AUDIO SOURCE SEPARATION INFORMED BY REDUNDANCY WITH GREEDY MULTISCALE DECOMPOSITIONS

Manuel Moussallam, Gaël Richard, Laurent Daudet

AUDIO SOURCE SEPARATION INFORMED BY REDUNDANCY WITH GREEDY MULTISCALE DECOMPOSITIONS

Manuel Moussallam[1,2] *, Gaël Richard*[1] *, Laurent Daudet*[2] *

[1]Institut Mines-Telecom - Telecom ParisTech
CNRS/LTCI - UMR 5141

[2]Institut Langevin - ESPCI ParisTech
Paris Diderot Univ. - UMR 7587

ABSTRACT

This paper describes a greedy algorithm for audio source separation of repeated musical patterns. The problem is understood as retrieving from a set of mixtures the part that is redundant among them and the parts that are specific to only one mixture. The key assumption is the sparsity of all the sources in the same multiscale dictionary. Synthetic and real life examples of source separation of hand cut repeated musical patterns are exposed. Results shows that the proposed method succeeds in simultaneously providing a sparse approximant of the mixtures and a separation of the sources.

Index Terms— Simultaneous sparse approximation; audio source separation; greedy decompositions

1. INTRODUCTION

There are at least two specific cases of audio source separation problems where redundancy plays a fundamental role. *Common signal separation* is a problem where, from a set of mixtures, one tries to recover a source that is shared among all of them. Practical applications range from film music extraction [1] to multichannel denoising [2, 3]. *Repeating pattern separation* [4], focuses on separating a varying component (e.g. the singing voice) from a repeating background (e.g. musical accompaniment).

These two separation problems can be linked because they share the same underlying source model. A mixture X_i indexed by i, is understood as a combination of an *individual* source P_i, specific to the mixture, and a source component X_c that is *shared* among all the mixtures (though potentially distorted in a different manner in each mixture).

In the common signal separation problem, the individual sources are often considered as noise and the shared component is the signal of interest. Redundancy in this case is the result of a multisensor acquisition [3] or the existence of multiple versions [1]. In the repeating pattern separation framework, the shared source is the musical background (e.g accompaniment) that remains stable while occurring several times in the music and the individual sources are the parts varying among these occurrences (e.g solo instrument, singing voice). Redundancy is here the consequence of musical repetitions.

State of the art methods addressing the problem of *Repeating pattern separation* (e.g the REPET algorithm [4, 5]) are based on element-wise classification of a Time-Frequency (TF) representation. A TF mask (usually based on the power spectral density of the mixtures) is constructed for the repeating musical background and the separation is performed by means of Wiener filtering relative to this mask. Often (e.g in [5]) an assumption is made on the individual sources, namely that they are *sparse* in the TF domain. In the same spirit, in [6], the authors also consider the individual sources to be sparse while the shared component is captured in a low-rank approximant of the spectrogram, a matrix factorization scheme known as Robust PCA [7].

Interestingly, the same sparsity hypothesis is also at the core of methods addressing the *Common signal separation* problem. The basic assumption (e.g in Sparse Component Analysis (SCA) [8], or in simultaneous approximation problems [2, 3]) is that the shared component has a sparse expansion in a dictionary Φ of waveforms called atoms.

In this work, we address the *Repeating pattern separation* problem using a sparse decomposition of the mixtures in a redundant dictionary. However, we consider that the shared source and the individual ones are no different in nature, and thus may all be sparsely decomposed in the *same* dictionary.

Section 2 details the problem formulation and the sparse source models adopted. Section 3 introduces the greedy algorithm proposed in this work. Section 4 presents a comparison of behavior with TF based methods on synthetic and real-life examples. Finally Section 5 exposes the proposed separation scheme as a byproduct of a more general compression system.

2. SIMULTANEOUS APPROXIMATION PROBLEM

Let us formulate the source separation problem as a simultaneous approximation paradigm. Indeed, the separation is obtained from jointly estimating both the shared source and the individual ones.

This work was partly supported by the QUAERO Programme, funded by OSEO, French State agency for innovation.

