HYBRID MODEL AND STRUCTURED SPARSITY FOR UNDER-DETERMINED CONVOLUTIVE AUDIO SOURCE SEPARATION

Fangchen Feng and Matthieu Kowalski

CNRS-SUPELEC-Univ Paris-Sud
Gif-sur-Yvette, France
{fangchen.feng, matthieu.kowalski}@lss.supelec.fr

ABSTRACT

We consider the problem of extracting the source signals from an under-determined convolutive mixture, assuming known filters. We start from its formulation as a minimization of a convex functional, combining a classical ℓ_2 discrepancy term between the observed mixture and the one reconstructed from the estimated sources, and a sparse regularization term of source coefficients in a time-frequency domain. We then introduce a first kind of structure, using a hybrid model. Finally, we embed the previously introduced Windowed-Group-Lasso operator into the iterative thresholding/shrinkage algorithm, in order to take into account some structures inside each layers of time-frequency representations. Intensive numerical studies confirm the benefits of such an approach.

Index Terms— structured sparsity; audio source separation; convolutive mixture

1. INTRODUCTION

In many situations, such as a concert for music or the so called cocktail party problem for speech, the recorded sound signals are issued from mixtures of several sound sources. In this article, we consider the *reverberant under-determined* setting. The difficulty is then twofold: the number of sources is larger than the number of mixture channels, and the reverberation is modeled as a convolution. We focus on the estimation of the source signals assuming that the mixing filters are known. The blind separation problem in that case is still a challenging open problem [1].

In the under-setting case, source separation problem can be adressed using time-frequency masking techniques (see [2] and references therein for example). Moreover, in order to deal with the convolution, the short time frequency transform (STFT), or Gabor transform, allows to approximate the convolution by several instantaneous mixture, depending on the frequency band. In [3], the considered inverse problem is formulated as a convex optimization problem, where a wideband ℓ_2 mixture fitting cost is used directly in the time domain, in addition of a ℓ_1 source sparsity cost in the time-frequency domain. Such an idea has been exploited using a analsys prior in [4], which confirms the benefit of such an approach on various type of audio mixture, over the classical time-frequency masking.

This article provides three contributions. Firstly, motivated by the research about hybrid model for signal [5], also known as Morphological Component Analysis [6], we investigate the use of the union of two Gabor frames, each adapted to the "morphological layer", for the signal time-frequency representation in the problem of source separation. Secondly, we link the Windowed-Group-Lasso [7] to the problem of source separation to obtain a more reliable sparse representation. Windowed group-Lasso is a convenient way to take into account some neighborhood information for a *structured* sparse approximation. It is the first time, to our knowledge, that the structured sparsity is used in the problem of convolutive source separation. Finally, we compare the proposed methods with the state-of-the-art method and we thereby conclude the favorable conditions for speech and music sources.

The rest of the article is organized as follows. Next section 2 introduces the notation, the mathematical models and proposed algorithms. Section 3 presents all the experiments done on various speech and music mixtures, in order to show the benefit of both hybrid model and structured sparsity. The last section 4 concludes the paper.

2. MATHEMATICAL MODEL AND ALGORITHMS

After the introduction of the general mixture model, this section presents the wideband convex problem under consideration with the hybrid model. Then, the structured shrinkage operators are presented, as well as the practical algorithms for source separation.

We consider the source separation problem for convolutive mixtures of the form

$$x_m(t) = \sum_{n=1}^{N} A_{mn} \star s_n(t) + e_m(t) \,, \tag{1}$$

with N source signals s_n of duration T and M ($M < N$) microphones, yielding M mixture channels x_m, $\star$ denotes the convolution. The effect of acoustic propagation between the sources and the microphones is modeled by a set of mixing filters $A_{mn}(t)$ of length P. Denoting by $\mathbf{x} \in \mathbb{R}^{M\times T}$ and $\mathbf{s} \in \mathbb{R}^{N\times T}$ the matrices of mixture channels and source signals and by $\mathbf{A} \in \mathbb{R}^{M\times N\times P}$ the three-way array of mixing filters, the mixing process (1) can be rewritten more concisely in matrix form as

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{e} \,, \tag{2}$$

where $\mathbf{e} \in \mathbb{R}^{M\times T}$ models the background noise. Since $M < N$, $\mathbf{A}$ is not invertible, hence the need for suitable approaches to estimate $\mathbf{s}$ given $\mathbf{x}$ and $\mathbf{A}$.

