HAL archives-ouvertes.fr

A compressed sensing approach to the simultaneous recording of multiple room impulse responses

Alexis Benichoux, Emmanuel Vincent, Rémi Gribonval

To cite this version:

Alexis Benichoux, Emmanuel Vincent, Rémi Gribonval. A compressed sensing approach to the simultaneous recording of multiple room impulse responses. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct 2011, New Paltz, NY, United States. pp.1, 2011. <hal-00612911v2>

HAL Id: hal-00612911

https://hal.archives-ouvertes.fr/hal-00612911v2

Submitted on 5 Aug 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A COMPRESSED SENSING APPROACH TO THE SIMULTANEOUS RECORDING OF MULTIPLE ROOM IMPULSE RESPONSES

Alexis Benichoux

Université Rennes 1, IRISA - UMR6074
Campus de Beaulieu 35042 Rennes, FR
alexis.benichoux@irisa.fr

Emmanuel Vincent, Rémi Gribonval

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu 35042 Rennes, FR
{emmanuel.vincent, remi.gribonval}@inria.fr

ABSTRACT

We consider the estimation of multiple room impulse responses from the simultaneous recording of several known sources. Existing techniques are restricted to the case where the number of sources is at most equal to the number of sensors. We relax this assumption in the case where the sources are known. To this aim, we propose statistical models of the filters associated with convex log-likelihoods, and we propose a convex optimization algorithm to solve the inverse problem with the resulting penalties. We provide a comparison between penalties via a set of experiments which shows that our method allows to speed up the recording process with a controlled quality tradeoff.

***Index Terms*—** Room impulse response recording, convex optimization, compressed sensing

1. INTRODUCTION

We focus on the recording of multiple room impulse responses. Up to now this is typically achieved by activating each loudspeaker or *source* in turn, with a silent interval equal to the expected duration of the impulse response in between [1]. The total recording duration is then $N(D + K - 1)$ where N is the number of sources, D the chirp duration and K the impulse response length in samples. An improvement [2] is to use time-overlapping but time-frequency disjoint chirps, which reduces the recording duration down to $NK + D - 1$ when the system is linear. These techniques remain time-consuming *e.g.* in the context of the calibration of high-end 3D audio systems or the collection of binaural room impulse responses involving hundreds of loudspeakers. We investigate here possible improvements using state-of-the-art system inversion tools. This problem is equivalent to the estimation of the *mixing filters* in the context of convolutive source separation [3].

The techniques in [4] and [3] for mixing filter estimation assume each source to be active alone in a certain time interval. Once this time interval has been localized, the corresponding filters are estimated using a subspace method [4], or convex optimization [3]. Alternative Convolutive Independent Component Analysis techniques [5] assume the number of sources to be at most equal to the number of sensors. Our work is to our knowledge the first to get rid of these two assumptions. We propose to take advantage of the *a priori* temporal structure of the filters to improve the iterative inversion of the linear system. In addition to the sparse prior introduced in [6] for single-source blind channel identification, we propose four new priors and a new multi-source inversion algorithm. Our approach is an example of *compressed sensing* [7][8], that is an emerging general approach to the recovery of structured signals from a smaller number of measurements. We show theoretically that white noise sources provide the most convenient system for inversion.

The structure of the paper is as follows. The formalization of the problem is described section 2. Section 3 corresponds to the study of the *a priori* structure of the filters. The implementation of the algorithm is detailed Section 4. The results shown in Section 5 show the potential of the proposed method.

2. APPROACH

The problem is formalized as follows : we represent the N sources of length T by the matrix $\mathbf{S} \in \mathbb{R}^{N \times T}$, the filters of length K by the three dimensional array $\mathbf{A} \in \mathbb{R}^{M \times N \times K}$ and the M observations by $\mathbf{X} \in \mathbb{R}^{M \times (T+K-1)}$. Assuming that the loudspeaker are linear, the convolutive matrix product $\star$ allows us to write

$$\mathbf{X} = \mathbf{A} \star \mathbf{S} = \left(\sum_{n \le N} A_{mn} * S_n \right)_{m \le M}. \tag{1}$$

Earlier work [9] used convex optimization tools to recover $\mathbf{S}$ when $\mathbf{A}$ is known, using a sparsity prior on the sources.