LD is on a joint position between Univ. Paris Diderot and Institut Universitaire de France

2.1. General formulation

Let us now denote $\mathbf{X} \in \mathbb{R}^{I \times N}$ the matrix of I mixtures $X_i \in \mathbb{R}^N$ each being of dimension N, and $\mathbf{\Phi} \in \mathbb{R}^{D \times N}$ an overcomplete dictionary of D unit-normed waveforms called *atoms*. An approximant $\widetilde{\mathbf{X}}$ of $\mathbf{X}$ on $\mathbf{\Phi}$ is of the form: $\widetilde{\mathbf{X}} = \mathbf{C_X} \cdot \mathbf{\Phi}$ where $\mathbf{C_X} \in \mathbb{R}^{I \times D}$ is sparse, meaning a large part of its values are zeros. The simultaneous approximation problem consists in jointly minimizing the divergence between data and approximant, and the number of non-zero elements. One formulation is:

$$\min \|\mathbf{C_X}\|_0 \text{ s.t } f(\mathbf{X} - \mathbf{C_X} \cdot \mathbf{\Phi}) \leq \epsilon \quad (1)$$

where f is a divergence measure of interest (e.g. a squared reconstruction error), $\|.\|_0$ is the l_0 pseudo-norm[1] and ϵ is a desired level of precision. Since a strict l_0 problem is NP-hard to solve, a commonly adopted reformulation is a penalized version:

$$\widehat{\mathbf{C_X}} = \arg\min f(\mathbf{X} - \mathbf{C_X} \cdot \mathbf{\Phi}) + \lambda \cdot \|\mathbf{C_X}\|_{p,q} \quad (2)$$

where $\|.\|_{p,q}$ is a mixed norm[2]. p and q can be chosen depending on the desired sparsity and λ is a parameter that controls the weight of the sparsity constraint. It has been shown [9] that mixed norms can enforce structured sparsity. Actually, a column of $\mathbf{C_X}$ filled with non-zero elements indicates that the corresponding atom can be found in all the mixtures and thus belongs to the shared source.

While convex optimization algorithms have been proposed along with structured sparsity priors [8, 9], greedy methods solving this problem are variants of Simultaneous Orthogonal Matching Pursuit (SOMP) [2]. This formulation is adapted when one tries to recover a shared component that is sparse in the dictionary and implicitly makes the assumptions that components from the individual sources will not be selected. In this context, the separation can be explained as a denoising of a multichannel signal based on inter-channel redundancies.

2.2. Distinguishing two different sparsities

In some situations, including music source separation, the previous formulation is not fully satisfactory. While a separation of the background is still desirable, the assumption that the individual sources cannot be sparsely represented in the same dictionary as the shared one does not hold any more. Without any knowledge of the sources characteristics or production mechanism (e.g source/filter modeling for singing voice) there is no reason to consider the shared and the individual components to be of a different kind.

Actually, it has been shown [10] that most musical signals are efficiently and sparsely decomposed in Fourier-based dictionaries (e.g. Gabor frames).

Although the shared source and the individual ones can be sparsely decomposed in the same $\mathbf{\Phi}$, atoms used to represent them will exhibit different *kinds* of sparsity. In a recent paper, [7] surveillance video frames were modeled as the sum of a low-rank and a sparse matrix. In a similar fashion, we can decompose $\mathbf{C_X}$ in a sum of two components: $\mathbf{C_X} = \mathbf{B_X} + \mathbf{P_X}$ where $\mathbf{B_X}$ is a *structured* sparse matrix and $\mathbf{P_X}$ an *unstructured* sparse matrix. $\mathbf{B_X}$ has a small number of columns of non zero elements. Each column indicates an atom that is spread among all mixtures. $\mathbf{P_X}$ on the opposite, contains at least one zero per column. Its non-zero elements denote atoms that only belong to a subset of mixtures, hence to a subset of individual sources.

The interest of such model is obvious for source separation, the shared source can be modeled as $X_c = \sum \mathbf{B_X} \cdot \mathbf{\Phi}$ and the individual sources are the rows in the product $\mathbf{P_X} \cdot \mathbf{\Phi}$. We can rewrite the problem so as to take this two-sparsities model into account:

$$\widehat{\mathbf{C_X}} = \arg\min f(\mathbf{X} - \mathbf{C_X} \cdot \mathbf{\Phi}) + \lambda \cdot \|\mathbf{B_X}\|_{p,q} + \gamma \cdot \|\mathbf{P_X}\|_{p',q'} \quad (3)$$

which allows to put different constraints (by means of λ, p, q and γ, p', q') on the matrices according to the desired sparsities for $\mathbf{B_X}$ and $\mathbf{P_X}$. We could have designed a pseudo-convex optimization algorithm to specifically solve this problem (in the spirit of the Principal Component Pursuit proposed in [7]), however these algorithms are computationally intensive and memory consuming. In order to process real scale audio data, we propose a simple greedy algorithm.

