Optimisation convexe pour l'estimation simultanée de réponses acoustiques

Alexis Benichoux[1], Emmanuel Vincent[2], Rémi Gribonval[2]

[1]Université Rennes 1, IRISA - UMR6074, Campus de Beaulieu, 35042 Rennes Cedex, France

[2]INRIA, Centre de Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France

[1] alexis.benichoux@irisa.fr

[2]{emmanuel.vincent,remi.gribonval}@inria.fr

Résumé − On s'intéresse à l'estimation de réponses acoustiques à partir de l'enregistrement simultané de plusieurs sources connues. Les techniques existantes nécessitent de pouvoir se ramener au cas où le nombre de sources, connues ou inconnues, est au plus égal au nombre de capteurs. Notre méthode s'affranchit de cette hypothèse dans le cas où les sources sont connues. Pour cela, on propose des modèles statistiques de filtres associés à des log-vraisemblances convexes, puis on met en place des algorithmes d'optimisation convexe pour résoudre le problème inverse, avec les pénalités qui résultent de ces modèles. On fournit une comparaison des différentes pénalités via un jeu d'expériences qui montre que la méthode proposée permet d'obtenir une estimation robuste, et augmente nettement les performances par rapport à l'approche naïve.

Abstract − We consider the estimation of acoustic 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 a convex log-likelihood, and we propose a convex optimization algorithm to solve the inverse problem, with the resulting penalties. We provide a comparaison between penalties via a set of experiments which shows that the proposed method allows to obtain a robust estimation, and greatly improves performance compared to the naive approach.

1 Introduction

L'enregistrement de réponses de salle s'effectue à ce jour par l'activation successive de sources à des positions différentes [6]. Dans le cas de l'enregistrement des BRIRs (*Binaural Room Impulse Responses*), l'estimation simultanée de plusieurs réponses à partir de l'enregistrement simultané de plusieurs sources permettrait de gagner du temps lorsque le nombre de positions est élevé.

Les méthodes [2] et [12] pour l'estimation des paramètres des mélanges convolutifs supposent que chaque source est présente seule sur un certain intervalle de temps. Une fois cet intervalle trouvé, les réponses acoustiques (ou filtres) associées sont estimées par la méthode des sous-espaces pour [2] et par optimisation convexe pour [12]. L'Analyse en Composantes Indépendantes (ACI) convolutive [10] suppose quant à elle que le nombre de sources est au plus égal au nombre de capteurs.

Nos travaux sont à notre connaissance les premiers qui s'affranchissent de ces deux hypothèses. On propose de prendre en compte la structure *a priori* des filtres pour faciliter l'inversion, itérative, du système.

La formalisation du problème est décrite dans la partie 2. La partie3 correspond à la recherche de l'*a priori* de structure des filtres. La mise en œuvre de l'algorithme résolvant le problème posé est détaillée dans la partie 4. Les résultats exposés dans la partie 5 valident la démarche proposée, notamment lorsque le système est sous-determiné, ou seulement faiblement surdéterminé.

2 Approche

Le problème est formalisé comme suit : on représente les N sources de longueur T par la matrice $\mathbf{S} = (S_n[t]) \in \mathbb{R}^{NT}$, les filtres de longueur K par $\mathbf{A} = (A_{mn}[t]) \in \mathbb{R}^{MNK}$ et les M observations par $\mathbf{X} = (X_m[t]) \in \mathbb{R}^{M(T+K-1)}$. Le produit matriciel convolutif $\star$ permet d'écrire

$$\mathbf{X} = \mathbf{A} \star \mathbf{S}. \tag{1}$$

Des travaux antérieurs [7] estiment $\mathbf{S}$ lorsque $\mathbf{A}$ est connu en exploitant une parcimonie des sources avec des techniques d'optimisation convexe. Ici, on adapte cette démarche pour estimer $\mathbf{A}$ lorsque $\mathbf{S}$ est connu en calculant $\lim_{\lambda \to 0} \mathbf{A}_\lambda$ avec

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

Si $\mathcal{P}$ est convexe, cette limite est alors solution de

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

Cette démarche revient choisir à la solution du système la plus vraissemblable au sens de $\mathcal{P}$. On choisit pour pénalité $\mathcal{P}$ l'opposé de la log-vraisemblance d'une distribution suggérée par l'analyse statistique d'une famille de filtres.