Here we adapt the method in [9] to estimate $\mathbf{A}$ when $\mathbf{S}$ is known, by estimating $\lim_{\lambda \to 0} \mathbf{A}_\lambda$ where

$$\mathbf{A}_\lambda = \mathrm{argmin}_{\mathbf{A}} \left\{ \frac{1}{2} \|\mathbf{X} - \mathbf{A} \star \mathbf{S}\|_2^2 + \lambda \mathcal{P}(\mathbf{A}) \right\}. \tag{2}$$

This limit is the solution of the constrained minimization problem

$$\min_{\mathbf{A}} \mathcal{P}(\mathbf{A}) \quad \text{s.t.} \quad \|\mathbf{X} - \mathbf{A} \star \mathbf{S}\|_2^2 = 0. \tag{3}$$

We choose for $\mathcal{P}$ the negative log-likelihood of a distribution suggested by the statistical analysis of a large family of filters.

3. STATISTICAL ANALYSIS OF A FAMILY OF FILTERS

The statistical theory of room acoustics [10] treats each filter as a random i.i.d. signal whose amplitude envelope $\rho(t)$ decays exponentially according to

$$\rho(t) = \sigma \, 10^{-3t/\mathrm{t_R}}, \tag{4}$$

where $\mathrm{t_R}$ is the room reverberation time in samples, and σ a scaling factor. This theory assumes that a filter $A \in \mathbb{R}^K$ follows a Gaussian distribution. In other work [6], $A(t)$ is instead assumed to have a constant amplitude envelope and to be sparse, as it is formed by

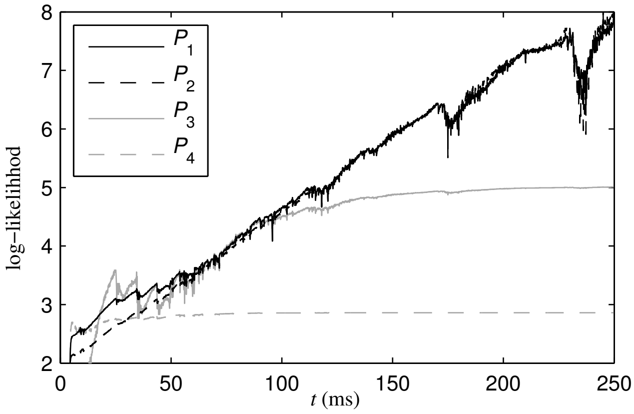

Figure 1: Comparison of the statistical models (5) to (8) over a set of generated filters for a reverberation time of 250 ms.

echoes at distinct instants. In order to evaluate the respective impact of both the envelope model and the sparsity model, we consider the following distributions : Laplacian with decaying envelope

$$P_1(t) = \frac{1}{2\rho(t)} e^{-|A(t)|/\rho(t)}, \tag{5}$$

Gaussian with decaying envelope

$$P_2(t) = \frac{1}{\sqrt{2\pi}\rho(t)} e^{-A^2(t)/2\rho^2(t)}, \tag{6}$$

Laplacian with constant envelope

$$P_3(t) = \frac{1}{2\sigma} e^{-|A(t)|/\sigma}, \tag{7}$$

Gaussian with constant envelope

$$P_4(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-A^2(t)/2\sigma^2}. \tag{8}$$

Figure 1 compares the average negative log-likelihoods of these four models over a set of 10 000 filters simulated by the image method [11] for one source and one microphone at random positions spaced by 1 m, in a rectangular room of dimensions $10 \times 8 \times 4$ m with $\mathrm{t_R} = 250$ ms. For each model, the scaling factor σ is set in the maximum likelihood sense. Envelope modeling appears to be crucial : the likelihood of models P_3 and P_4 is much larger than that of P_1 and P_2 for large t. Sparsity has a weaker impact : the likelihood of P_1 (and to a lesser extent that of P_2) is larger than that of P_2 for $t \leq 60$ ms, but becomes similar for $t > 60$ ms. These observations lead us to propose a fifth hybrid model

$$P_5(t) = \begin{cases} P_1(t) & \text{if } t \leq 60 \text{ ms} \\ P_2(t) & \text{i } t > 60 \text{ ms.} \end{cases} \tag{9}$$

Assuming Gaussian white additive noise, maximum *a posteriori* estimation of the filters is equivalent to (2) with $\mathcal{P}_i = -\log P_i$.