3. JOINTLY ADAPTIVE MATCHING PURSUIT

We propose the use of a fast greedy algorithm of the Matching Pursuit [11] family to find (potentially suboptimal) solutions to (3). The separation could be addressed in a post-processing step, for example by clustering the selected atoms according to their projections across mixtures. However, we have found that much better results can be obtained when the separation process is integrated in the greedy algorithm. This integration takes the form of two modifications of the basic algorithm. These changes are: i) the atom selection criterion has been changed, and ii) after an atom is chosen, a decision mechanism is added, attributing it either to the shared source or to the (or multiple) individual sources.

3.1. Structure

A matrix of residuals is initialized from the matrix of mixtures $\mathbf{R}^0 = \mathbf{X}$. The algorithm iteratively builds the two matrices $\mathbf{B_X}$ and $\mathbf{P_X}$ by selecting an atom in $\mathbf{\Phi}$ according to a criterion $\mathcal{C}(\mathbf{\Phi}, \mathbf{R}^n)$. Then a decision is taken whether to attribute the selected atom to the shared source or to a subset of individual sources.

[1] The l_0 pseudo-norm $\|\mathbf{X}\|_0$ counts the number of nonzero entries of $\mathbf{X}$.

[2] See [9] for proper definition

Algorithm 1 Jointly Adaptive Matching Pursuit (JAMP)

Input: $\mathbf{X}$, $\mathbf{\Phi}$

$\mathbf{R}^0 := \mathbf{X}$, $n = 0$

repeat

  Step 1 : Select atom $\phi_k \leftarrow \mathcal{C}(\mathbf{\Phi}, \mathbf{R}^n)$

  Step 2 : Decide if ϕ_k is background or not

  if ϕ_k is background **then**

    $\forall i, \mathbf{B}_\mathbf{X}[i,k] = \langle \phi_k, R_i^n \rangle$

  else

    Find which channels $J \subset I$, ϕ_k belongs to.

    $\forall j \in J, \mathbf{P}_\mathbf{X}[j,k] = \langle \phi_k, R_j^n \rangle$

  end if

  Step 3 : Update residual :

  $\mathbf{R}^n = \mathbf{X} - (\mathbf{B}_\mathbf{X} \cdot \mathbf{\Phi} + \mathbf{P}_\mathbf{X} \cdot \mathbf{\Phi})$

  $n \leftarrow n + 1$

until a stopping condition is met

Output: $\mathbf{R}^n$, $\mathbf{B}_\mathbf{X}$ and $\mathbf{P}_\mathbf{X}$

3.2. STEP 1 : Atom Selection

For the sake of clarity, we denote $r_i^n(\phi)$ the squared absolute value of the projection of an atom ϕ onto R_i^n, the residual of the i-th mixture at the n-th iteration: i.e $r_i^n(\phi) = |\langle R_i^n, \phi \rangle|^2$. Four criteria have been investigated in this work:

$$\mathcal{C}_S(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} \sum_{i=0}^{I-1} r_i^n(\phi)$$

$$\mathcal{C}_M(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} \min_i r_i^n(\phi)$$

$$\mathcal{C}_W(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} w(\phi, \mathbf{R}^n) \cdot \sum_{i=0}^{I-1} r_i^n(\phi)$$

$$\mathcal{C}_P(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} \sum_{i=0}^{I-1} r_i^n(\phi) + \sum_{i \neq j} |r_i^n(\phi) - r_j^n(\phi)|$$

$\mathcal{C}_S$ is simply an energetic criterion, it does not influence the choice of an atom from the background or the foreground. $\mathcal{C}_M$ is a criterion that minimizes the risk of selecting an atom not belonging to the background. $\mathcal{C}_W$ is a weighted variant, the weight being defined by the spectral flatness of the distribution of the atom projections on the I residuals. This flatness is the ratio of the geometric mean over the arithmetic mean. This criterion penalizes the selection of an atom if it does not belong to the background. Finally $\mathcal{C}_P$ encourages the selection of atom from the individual sources by adding the inter-channel atom projection differences to the plain energetic criterion.