Let us denote by $\mathbf{\Phi} \in \mathbb{C}^{T\times B}$ the matrix representing an energy-preserving STFT operator (or Parseval Gabor frame), the sources $\mathbf{s}$ can be resynthesized from their estimated STFT coefficients $\boldsymbol{\alpha} \in$

This work benefited from the support of the "FMJH Program Gaspard Monge in optimization and operation research", and from the support to this program from EDF.

$\mathbb{C}^{N\times B}$ by

$$\mathbf{s} = \boldsymbol{\alpha}\boldsymbol{\Phi}^* \tag{3}$$

where $\boldsymbol{\Phi}^* \in \mathbb{C}^{B\times T}$ is the adjoint operator of $\boldsymbol{\Phi}$, that is its Hermitian transpose.

2.1. Wideband Hybrid Lasso

One possible assumption is that the signals are sparse in the time-frequency domain [8, 9]. Under this assumption, we can recast the source separation problem into a convex optimization framework. This assumption relies on the following functional [3]

$$\min_{\boldsymbol{\alpha}\in\mathbb{C}^{N\times B}} \frac{1}{2}\|\mathbf{x} - \mathbf{A}\star\boldsymbol{\alpha}\Phi^*\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{4}$$

In [5] for audio, and in [6] for images, hybrid model or morphological component analysis suppose that a signal can be expressed as the sum of two layers: a tonal one and a transient. The underlying assumption can be expressed as the expectation that each class of components has a sparse decomposition within one, of the frames in the union. Then, a complementary approach to the model described in (4) is to consider a union of two frames or bases, each adapted to the "morphological layer". The hybrid model is given as follow:

$$\mathbf{s} = \mathbf{s}_{ton} + \mathbf{s}_{trans} = \boldsymbol{\alpha}_{ton}\boldsymbol{\Phi}_{ton} + \boldsymbol{\alpha}_{trans}\boldsymbol{\Phi}_{trans}$$

where $\boldsymbol{\Phi}_{ton} \in \mathbb{C}^{T\times B_{ton}}$ is a Gabor frame adapted for the tonal layer, and $\boldsymbol{\Phi}_{trans} \in \mathbb{C}^{T\times B_{trans}}$ adapted for the transient. The reader can refer to [10] for a more theoretical study of hybrid decompositions.

Given the model above, a natural way of circumventing the wideband Lasso (4) is to replace the decomposition using one Gabor frame by a decomposition using two Gabor frames

$$\begin{aligned}\min_{(\boldsymbol{\alpha}_{ton},\boldsymbol{\alpha}_{trans})} &\frac{1}{2}\|\mathbf{x} - \mathbf{A}\star(\boldsymbol{\alpha}_{ton}\Phi_{ton}^* + \boldsymbol{\alpha}_{trans}\Phi_{trans}^*)\|_2^2 \\ &+\lambda\left(\mu\|\boldsymbol{\alpha}_{ton}\|_1 + (1-\mu)\|\boldsymbol{\alpha}_{trans}\|_1\right)\end{aligned} \tag{5}$$

where $\lambda > 0$ is an hyperparameter balancing the data term and the regularizer, and $0 \leq \mu \leq 1$ is a hyperparameter balancing between the tonal and the transient layers.

Minimization of convex functions like (5) relies on the so-called proximity operator of convex penalties. The proximity operators typically lead to shrinkage/thresholding operator known as the soft-thresholding for the ℓ_1 norm.