4. ALGORITHM

To solve (2), we use the FISTA (Fast Iterative Shrinkage-Thresholding) algorithm [12], which exploits the differentiability of the data fidelity term

$$\mathcal{L} : \mathbf{A} \mapsto \|\mathbf{X} - \mathbf{A} \star \mathbf{S}\|_2^2, \tag{10}$$

and the convexity and semicontinuity of $\mathcal{P}_i$. So-called proximity operators are employed to overcome the non-differentiability of $\mathcal{P}_i$.

Definition 1 *For $\mathcal{P} : E \to \mathbb{R}$ semicontinuous and convex the proximity operator associated with $\mathcal{P}$ is the function*

$$\mathrm{prox}_{\mathcal{P}} : \mathbf{x} \in E \mapsto \mathrm{argmin}_{\mathbf{y} \in E} \left\{ \mathcal{P}(\mathbf{y}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \right\}$$

The general steps of FISTA are described in Algorithm 1. It relies on the computation of the gradient of $\mathcal{L}$, its Lipschitz constant L, and the proximity operator of the scaled penalty $\alpha\mathcal{P}$.

Algorithm 1 FISTA

1: $\mathbf{A}^0 \in \mathbb{R}^{MNK}, \tau^0 = 1$
2: **for** $k \leq k_{\max}$ **do**
$\tilde{\mathbf{A}}^k = \mathrm{prox}_{\frac{\lambda}{L}\mathcal{P}} \left(\mathbf{A}^{k-1} - \frac{\nabla\mathcal{L}(\mathbf{A}^{k-1})}{L} \right)$
$\tau^k = \frac{1+\sqrt{1+4(\tau^{k-1})^2}}{2}$
$\mathbf{A}^k = \tilde{\mathbf{A}}^k + \frac{\tau^{k-1}-1}{\tau^k}(\tilde{\mathbf{A}}^k - \tilde{\mathbf{A}}^{k-1})$
3: **end for**

The computation of the gradient of $\mathcal{L}$ requires the introduction of the adjoint of the linear operator $\mathbf{A} \mapsto \mathbf{A} \star \mathbf{S}$. Denoting by $\tilde{S}_n \in \mathbb{R}^T$ the time reversal of S_n, *i.e.* for $t \leq T$, $\tilde{S}_n(t) = S_n(T - t + 1)$, the adjoint operator is expressed using $S^* := \left(\tilde{S}_1, \ldots, \tilde{S}_N\right)$ as

$$\mathbf{X} \mapsto \mathbf{X} \star \mathbf{S}^* := \left((\tilde{S}_n * X_m)(t) \right)_{m \leq M, n \leq N, 1 \leq t \leq K}. \tag{11}$$

One may then write the gradient as

$$\nabla\mathcal{L}(\mathbf{A}) = (\mathbf{X} - \mathbf{A} \star \mathbf{S}) \star \mathbf{S}^*. \tag{12}$$

The Lipschitz constant of $\nabla\mathcal{L}$ is the greatest eigenvalue of the operator $\mathbf{A} \mapsto \mathbf{A} \star \mathbf{S} \star \mathbf{S}^*$. We obtain this value using the power iteration algorithm as in [9, Algorithm 5].

