP).

To solve this problem, various techniques have been utilized in the past. Interpolation techniques based on autoregressive (AR) modelling were developed [2] followed by wavelet transform [3] or neural network based methods [4].

Currently, signal processing methods referred as the *Sparse Representations* are increasingly popular for solving underdetermined linear equation systems. Solving the problem of missing samples in sound signal processing is called the Audio Inpainting [5].

This paper shows that sparse representations are able to improve the missing signal restoration problem in case of objective reconstruction assessment in contrast to state-of-the-art interpolation methods. Applying the structured sparsity in time-frequency (TF) domain is presented and current results are provided.

2 SPARSE SIGNAL REPRESENTATIONS

We have a set of vectors $\{\mathbf{d}_j\}$, $j = 0, 1, \ldots, K_\mathrm{D}$ called atoms. These atoms form together a non-orthogonal system called a frame. This frame has a meaning of a dictionary

$$\mathbf{D} \in \mathbb{R}^{N \times K_D}, \tag{1}$$

where N is a length of single atom and K_D is a number of atoms. The condition $N \leq K_\mathrm{D}$ must be preserved. Each atom is one 'word' from a dictionary. With this dictionary and appropriate sequence of coefficients $\mathbf{x}_i \in \mathbb{R}^{K_\mathrm{D}}$, the observed signal y_i is approximated

$$\widehat{\mathbf{y}}_i \approx \mathbf{D}\mathbf{x}_i \tag{2}$$

for each segment $i = \{0, 1, \ldots, M\}$. Sparsity means, that the vector $\mathbf{x}_i$ has only a few nonzero coefficients compared to N.

Sparse vector $\mathbf{x}_i$ from the original signal is obtained by solving an optimization problem. Generally there are two groups of solvers: *Greedy* algorithms and algorithms based on ℓ_1-*minimization* [6]. Audio signals are typically sparse in TF domain, therefore can be approximated by a few coefficients using proper dictionary. This was the main motivation for existing and future research in this field.

3 AUDIO INPAINTING

Audio Inpainting is based on approximating missing or distorted information with atoms $\{\mathbf{d}_j\}$ from the dictionary. At the begining the signal is segmented into segments of defined length same as the atom length N and appropriate overlap. For those segments where wrong samples are detected (either specified a-priori by the user or some detection method), individual inpainting is proceeded. Wrong samples are identified in a diagonal matrix $\mathbf{M}^{\mathrm{m}}$ consisting of zeros and ones. The missing samples (meaning rows in the dictionary) are represented as ones and reliable samples as zeros. Matrix $\mathbf{M}^{\mathrm{r}}$ represents the reliable samples matrix with values in opposite to the $\mathbf{M}^{\mathrm{m}}$ matrix.

Reliable samples can be obtained by

$$\mathbf{y}^{\mathrm{r}} = \mathbf{M}^{\mathrm{r}}\mathbf{D}\mathbf{x}, \tag{3}$$

while recovering of unknown samples $\widehat{\mathbf{y}}^{\mathrm{m}}$ is performed by estimating $\widehat{\mathbf{x}}$ as a sparse vector

$$\widehat{\mathbf{y}}^{\mathrm{m}} = \mathbf{M}^{\mathrm{m}}\mathbf{D}\widehat{\mathbf{x}}. \tag{4}$$

4 ℓ_1-MINIMIZATION

Natural solution of counting a number of non-zero coefficients is provided by ℓ_0 norm. However, ℓ_0 norm is not a convex function, thus it is not convenient to use any of available convex optimization algorithms. While ℓ_p norms are convex for $p \geq 1$, using the closest convex norm ℓ_1 is a natural attempt. In the case of noisy data, problem known as LASSO is defined as

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{D}\mathbf{x} - \mathbf{y}\|_2 \leq \delta, \tag{5}$$

where a correct solution error δ is permitted. In most cases, solutions of this problem using ℓ_0 and ℓ_1 norm are equal [6]. Problem (5) can be formulated in a general convex sense as unconstrained version

$$\widehat{\mathbf{y}} = \mathbf{D} \arg\min_{\mathbf{x} \in \mathbb{C}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_{p,p} \tag{6}$$

where $\lambda\|x\|_{p,p}$ is a regularization term which penalizes certain types of solutions and λ is a weighting coffecient controlling strength of the term. Different kind of sparsity or structure natural for real signals can be enforced choosing proper penalty. The minimizer of the convex functional (6) can be computed by proximal algorithms. Currently, the most efficient proximal algorithm is FISTA (Fast Iterative Shrinkage/Thresholding Algorithm) [7].