Denoting α_{tf} the coefficient in each time-frequency bin, the soft-thresholding operator reads

$$\tilde{\alpha}_{tf} = \mathbb{S}_\lambda(\alpha_{tf}) = \alpha_{tf}\left(1 - \frac{\lambda}{|\alpha_{tf}|}\right)^+ \tag{6}$$

Then, one can minimize (5) thanks to Iterative Shrinkage/Thresholding Algorithm (ISTA). We provide the general form of its accelerated version (FISTA) [11] in Algorithm 1, where the data term $\mathcal{L}(\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{x} - \mathbf{A}\star(\boldsymbol{\alpha}_{ton}\Phi_{ton}^* + \boldsymbol{\alpha}_{trans}\Phi_{trans}^*)\|_2^2$ is L-Lipschitz differentiable with gradient

$$\nabla\mathcal{L}_{ton}(\boldsymbol{\alpha}) = [\mathbf{A}^\star\star(\mathbf{x} - \mathbf{A}\star(\boldsymbol{\alpha}_{ton}\Phi_{ton}^* + \boldsymbol{\alpha}_{trans}\Phi_{trans}^*))]\Phi_{ton} \tag{7}$$

$$\nabla\mathcal{L}_{trans}(\boldsymbol{\alpha}) = [\mathbf{A}^\star\star(\mathbf{x} - \mathbf{A}\star(\boldsymbol{\alpha}_{ton}\Phi_{ton}^* + \boldsymbol{\alpha}_{trans}\Phi_{trans}^*))]\Phi_{trans} \tag{8}$$

Algorithm 1: FISTA for solving (5)

Initialization: $\boldsymbol{\alpha}_{ton}^{(0)} \in \mathbb{C}^{N\times B_{ton}}$, $\boldsymbol{\alpha}_{trans}^{(0)} \in \mathbb{C}^{N\times B_{trans}}$, $\mathbf{z}_{ton}^{(0)} = \boldsymbol{\alpha}_{ton}^{(0)}$, $\mathbf{z}_{trans}^{(0)} = \boldsymbol{\alpha}_{trans}^{(0)}$, $\tau^{(0)} = 1$, $k = 1$.

repeat

$\boldsymbol{\alpha}_{ton}^{(k)} = \mathbb{S}_{\lambda/L}\left(\mathbf{z}_{ton}^{(k-1)} - \frac{\nabla\mathcal{L}_{ton}(\mathbf{z}_{ton}^{(k-1)}, \mathbf{z}_{trans}^{(k-1)})}{L}\right)$;

$\boldsymbol{\alpha}_{trans}^{(k)} = \mathbb{S}_{\lambda/L}\left(\mathbf{z}_{trans}^{(k-1)} - \frac{\nabla\mathcal{L}_{trans}(\mathbf{z}_{ton}^{(k-1)}, \mathbf{z}_{trans}^{(k-1)})}{L}\right)$;

$\tau^{(k)} = \frac{1+\sqrt{1+4\tau^{(k-1)2}}}{2}$;

$\mathbf{z}_{ton}^{(k)} = \boldsymbol{\alpha}_{ton}^{(k)} + \frac{\tau^{(k-1)}-1}{\tau^{(k)}}(\boldsymbol{\alpha}_{ton}^{(k)} - \boldsymbol{\alpha}_{ton}^{(k-1)})$;

$\mathbf{z}_{trans}^{(k)} = \boldsymbol{\alpha}_{trans}^{(k)} + \frac{\tau^{(k-1)}-1}{\tau^{(k)}}(\boldsymbol{\alpha}_{trans}^{(k)} - \boldsymbol{\alpha}_{trans}^{(k-1)})$;

$k = k + 1$

until *convergence*;

where the adjoint $\mathbf{A}^*$ of $\mathbf{A}$ is obtained by transposition of source and channel indexed and time reversal of the filters.