The log-likelihood of the distributions introduced previously correspond to the ℓ_1 and ℓ_2 norms, whose proximity operators are well-known [9]. Denoting $x^+ := \max(x, 0)$for $x \in \mathbb{R}$, we obtain

$$\mathrm{prox}_{\alpha\mathcal{P}_1}(\mathbf{A})_{m,n,t} = \frac{\rho(t)A_{mn}(t)}{|\rho(t)A_{mn}(t)|} \left(|A_{mn}(t)| - \frac{\alpha}{\rho(t)} \right)^+ \tag{13}$$

$$\mathrm{prox}_{\alpha\mathcal{P}_2}(\mathbf{A})_{m,n,t} = \frac{A_{mn}(t)}{1 + \alpha/\rho^2} \tag{14}$$

$$\mathrm{prox}_{\alpha\mathcal{P}_3}(\mathbf{A})_{m,n,t} = \frac{A_{mn}(t)}{|A_{mn}(t)|} (|A_{mn}(t)| - \alpha)^+ \tag{15}$$

$$\mathrm{prox}_{\alpha\mathcal{P}_4}(\mathbf{A})_{m,n,t} = \frac{A_{mn}(t)}{1 + \alpha}. \tag{16}$$

Concerning the hybrid model (9), we use (13) or (14) depending on the value of t.

We estimate the minima $\mathbf{A}_\lambda$ for $\lambda \in \{1, 10^{-1}, \ldots, 10^{-14}\}$, initializing each FISTA step at the minimum obtained for the previous value. We keep the last minimum obtained for $\lambda = 10^{-14}$, and consider it as an estimate of the limit $\lim_{\lambda\to 0} \mathbf{A}_\lambda$, *i.e.* the solution of (3). Note that the ℓ_2 penalization $\mathcal{P}_4$ corresponds to the definition of the Moore-Penrose pseudo inversion.

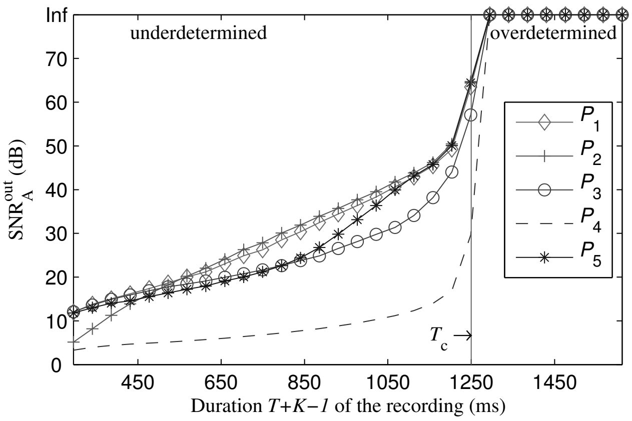

Figure 2: Performance of the estimation of $\mathbf{A}$ with $N = 5$ white noise sources, depending on the duration of the signal.

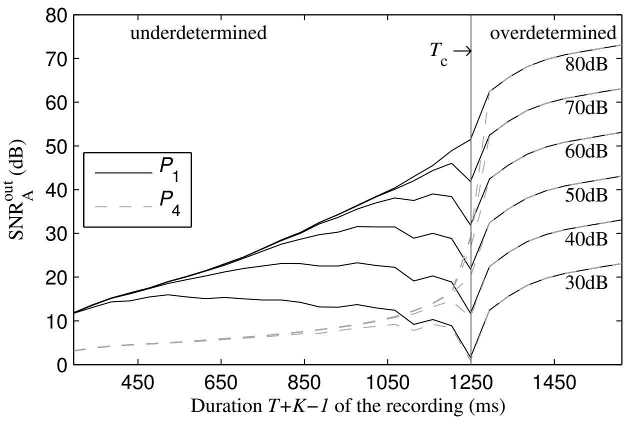

Figure 3: Estimation of $\mathbf{A}$ for six different additive noise levels, depending on the recording duration with $N = 5$ sources

5. EXPERIMENTAL RESULTS

The Matlab code allowing to reproduce the following experiments is available at the following address [13].

5.1. Role of the condition number

First we wish to study the contribution of the penalty depending of the invertibility of the problem. The system is composed of $M(T+K-1)$ equations for MNK variables, therefore it is underdetermined if and only if the recording duration in samples satisfies

$$T + K - 1 < T_c := NK. \tag{17}$$

Note that T_c is smaller than $NK + D - 1$ which is the length of the recording required in [2].

Performance does not depend on the number M of microphones, in fact each microphone brings an independent problem.