5 STRUCTURED SPARSITY

Every typical spectrogram of a musical signal is naturally structured. Considering this, algorithm for sparse signal modelling incorporating information about a structure (evaluation of the coefficient on the strength of its neighborhood) in an analysis stage of processing would be an advantage compared to the regular sparse modelling where coefficients are treated independently. While ℓ_1-norm performs

individually on each coefficient, mixed norms can substitute this norm to perform independently on a group of coefficients. Keeping or discarding particular coefficient under consideration is decided up to certain neighborhood of the coefficient. Further improvement called the *Social Sparsity* means weighting of the coefficients in the neighborhood [8].

The neighborhood should be chosen according to the specific signal class under investigation, e.g. focused on tonal/transient part. According to this, structured shrinkage operators representing a neighborhood system have to be defined. The convex optimization problem for Audio Inpainting with mixed norms is reformulated as

$$\widehat{\mathbf{y}}^{\mathrm{m}} = \mathbf{M}^{\mathrm{m}}\mathbf{D} \cdot \arg\min_{\mathbf{x}\in\mathbb{C}^N} \left(\frac{1}{2}\|\mathbf{M}^{\mathrm{r}}\mathbf{y} - \mathbf{M}^{\mathrm{r}}\mathbf{D}\mathbf{x}\|_2^2 + \lambda\|x\|_{p,q} \right) \tag{7}$$

where p represents a within-group penalty na q is across-group penalty.

Due to the non-stationariry of sound signals, windowing with overlapping and weighting is incorporated. After running some experiments, following neigborhoods were evaluated as the most promising:

1. *Windowed-Group-Lasso*: Keeping a coefficient if the energy of its neighborhood is large enough (positive correlation). $p = 2$, $q = 1$,
2. *Persistent-Elitist-Lasso*: Coefficient will be kept if its neighborhood is energetic enough compared to the others. $p = 1$, $q = 2$ and one more index for sophisticated treatment are utilized.

6 EXPERIMENTS

Illustration of structured sparsity algorithms performance applied to solving an Audio Inpainting problem is provided. Our experiments were performed using Matlab BWItoolbox[1] developed by Brno University of Technology and University of Vienna in a bilateral project, core engine of Structured Sparsity algorithms[2] is provided by University of Vienna.

First, different neighborhoods representing mixed norms were compared. Some of them provided unusable results (e.g. Persistent-Group-Lasso). Among several possibilities, two most promising results (presented in last section) are presented here in Fig. 1. Experiments were performed on a musical file *music_08* from the BWItoolbox with sampling frequency $f_{\mathrm{s}} = 16\,\mathrm{kHz}$, the signal gap beginning sample number 9528 with duration of 164 samples ($10.3\,\mathrm{ms}$).

[1] http://www.stud.feec.vutbr.cz/ xmachv00/bwitoolbox/bwitoolbox_v01.zip
[2] http://homepage.univie.ac.at/monika.doerfler/StrucAudioToolboxV02.rar

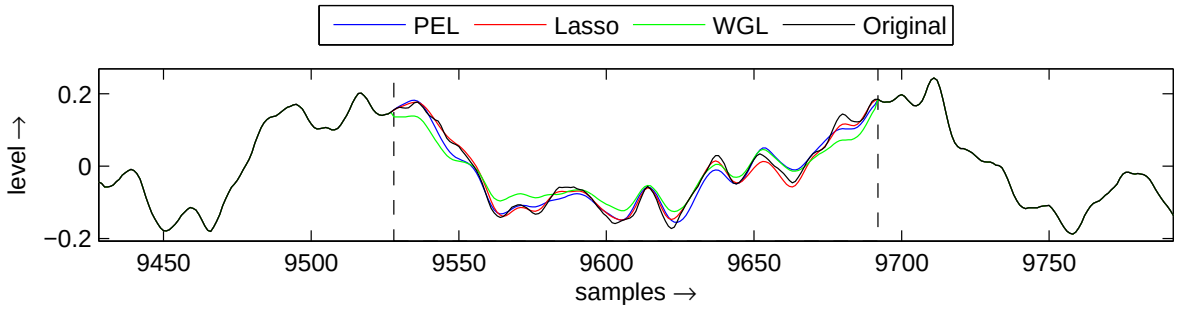

Figure 1: Comparison of Structured Sparsity methods: Persistent-Elitist-Lasso, Lasso (without persistence) and Windowed-Group-Lasso.

Table 1: Best parameters for each of structured sparsity methods

Method	Atom length (samples)	Time shift (samples)	Neighb. size (coefficients)	Neighb. size (ms)	Weighting center (coefficients)
PEL	4096	1024	5	512	3
WGL	1024	256	7	160	4

The parameters of structured sparsity experiments were selected according to complex parameters testing results for each of the methods (see Table 1), parameter $\lambda = 0.01$ and an overcomplete Gabor dictionary **D**. Pure Lasso method was performed with a non-persistent (no neighborhood) setup.