3 Analyse statistique d'une famille de filtres

La théorie statistique de l'acoustique des salles [8] traite chaque filtre $a(t)$ comme un signal aléatoire non i.i.d. dont l'amplitude moyenne $\rho(t)$ décroît exponentiellement selon

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

où $\mathrm{t_R}$ est le temps de réverbération de la salle exprimé en échantillons et σ un facteur d'échelle. Cette théorie suppose par ailleurs que $a(t)$ suit une distribution gaussienne. D'autres travaux [9] supposent au contraire que $a(t)$ a une amplitude moyenne constante et est parcimonieux car constitué d'échos à des instants distincts. Afin d'évaluer l'impact respectif de la décroissance d'amplitude et de la parcimonie, nous considérons les quatre distributions suivantes : laplacienne à amplitude décroissante

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

gaussienne à amplitude décroissante

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

laplacienne à amplitude constante

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

gaussienne à amplitude constante

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

La figure 1 compare la log-vraisemblance moyenne de ces quatre modèles sur un ensemble de 10 000 filtres simulés par la méthode des sources images [3] pour une source et un micro positionnés aléatoirement à 1 m l'un de l'autre dans une salle de taille $10 \times 8 \times 4$ m avec $\mathrm{t_R} = 250$ ms. Pour chaque modèle, le facteur d'échelle σ est fixé au sens du maximum de vraisemblance. La modélisation de la variation d'amplitude au cours du temps apparaît comme cruciale : en effet, la vraisemblance des modèles P_3 et P_4 est beaucoup plus faible que celle des modèles P_1 et P_2 pour t grand. La modélisation de la parcimonie a un impact plus faible : la vraisemblance de P_1 (et dans une moindre mesure celle de P_3) surpasse légèrement celle de P_2 pour $t \leq 60$ ms, mais devient similaire pour $t > 60$ ms. Un zoom montre néanmoins que P_2 est légèrement meilleur que P_1 pour $t > 60$ ms. Ces observations nous conduisent à proposer un cinquième modèle hybride

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

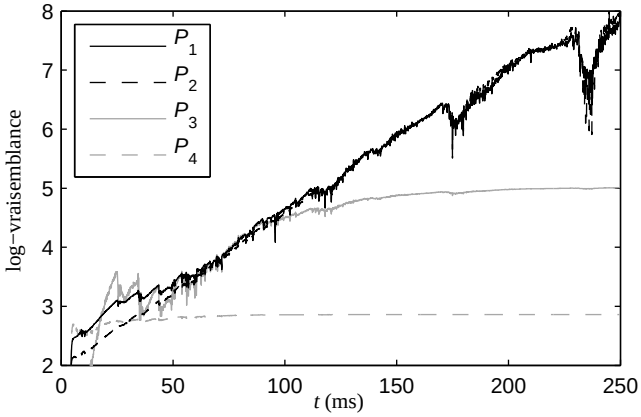

FIG. 1: Comparaison des modèles statistiques de filtres (5) à (8) sur un ensemble de filtres générés pour une salle de temps de réverbération égal à 250 ms.

4 Algorithme

Pour résoudre (2) on utilise l'algorithme de seuillage itératif doux FISTA (*Fast Iterative Shrinkage-Thresholding Algorithm*, [4]) qui exploite d'une part la différentiabilité de

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

et la convexité et la semi-continuité de $\mathcal{P}_i = -\log P_i$ d'autre part. Les opérateurs proximaux [11] permettent de contourner la non différentiabilité de $\mathcal{P}$ pour construire des algorithmes efficaces.

Definition 1 *Pour $\mathcal{P} : E \to \mathbb{R}$ semi-continue et convexe on a appelle opérateur proximal associé à $\mathcal{P}$ la fonction*

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

Les étapes de FISTA sont decrites dans l'algorithme 1. On a besoin de connaître le gradient de $\mathcal{L}$, sa constante de Lipschitz L, et l'opérateur proximal de $\mathcal{P}$.

Algorithme 1 FISTA

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

Le calcul du gradient de $\mathcal{L}$ conduit à l'introduction d'un opérateur adjoint correspondant aux règles de calcul de la convolution $\star$ des matrices à trois dimensions $\mathbf{A}$ et $\mathbf{S}$. En notant $\tilde{S}_k \in \mathbb{R}^T$ le retournement de S_k, *i.e.* pour $1 \leq t \leq T$, $\tilde{S}_k[t] = S_k[T - t + 1]$, l'adjoint $\mathbf{S}^*$ vérifie

$$\mathbf{X} \star \mathbf{S}^* = \left((\tilde{S}_n * X_m)[t] \right)_{\substack{m \leq M \\ n \leq N \\ 1 \leq t \leq K}}. \quad (11)$$

Le gradient de $\mathcal{L}$ s'ecrit alors

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

La constante de Lipschitz de $\nabla\mathcal{L}$ est la plus grande valeur propre de l'opérateur $\mathbf{A} \mapsto \mathbf{A}\star\mathbf{S}\star\mathbf{S}^*$, et on l'obtient numériquement par l'algorithme des puissances itérées, [7, Algorithme 5].