3.3. STEP 2 : Separating the sources

The decision making obviously depends on the chosen criterion. Using $\mathcal{C}_M$, no atoms from the individual sources should be selected (at least until the background has been approximated to a good precision), thus no specific mechanism is required. Using any other criterion, on the other hand, forces us to add an additional step.

	Background		Foreground	
	SDR	SIR	SDR	SIR
$\mathcal{C}_S$	6.0 (1.4)	35.8 (8.6)	6.2 (1.5)	17.6 (4.6)
$\mathcal{C}_M$	1.2 (0.9)	16.9 (6.4)	1.2 (1.6)	2.8 (2.0)
$\mathcal{C}_W$	6.0 (1.4)	34.2 (7.7)	6.6 (1.8)	17.3 (4.8)
$\mathcal{C}_P$	7.8 (1.8)	35.0 (6.3)	7.1 (1.7)	20.4 (5.5)

Table 1. Separation scores (mean and std) after 1000 iterations of JAMP with various selection criteria

Let ϕ_k be the chosen atom, the distribution of the $r_i^n(\phi_k)$ is informative. If ϕ_k is efficient in representing the background, then this distribution will be flat (e.g. the $r_i^n(\phi_k)$ have small empirical variance). On the opposite, if there are great disparities in the values of $r_i^n(\phi_k)$, then one can assume that ϕ_k should not be assigned to the background, but to a subset (potentially only one) of the individual sources.

Any statistical measure of the dispersion of the $r_i^n(\phi_k)$ values can thus be used. In this work we have used a simple relative standard deviation $D = \frac{\sigma}{\mu}$. This value is low when the dispersion is weak, thus a threshold τ can be defined so as to make a decision on the atom appertaining to the background or the individual sources. In this work, we have set $\tau = 0.5$ and have not tried to optimize this parameter.

4. EXPERIMENTS

4.1. Comparing Selection criteria

To evaluate separation performances of the various selection criteria we have designed the following experiment. 4 short audio excerpt (5 seconds) were used to create 12 sets of 3 mixtures. For each set, one of the excerpts is used as the background source and is present in all the mixtures without distortion. Three different Individual-to-Shared source energy ratios were used, namely 5, 0 and -5 dB, so that a variety of mixing situations are tested. Performance is assessed by means of the widely adopted measures presented in [12]. Table 1 gives the results in terms of Sources-to-distortion ratio (SDR) and Sources-to-Interferences ratio (SIR).

The $\mathcal{C}_M$ criteria is used without any decision mechanism, the foreground sources are being estimated directly from the residual since only atoms from the background should be selected by the algorithm. We can see that this method gives substantially lower results. Interestingly the best technique appears to be using $\mathcal{C}_P$.

4.2. Comparing with Time-Frequency based Techniques

In [5], the authors have presented a simple separation technique based on an estimation of a Time-Frequency mask for the background source as the median (respectively the minimum) of the mixture spectrogram. We have implemented this method, labeled REPET-Median (resp. REPET-Min) and

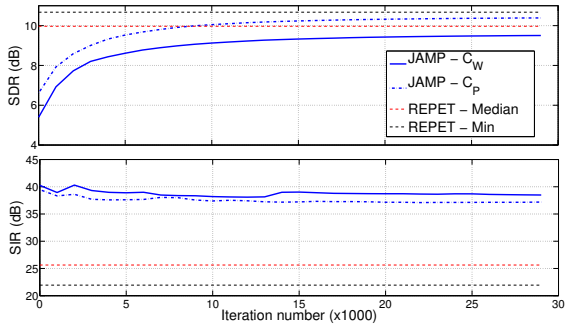

Fig. 1. BSS eval mean score for synthetic examples. Comparing TF masking technique from [5] with two Joint Matching Pursuit with criteria Weighted and Penalized