Introducing the linear operator $\mathcal{T} : \mathbb{C}^{N\times(B_{ton}+B_{trans})} \to \mathbb{R}^{M\times T}$ defined by

$$\mathcal{T}(\boldsymbol{\alpha}) = \mathcal{T}(\boldsymbol{\alpha}_{ton}, \boldsymbol{\alpha}_{trans}) = \mathbf{A}\star(\boldsymbol{\alpha}_{ton}\Phi_{ton}^* + \boldsymbol{\alpha}_{trans}\Phi_{trans}^*) \tag{9}$$

and $\mathcal{T}^*$ its adjoint, the Lipschitz constant L is given by

$$L = ||\mathcal{T}^*\mathcal{T}||_{2op} \tag{10}$$

with $||.||_{2op}$ denoting the operator norm. It can be well approximated thanks to a classical power iteration algorithm.

2.2. Structured shrinkage operator

When one looks at the time-frequency analysis coefficients of an audio signal, one can notice that there is a grouping effect of the coefficients in both time and frequency-direction. Then, one of the main limitations of the Lasso estimate is that all the coefficients are treated independently. However, the use of a group-Lasso penalty [12] is not directly possible on the time-frequency coefficients. Indeed, one cannot define independent groups as a prior. To avoid this, some authors have studied various kinds of *Group-Lasso with overlap*, such as in [13, 14]. However, the main practical limitations of such approaches is the computational cost. The strategy chosen here in order to obtain a more reliable sparse representation, is the use of new thresholding operators as the *Windowed-Group-Lasso* (WG-Lasso). Windowed-Group-Lasso was first defined in [15] and was deeper studied in [7]. The idea is to use the neighborhood information of a given coefficient inside the shrinkage operators, in order to exploit the time-frequency persistence properties. Using this neighborhood structure, WG-Lasso is defined by the following operator, for each time-frequency index (t, f):

$$\tilde{\alpha}_{tf} = \mathbb{S}_\lambda^{WGL}(\alpha_{tf}) = \alpha_{tf}\left(1 - \frac{\lambda}{\sqrt{\sum_{t',f'\in\mathcal{N}(t,f)}|\alpha_{t'f'}|^2}}\right)^+ \tag{11}$$

where $\mathcal{N}(t, f)$ denotes the time-frequency neighborhood of the time-frequency index (t, f). The idea of this shrinkage operator is to select a coefficient if the energy of its neighborhood is sufficiently large. Consequently, an isolated "big" coefficient can be discarded, but a "small" coefficient in the middle of big ones can be kept.

In the case of two Gabor frames, as the transform adapted for the tonal part is well localized in frequency and the transform adapted for the transient part is well localized in time, the structures in the time-frequency plan for the tonal part and the transient part seem different. Empirically, we choose the neighborhood extending in time for the tonal layer and the neighborhood extending in frequency for the transient layer.

3. EXPERIMENTS

After the presentation of the experimental setup, we describe in this section the influences of various choices for the parameters, as the size of the Gabor windows, the size of the neighborhood, and the benefit of the hybrid model.

3.1. Experimental setup

For all the experiments, the signals were sampled at 11 kHz, and the mixing filters were room impulse responses simulated via the image technique [16] using the Roomsim software [17]. The number of microphones and the number of sources were respectively set to 2 and 4. We provide results for the following configuration: the microphone spacing was set to $d = 1$m and the reveberation time was $RT_{60} = 250$ms for ten different sets of male and/or female speech sources from various nationalities and ten different sets of music sources (including singing voice and various instruments).

In all the experiments, the STFT was computed with half-overlapping tight windows using the ltfat toolbox[18]. The center of neighborhoods is always the considered sample for the windowed group-Lasso. In order to only evaluate the different methods in the light of the source separation efficiency, we did not add any simulated noise. In order to avoid complex evaluation of the hyper-parameters λ and μ in (5), we choose the most "natural" setting, i.e. $\lambda \to 0$[1] in order to obtain a perfect reconstruction of the mixture (and then, do not performing any denoising) and $\mu = 0.5$ in order to not favor a specific layer. One can surely improve the results by playing with these hyperparameters, but the price is a very expensive computational cost.

The separation performance was assessed using the now popular Signal to Distorsion Ratio (SDR) and Signal to Interference Ratio (SIR) [19]. The SDR indicates the overall quality of each estimated source compared to the target, while the SIR reveals the amount of residual crosstalk from the other sources. A larger value of SDR/SIR means a better quality of the separation. These measures were subsequently averages over all sources for each mixing condition. The wideband Lasso method was performed as a baseline.