The MN filters are a solution of the linear system, for $m \leq M$

$$X_m = S_1 \star A_{m1} + S_2 \star A_{m2} + \ldots + S_N \star A_{mN}. \tag{18}$$

In order to guide the choice of the source signals, we first show theoretically that the system is well conditioned if the sources are uncorrelated. To this aim we compute the condition number of the system *i.e.* the ratio of highest to lowest singular value. Note that this is only defined for a full rank system, hence for $T \geq T_c$. In the under-determined setting, similar estimates can easily be obtained by exploiting the sparsity of the filters $\mathbf{A}$.

The relation between the correlation of the sources and the condition number of the system is detailed in the following lemma proved in the Appendix. Using the usual cross-correlation function for $n, n' \leq N$:

$$r_{nn'}(k) = \sum_{t=1}^{T-k} S_n(t) S_{n'}(t+k), \quad 0 \leq k \leq K-1. \tag{19}$$

we introduce a measure of the maximum correlation between the sources

$$r := \max_{n \neq n' \text{ or } k \neq 0} |r_{nn'}(k)|, \tag{20}$$

Lemma 1 *For small r, the condition number of* (18) *obeys*

$$1 \leq c \leq \frac{\max_n r_{nn}(0) + r(NK-1)}{\min_n r_{nn}(0) - r(NK-1)}. \tag{21}$$

Note that for white noises, r is small, and c is close to 1, leading to a well conditioned system.

5.2. Performance as a function of the recording duration

In previous work [14, fig.2], we used human voice recordings, and we observed a transitory regime for $T > T_c$ where the penalties still had an impact due to a large condition number. In this paper we want to choose the sources such that the system is well-conditioned, therefore we used white Gaussian noise, motivated by the above theoretical guarantees. We observed experimentally (results not shown here) that this choice remains experimentally valid for underdetermined systems.

As a measurement of the error between the estimated filters $\mathbf{A}_\lambda$ and the true filters $\mathbf{A}$, we define the following ratio in decibels

$$\text{SNR}_{\mathbf{A}}^{\text{out}}(\mathbf{A}_\lambda) = 10 \log_{10} \frac{\|\mathbf{A}\|_2^2}{\|\mathbf{A}_\lambda - \mathbf{A}\|_2^2}. \tag{22}$$

When the solution is not unique, we expect to observe better results with the proposed regularizations : we then run the algorithm for several values of T.

The results shown in Figure 2 correspond to the case of $N = 5$ sources, $M = 2$ sensors, with filters of length $K = 2753$ (250 ms sampled at 11025 Hz) synthesized as in Section 2. For readability we express all the durations in ms the following. We obtain the critical value $T_c = 1250$ ms beyond which the system is overdetermined. We vary the length of the sources from $T = 45$ ms to $T = 1500$ ms.

We observe in Figure 2 a clear jump around the critical value T_c after which the inversion made by all regularized algorithms yields the same solution, up to machine precision.

For $T < T_c$ we observe the clear impact of all regularizations compared to the pseudo-inverse $\mathcal{P}_4$. The new penalties $\mathcal{P}_1$ and $\mathcal{P}_2$, corresponding to the Laplacian and Gaussian distributions with decaying envelope give the best results. For a speed-up of 50% in recording duration compared to [2], we achieve a recovery $\text{SNR}_{\mathbf{A}}^{\text{out}} = 25$ dB.

5.3. Robustness to noise

We now add Gaussian white additive noise to the mixtures,

$$\mathbf{X} = \mathbf{A} \star \mathbf{S} + \mathbf{W}. \tag{23}$$

For each penalty $\mathcal{P}_i$ and each duration $T + K - 1$ of the recordings, six experiments were made for a signal-to-noise ratio of $30, 40, 50, 60, 70$ and 80 dB. We observe in Figure 3 that the noise decreases the overall performance, but has a smaller impact on the ℓ_1 re-scaled penalty $\mathcal{P}_1$ than on the common pseudo-inversion $\mathcal{P}_4$. Not surprisingly, the two penalties lead to the same result once the sources are long enough for the solution to be unique.

This experiment confirms the possibility to speedup the recordings even in the presence of noise. With an input SNR of 50 dB, the estimation fidelity $\text{SNR}_{\mathbf{A}}^{\text{out}}$ is still 25 dB.