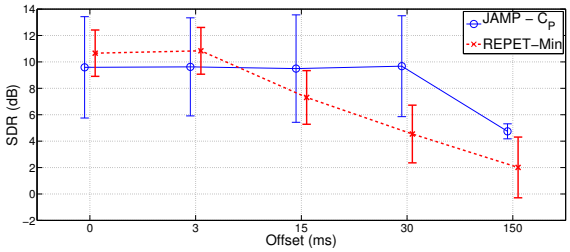

Fig. 2. BSS eval mean score for synthetic examples. Comparing TF masking technique from [5] with Joint Matching Pursuit with criterion Penalized for various offsets in background alignments

compared it to our own. Since JAMP is iterative, we can follow the evolution of the separation scores through the decomposition process. Figure 1 presents such results for analyzing mean performances on the same set of signals as above. JAMP with a $\mathcal{C}_P$ selection criterion can reach the same SDR level than REPET-Median in about 10000 iterations. In average, REPET-Min gives better SDR results, however the mean SIR values are much better using JAMP.

Additionally, the JAMP algorithm is designed to be more robust to the backgrounds being offset in the mixtures. Actually, a local optimization of atoms time localization for each mixture is performed. It effectively manages to reduce pre-echo artifacts [10]. Figure 2 presents the results of an experiment in which the background sources are offset in each mixture. Performances of the REPET-Min algorithm drops quite sharply for offsets of about 15ms, while they remain quite unchanged until 150ms for JAMP.

4.3. Real audio data

The real difficulty arises when the background is not perfectly identical (e.g when considering a repeated musical pattern). The experiment here consists of separating the singing voice from a repeated musical background. Due to variation in the execution, the background is not exactly the same nor perfectly aligned since tempo variation can occur, which makes it a difficult task.

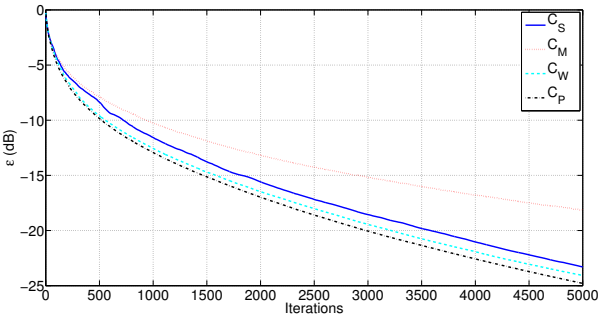

Fig. 3. Mean normalized reconstruction error for various criteria. The criteria $\mathcal{C}_P$ that maximizes the separation performances is also the one that minimizes the reconstruction error.

As in [5], we use audio material from the Beach Boys. We have cut musical excerpts in 5 songs by hand and constituted 5 sets of 4 mixtures. In each set, all the mixtures are occurrences of the same repeated pattern (usually from the verse) and last a few seconds (from 3 to 6). We have compared the results of the JAMP algorithm (run for 10000 iterations) with REPET using the min and median methods. We have also compared performances when using only $I = 3$ occurrences for the separation. Results are summarized in Table 2. Performances are quite comparable with a slight advantage to JAMP on the singing voice SDRs and SARs. For the background, REPET-Min gives the best scores except for SIRs where JAMP is clearly ahead.

These results are encouraging. While using a single generic dictionary, JAMP manages to sparsely decompose both the shared source and the individual components. Perceptively though, JAMP creates ringing artifacts, but those could be reduced (e.g. by pre-echo control methods [10]). Increasing the number of iteration leads to an increase of all JAMP scores but the Musical Background SIRs.

5. BEYOND SOURCE SEPARATION

The simultaneous approximation problem (i.e. finding good joint approximations of the signals) appears disjoint from the source separation problem addressed above. Actually, the global reconstruction error is not intuitively linked to the source separation performances.