One can listen some sound examples, for both speech and music demixing, on the webpage `http://webpages.lss.supelec.fr/perso/matthieu.kowalski/FK_icassp14/FK_icassp.html`.

3.2. Mono layer model

We first illustrate the benefit of structured sparsity over the simple wideband Lasso, i.e. in the case where only one Gabor dictionary is used. Then, it remains mainly two parameters to influence the quality of the separation: the size of the Gabor window and the size of the neighborhood.

The results are summarized in Figures 1 and 2, where the variations of the SDR and SIR are plotted as a function of the size of the neighborhood, for various size for the window. We recall that if the size of the neighborhood is 1, then WG-Lasso becomes Lasso. Notice that the SDR of the Lasso for the *speech sources* is maximum (7.9dB) when the Gabor window is 512 samples and the maximum (8.6dB) of the WG-Lasso is achieved with the same window length when the size of the neighborhood is 3. For the *music sources*, the maximum SDR (6.1dB) of Lasso method is achieved when the window is 1024 samples and the maximum (6.7dB) of WG-Lasso is reached with a window of 512 samples. However, the difference between the performance (SDR and SIR) of WG-Lasso with the window of 512 samples and 1024 samples is not significant. It is noticed that the algorithm is relatively robust with respect to the choice of the neighborhood: there is a significant increase in performance from the Lasso to WG-Lasso with the neighborhood of 3, but further enlarging the neighborhood does not improve the performance which suggests a quite robust choice. One of the most interesting thing, is that the same remarks apply for the SIR which is also improved.

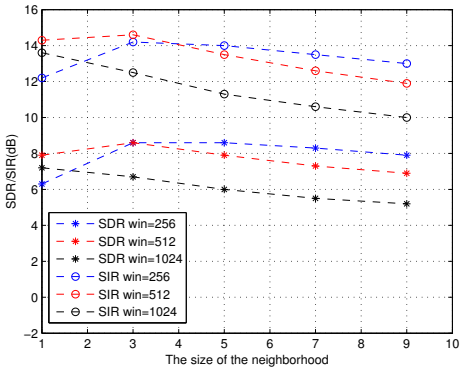

Fig. 1. Different size of neighborhood for WG-Lasso with speech source

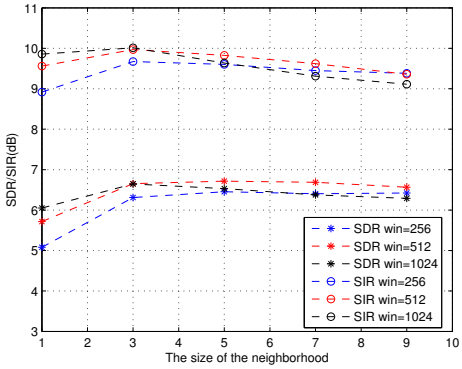

Fig. 2. Different size of neighborhood for WG-Lasso with music source

In conlusion, in the case of one Gabor frame, the best performance is realised when the Gabor window is 512 and the size of neighborhood is 3 for both speech and music mixtures.

3.3. Hybrid model

We present here the results for the hybrid model. In this case, the choices for the parameters are more tricky: one can play on the size of the windows for the two layers, as well as for the size of the neighborhood.

[1] $\lambda \to 0$ is not equivalent to set $\lambda = 0$, as one cannot invert the limit and the maximum operator.

3.3.1. Size of Gabor windows for Lasso with two Gabor frames

We first evaluate the performances without structured sparsity. In that case, the size of the Gabor window varies from 2^8 to 2^{11} for the tonal part and 2^5 to 2^8 for the transient layer. Table 1 and 2 illustrate the variation of the SDR and SIR. As shown in the tables, the best condition for the speech sources is a window of length 512 for tonal and 32 for transients; and for the music sources a window of length 2048 for tonal and 256 for transient. Besides, it is noticed that it is relatively robust with respect to the choice of the window for the transient part for both types of source. For the speech source, the trends observed when considering the choice of the window for tonal part is similar to the trends when one Gabor frame is used, but similar trends do not appear for the music source.