6. CONCLUSION

For the considered problem, the various *a priori* introduced as convex penalties provide better estimation of the filters than simple deconvolution using Moore-Penrose pseudo-inverse. The best results are achieved with the new proposed penalties based on a decaying envelope model. This method can speed up the recording, with a reasonable quality trade-off for noisy measurements. For large numbers of sources N in reverberant rooms, the expected speedup can be significant. Further experiments are needed to confirm the validity of the approach in such scenarii, by taking into account the nonlinearity of the loudspeakers, as well as other performance measures. Besides, we know that source separation informed by the filters provides better results [9] : this opens the way to the alternate estimation of both the sources and the filters with suitable penalties on the filters as opposed to [15].

Appendix

Denote

$$\Sigma_n := \begin{pmatrix} S_n(1) & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ & & & 0 \\ S_n(T) & & & S_n(1) \\ 0 & \ddots & & \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & 0 & S_n(T) \end{pmatrix} \in \mathbb{R}^{(T+K-1)\times K}.$$

We derive from (18) the block matrix notation

$$(\Sigma_n^T X_m)_{n\le N} = (\Sigma_n^T \Sigma_{n'})_{n,n'\le N} (A_{mn})_{n\le N}, \tag{24}$$

and the correlation of the sources (19) appears since we have

$$\Sigma_n^T \Sigma_{n'} = \begin{pmatrix} r_{nn'}(0) & \cdots & r_{n'n}(K-1) \\ \vdots & \ddots & \vdots \\ r_{nn'}(K-1) & \cdots & r_{nn'}(0) \end{pmatrix}. \tag{25}$$

Now c is the condition number of the $NK \times NK$ block matrix $\mathbf{R} = (\Sigma_n^T \Sigma_{n'})_{n,n'\le N}$, and the key is to choose the sources so that this matrix is highly diagonally dominant. Using Gerschogorin's disc theorem [16], we control its eigenvalues. For $\lambda \in \text{Sp}(\mathbf{R})$, there exist $n \le N$ such that

$$\begin{aligned} |\lambda - r_{nn}(0)| &\le \sum_{k=1}^{K-1} r_{nn}(k) + \sum_{k=0}^{K-1} \sum_{n\ne n'} r & (26) \\ &\le r(NK-1) & (27) \end{aligned}$$

Then if $r(NK-1) < \min_n r_{nn}(0)$ we can conclude that

$$c = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \le \frac{\max_n r_{nn}(0) + r(NK-1)}{\min_n r_{nn}(0) - r(NK-1)}. \tag{28}$$

This work was supported by the EU FET-Open project FP7-ICT-225913-SMALL and the ANR project 08-EMER-006 ECHANGE.

7. REFERENCES

[1] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. AES 108th Convention*.

[2] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *Journal Audio Engeneering Society*.

[3] P. Sudhakar, S. Arberet, and R. Gribonval, "Double sparsity: Towards blind estimation of multiple channels," *in Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*.

[4] A. Aissa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Blind separation of underdetermined convolutive mixtures using their time-frequency representation," *IEEE Transactions on Audio Speech and Language Processing*.

[5] S. Makino, T. Lee, and H. Sawada, *Blind speech separation*.

[6] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind channel identification for speech dereverberation using ℓ_1-norm sparse learning," in *Advances in Neural Information Processing Systems 20*.

[7] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*.

[8] E. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*. Citeseer.

[9] 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*.

[10] H. Kuttruff, *Room Acoustics*, 4th ed., New York.

[11] J. Allen and A. Berkeley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, Apr.

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*.

[13] http://hal.inria.fr/inria-00594252/.

[14] A. Benichoux, E. Vincent, and R. Gribonval, "Optimisation convexe pour l'estimation simultanée de réponses acoustiques," in *Actes du* 23^e *colloque GRETSI*.

[15] D. Barchiesi and M. D. Plumbley, "Dictionary learning of convolved signals," *In Proc. Conf. on Acoustics, Speech and Signal Processing*.

[16] D. Serre, *Matrices: Theory and applications*.