With the synthetic dataset described in 4.2, we have found that the criteria $\mathcal{C}_P$ gave the best separation performances. Figure 3 shows that it is also the one that minimizes the reconstruction error ϵ:

$$\epsilon = 10 \log \left(\frac{\|\mathbf{X} - \mathbf{C}_\mathbf{X} \mathbf{\Phi}\|_F^2}{\|\mathbf{X}\|_F^2} \right)$$

where $\|.\|_F^2$ is the squared Frobenius matrix norm or the sum of the squares of the entries in the matrix. This comes as a surprise since source separation and error minimization could

Method	3 Versions SDR (dB)	3 Versions SIR (dB)	3 Versions SAR (dB)	4 Versions SDR (dB)	4 Versions SIR (dB)	4 Versions SAR (dB)
	Musical Background					
REPET-Min	**3.16** $\pm$ 1.7	3.41 $\pm$ 5.8	**10.03** $\pm$ 1.9	**3.47** $\pm$ 1.2	3.29 $\pm$ 4.7	**11.23** $\pm$ 2.0
REPET-Med	2.49 $\pm$ 0.6	8.08 $\pm$ 6.4	3.28 $\pm$ 1.5	2.62 $\pm$ 0.7	7.61 $\pm$ 6.3	4.23 $\pm$ 1.6
JAMP - $\mathcal{C}_P$	1.96 $\pm$ 0.6	**19.14** $\pm$ 7.2	-0.87 $\pm$ 2.2	2.06 $\pm$ 0.6	**17.42** $\pm$ 6.0	-0.60 $\pm$ 2.3
	Singing Voice					
REPET-Min	1.67 $\pm$ 0.9	**9.96** $\pm$ 3.2	0.25 $\pm$ 3.0	1.39 $\pm$ 0.7	**11.17** $\pm$ 3.1	-0.55 $\pm$ 2.4
REPET-Med	2.91 $\pm$ 0.6	5.47 $\pm$ 2.7	4.71 $\pm$ 1.8	2.92 $\pm$ 0.4	5.40 $\pm$ 2.2	**4.96** $\pm$ 1.5
JAMP - $\mathcal{C}_P$	**3.62** $\pm$ 0.8	5.94 $\pm$ 2.5	**5.21** $\pm$ 1.8	**3.48** $\pm$ 1.0	6.03 $\pm$ 2.5	4.79 $\pm$ 2.3

Table 2. Separation scores on repeating musical segments from the Beach Boys. JAMP stopped after 10000 iterations

have been antagonistic optimization goals. The fact that the proposed method optimizes both objectives opens interesting perspectives. Minimizing ϵ is a desirable property in a compression context, hence, this work could be embedded in a broader distributed source coding scheme. Recent work on distributed compressive sampling [13] support this prospective. The source separation would then be a nice additional feature of the compression.

6. CONCLUSION

The joint modeling of the shared source and the individual ones accounts to modeling the redundant and the non-redundant parts of the signal. Although further theoretical studies must be conducted on the matter, it is worth noticing that efficiently separating those parts enables the compression of the redundant parts, but also succeed in minimizing a global reconstruction error on the original mixtures. For musical signals, one cannot always make the assumption that those parts have sparse expansions on different dictionaries. The proposed method overcomes this limitation. JAMP is a simple, fast pursuit algorithm. Hence, it provides an interesting alternative to existing methods in the context of musical repeated pattern separation. Artifacts reduction should be the next matter of concern, and future work will try to embed the model on a broader signal structuration scheme.

7. REFERENCES

[1] A. Liutkus and P. Leveau, "Separation of music+effects sound track from several international versions of the same movie," in *128th AES conv.*, 2010.

[2] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, "Simultaneous sparse approximation via greedy pursuit," *Proc. ICASSP*, 2005.

[3] R. Gribonval, B. Mailhe, H. Rauhut, K. Schnass, and P. Vandergheynst, "Average case analysis of multichannel thresholding," *Proc. ICASSP*, 2007.

[4] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," *Proc. ICASSP*, 2011.

[5] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proc. ICASSP*, 2012.

[6] P.S Huang, S.D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," *Proc. ICASSP*, 2012.

[7] E. Candès, X.Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM (85)*, 2011.

[8] R. Gribonval and S. Lesage, "A survey of Sparse Component Analysis for blind source separation: principles, perspectives, and new challenges," in *Proc. ESANN*, 2006.

[9] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. on Audio, Speech, Lang. Proc.(18)*, 2010.

[10] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Trans. on Audio, Speech, Lang. Proc.*, 2008.

[11] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Proc.(41)*, 1993.

[12] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, Lang. Proc.(14)*, 2006.

[13] D. Sundman, S. Chatterjee, and M. Skoglund, "Greedy pursuits of compressed sensing of jointly sparse signal.," *Proc. EUSIPCO*, 2011.