However, the use of hybrid model with the chosen specific choice for the hyperparameters, does not improve the quality of the separation compared to the single layer model: if the SDR is slightly improved by 0.1 to 0.2 dB, the SIR is degraded about 0.5 dB.

Table 1. SDR/SIR OF DIFFERENT SIZE OF WINDOWS FOR THE SPEECH SOURCES

		Size of window for tonal part			
		256	512	1024	2048
Size of window for transient part	32	6.5/11.9	**8.0/13.9**	7.7/13.5	6.3/12.0
	64	6.2/11.7	7.8/13.7	7.5/13.3	6.2/11.8
	128	6.0/11.4	7.7/13.6	7.4/12.3	6.2/12.0
	256	6.3/12.2	7.7/13.7	7.8/13.9	7.1/13.1

Table 2. SDR/SIR OF DIFFERENT SIZE OF WINDOWS FOR THE MUSIC SOURCES

		Size of window for tonal part			
		256	512	1024	2048
Size of window for transient part	32	4.6/8.0	5.3/8.5	5.8/8.9	5.9/8.9
	64	4.6/8.0	5.4/8.6	5.9/9.1	6.0/9.1
	128	4.9/8.2	5.4/8.7	6.0/9.2	6.2/9.3
	256	5.1/9.0	5.4/8.7	6.0/9.3	**6.2/9.3**

3.3.2. Size of neighborhoods for WG-Lasso in the case of two Gabor frames

In the case of two Gabor frames, we first set the size of the Gabor window for the tonal layer to 512 samples and 32 for the transient layer for the speech sources. The size of the neighborhoods for both layers varies from 1 to 9. The performance is shown in the upper tabular of Table 3. The best performance is reached when the neighborhood is 3 for the tonal layer and 5 for the transient. Moreover, it can be seen that, besides improving the SDR and SIR, the algorithm is also robust with respect to the choice of the neighborhoods when the neighborhoods are between 3 and 5.

The performances were also evaluated for a window of length 256 sample for the tonal layer, still with a window of 32 samples for the transient layer. As shown in the second tabular of Table 3, althought the maximum is achieved when the neighborhoods are (5,9), similar trends as in the previous setting can be observed. For the music sources, we set the sizes of the Gabor windows to (2048,256) and (512,32) respectively. The performances are illustrated in Table 4.

In conlusion, the best performance for speech source in the case of two Gabor frames is achieved when the Gabor windows are (512,32), and (2048,256) for the music source. The neighborhoods between 3 and 5 seem to be favorable for both.

We can summarized all these results in Table 5, which shows the variation of SDR/SIR for both Lasso and WG-Lasso in the case of two Gabor frames and one Gabor frame. One can remark that the two Gabor frames outperforms the method of one Gabor for both Lasso and WG-Lasso. It is interesting that the performances are improved by more than 1 dB on SDR without degrading the SIR for both speech and music, thanks to the structured hybrid model for speech sources.

Table 3. SDR/SIR OF DIFFERENT SIZE OF NEIGHBORHOOD FOR THE SPEECH SOURCE

Speech, windows 512-32		Size of neighborhood for tonal part			
		1	3	5	9
Size of neighborhood for transient part	1	8.0/13.9	8.4/14.0	7.7/13.1	6.7/11.9
	3	8.2/14.2	9.0/14.7	8.4/13.9	7.7/12.8
	5	8.2/14.2	**9.0/14.8**	8.5/14.0	7.8/12.9
	9	8.1/14.2	9.0/14.8	8.4/13.9	7.7/12.8

Speech, windows 256-32		Size of neighborhood for tonal part			
		1	3	5	9
Size of neighborhood for transient part	1	6.5/11.9	7.9/13.2	7.7/13.0	7.0/12.2
	3	6.7/12.2	8.6/14.0	8.6/13.9	7.8/12.8
	5	6.6/12.2	8.8/14.2	8.8/14.0	8.3/13.3
	9	6.6/12.2	8.7/14.1	**9.1/14.5**	8.4/13.4