Regarding Fig. 1 with inpainting experiments, all of tested methods were succesfully aproximating the original signal. However, WGL method result produces non-smooth transition at the beginning and end of the missing gap, which in fact is not perceptually audible.

For comparison, two state-of-the-art algorithms were selected. First one an interpolation method called Weighted Forward-Backward Interpolation (wFB) presented in [2]. Another method is Orthogonal Matching Pursuit, a greedy method solving underdetermined linear systems. Both of them produce less satisfying results then the novel approach via structured sparsity, see Fig. 2.

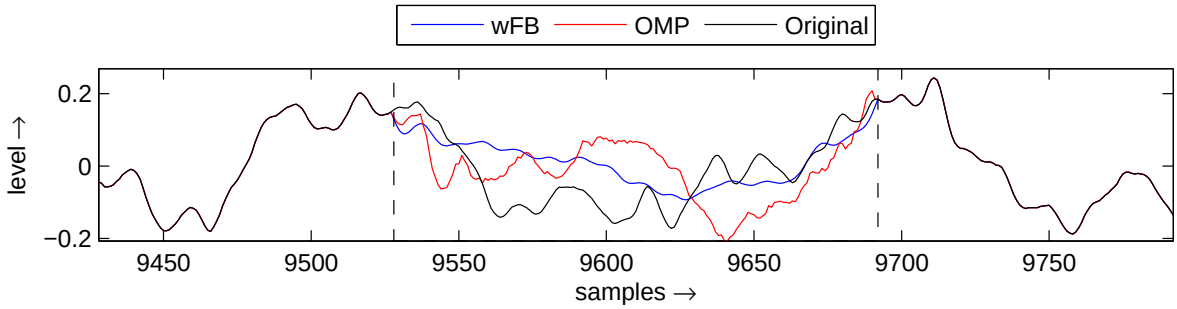

Figure 2: State-of-the-art interpolation methods: Audio Inpainting via Orthogonal Matching Pursuit and Weighted Forward-Backward Interpolation.

In Table 2 is presented an objective evaluation via SNR. The best result is reached using PEL Structured Sparsity method followed by pure Lasso algorithm. SNR results present comparison of a single gap reconstruction. In the sense of human sound perception the most disturbance was added by OMP algorithm. Comparing other methods, the inpainting of this short segment was almost inaudible.

Table 2: Objective evaluation of Inpanting methods via SNR

Method	SNR [dB]
wFB	2.33
OMP	–1.31
Lasso	19.50
PEL	20.70
WGL	13.88

7 CONCLUSION

This paper reviews the problem of missing audio signal interpolation and provides a novel approach for solving this problem using ℓ_1-minimization methods. Theoretical aspects like sparsity or inpainting issues statement were briefly described for basic understanding the problem as whole. Methods

based on ℓ_1-minimization outperform other state-of-the-art methods, however, more focus has to be concentrated on optimization of structed sparsity methods for proving the theoretical advantages of apriori information about the structure in the spectrogram. Another idea for future work is to perform complex testing on musical recordings of various genres while trying to fit the algorithm parameters to specific kinds of music.

ACKNOWLEDGEMENT

The research was performed in laboratories supported by the SIX project registration number CZ.1.05/2.1.00/03.0072, the specific junior interfaculty research grant no. FEKT/FSI-J-13-1903 and the international MOBILITY project reg. no. 7AMB13AT021.

REFERENCES

[1] V. Mach, "Digital restoration of recordings from the phonograph cylinders and their copies," in *As recorded by the phonograph: Slovak and Moravian songs recorded by Hynek Bím, Leoš Janáček and Františka Kyselková in 1909–1912*. Brno: The Institute of Ethnology of the Academy of Sciences of the Czech Republic, v.v.i., 2012, pp. 165–176.

[2] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1124–1135, 1996.

[3] P. Rajmic and J. Klimek, "Removing crackle from an LP record via wavelet analysis," in *Proceedings of the 7th international conference on digital audio effects DAFx04*, 2004, pp. 100–103. [Online]. Available: http://dafx04.na.infn.it/WebProc/Proc/P_038.pdf

[4] G. Cocchi and A. Uncini, "Subbands audio signal recovering using neural nonlinear prediction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 2, 2001, pp. 1289 –1292 vol.2.

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

[6] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[7] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[8] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators," *Signal Processing, IEEE Transactions on*, vol. 61, no. 10, pp. 2498–2511, 2013.