Table 4. SDR/SIR OF DIFFERENT SIZE OF NEIGHBORHOOD FOR THE MUSIC SOURCE

Music, window 2048-256		Size of neighborhood for tonal part			
		1	3	5	9
Size of neighborhood for transient part	1	6.2/9.3	6.7/9.6	6.7/9.5	6.5/9.3
	3	6.4/9.5	7.1/9.9	**7.2/10.0**	7.1/9.8
	5	6.3/9.4	7.0/9.9	7.1/9.9	7.1/9.7
	9	6.1/9.3	6.8/9.7	7.0/9.7	7.0/9.6

Music, window 512-32		Size of neighborhood for tonal part			
		1	3	5	9
Size of neighborhood for transient part	1	5.3/8.5	5.9/8.9	6.0/8.8	5.9/8.6
	3	5.4/8.7	6.3/9.3	6.4/9.3	6.4/9.1
	5	5.5/8.8	6.4/9.4	6.6/9.4	6.6/9.2
	9	5.4/8.7	6.4/9.4	**6.6/9.3**	6.6/9.1

Table 5. SDR/SIR: Two Gabors vs One Gabor

	Lasso +1Gabor	Lasso +2Gabors	WG-Lasso +1Gabor	WG-Lasso +2Gabors
Speech	7.9/14.3	8.0/13.9	8.6/14.6	**9.1/14.5**
Music	6.0/9.9	6.2/9.3	6.7/10.0	**7.2/10.0**

4. CONCLUSION

In this paper we developed several iterative algorithms to separate convolutive mixtures using sparse source models in a time-frequency dictionary, when the mixing filter system is supposed to be known. We showed that these approaches give interesting results compared to wideband Lasso by improving SDR, with a stable SIR. We only displayed the results for the specific setting $RT_{60} = 250ms$ for a distance between the two micro of 1 m, as it is the most favorable case for the wideband Lasso [3]. However, we also run experiments on the setting $RT_{60} = 50ms$, and we have observed very similar behavior, for the same magnitude of improvement on SDR, but also a similar improvement on the SIR.

The next step should be to consider the problem of blind source separation for underdetermined convolutive mixture, but the study performed in [1] suggests a very difficult and challenging problem.

5. REFERENCES

[1] A. Benichoux, E. Vincent, and R. Gribonval, "A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors," in *ICASSP - 38th International Conference on Acoustics, Speech, and Signal Processing - 2013*, Vancouver, Canada, Mar. 2013.

[2] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, August 2007.

[3] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio Speech and Language Processing, Special Issue on: "Processing Reverberant"*, vol. 17, no. 7, pp. 1818–1829, 2010.

[4] S. Arberet, P. Vandergheynst, R. Carrillo, J. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 21, no. 7, pp. 1391–1402, Jul. 2013.

[5] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Procesing*, vol. 82, no. 11, pp. 1595–1617, 2002.

[6] M. Elad, J.-L. Starck, D. L. Donoho, and P. Querre, "Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)," *Journal on Applied and Computational Harmonic Analysis*, vol. 19, pp. 340–358, November 2005.

[7] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE transactions on signal processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

[8] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[9] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, pp. 2353–2362, 2001.

[10] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[11] 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.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Serie B*, vol. 68, no. 1, pp. 49–67, 2006.

[13] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *ICML*, 2009.

[14] I. Bayram, "Mixed norms with overlapping groups as signal priors," in *Proc. International Conference on Audio Speech and Signal Processing (ICASSP)*, May 2011.

[15] M. Kowalski and B. Torrésani, "Structured sparsity: from mixed norms to structured shrinkage," in *Proceeding of Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, April 2009.

[16] U. Svensson and U. Kristiansen, "Computational modelling and simulation of acoustic spaces," in *Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio*, 2002, pp. 1–20.

[17] "Roomsim," http://media.paisley.ac.uk/~campbell/Roomsim/.

[18] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 4, June 2012.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.