La log-vraisemblance des distributions introduites correspond à des normes ℓ_1 ou ℓ_2, les opérateurs proximaux correspondants sont alors bien connus [7]. Les pénalités introduites sont séparables, c'est à dire que l'on peut raisonner coordonnée par coordonnée [5, Lemme 2.9]. Pour la norme ℓ_1 on obtient le seuillage doux [7, Proposition 1]. L'opérateur proximal de la norme ℓ_2 s'obtient directement en utilisant la différentiabilité, et pour l'introduction du facteur d'échelle on utilise la règle de calcul [5, Lemme 2.6]. En notant pour $x \in \mathbb{R}$

$$x^+ = \max(x, 0), \tag{13}$$

les opérateurs proximaux des log-vraisemblances des distributions (5) à (8) sont

$$\mathrm{prox}_{\lambda\mathcal{P}_1}(\mathbf{A})_{m,n,t} = \frac{\rho(t)A_{m,n,t}}{|\rho(t)A_{m,n,t}|}\left(|A_{m,n,t}| - \frac{\lambda}{\rho(t)}\right)^+ \tag{14}$$

$$\mathrm{prox}_{\lambda\mathcal{P}_2}(\mathbf{A})_{m,n,t} = \frac{A_{m,n,t}}{1 + \lambda/\rho^2} \tag{15}$$

$$\mathrm{prox}_{\lambda\mathcal{P}_3}(\mathbf{A})_{m,n,t} = \frac{A_{m,n,t}}{|A_{m,n,t}|}\left(|A_{m,n,t}| - \lambda\right)^+ \tag{16}$$

$$\mathrm{prox}_{\lambda\mathcal{P}_4}(\mathbf{A})_{m,n,t} = \frac{A_{m,n,t}}{1 + \lambda}. \tag{17}$$

Pour le modèle hybride (9) on utilise (14) ou (15) selon la valeur de t.

On calcule les minima $\mathbf{A}_\lambda$ pour $\lambda \in \{1, 10^{-1}, \ldots, 10^{-14}\}$ en initialisant l'algorithme FISTA au minimum obtenu pour la valeur précédente. On retient le dernier minimum obtenu, pour $\lambda = 10^{-14}$, qu'on considère une bonne approximation de la limite $\lim_{\lambda\to 0} \mathbf{A}_\lambda$. Contrairement à ce ce qu'on pourrait penser de prime abord, cela ne revient pas à annuler la régularisation mais (3) à minimiser le terme de régularisation $\mathcal{P}(\mathbf{A})$ sous la contrainte d'égalité.

Par exemple la pénalité $\mathcal{P}_4$ qui correspond à un régulation par la norme 2, fournit la solution de plus petite norme 2, et donc coïncide avec la pseudo-inversion de Moore-Penrose.

5 Résultats expérimentaux

Le code Matlab permettant de reproduire les deux expériences suivantes est disponible à l'adresse [1].

5.1 Performance en fonction de la durée du signal

On veut d'abord étudier l'apport des pénalités en fonction de l'inversibilité du problème. Le système comporte

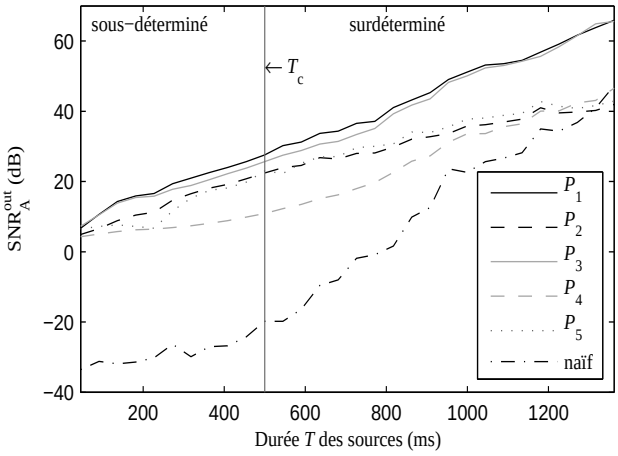

FIG. 2: Performance moyenne d'estimation de $\mathbf{A}$ sur dix enregistrements simulés de trois sources, en fonction de la durée du signal.

$M(T+K-1)$ équations pour MNK variables, il est sous-déterminé si et seulement si $T < (N-1)K+1$. Il s'agit de faire varier la durée T des sources pour mettre en évidence le rôle de la régularisation quand la solution du système n'est pas unique.

On se place dans le cas $N = 3$ sources et $M = 2$ capteurs, avec un filtre de $K = 2753$ échantillons correspondant à $\mathrm{t_R} = 250$ ms, synthétisé comme en Section 2. On obtient la valeur critique $T_c = 500$ ms au delà de laquelle le système est surdéterminé. On fait varier la taille des sources de $T = 45$ ms à $T = 1300$ ms. On dispose de 30 enregistrements de voix de 12 s échantillonnés à 11025 Hz, ce qui permet de moyenner les résultats sur 10 expériences menées avec 3 sources différentes.

Pour mesurer l'erreur entre les filtres estimés $\mathbf{A}_\lambda$ et les vrais filtres $\mathbf{A}$, on introduit le rapport en décibels

$$\mathrm{SNR}_{\mathbf{A}}^{\mathrm{out}}(\mathbf{A}_\lambda) = 10\log_{10}\frac{\|\mathbf{A}\|_2^2}{\|\mathbf{A}_\lambda - \mathbf{A}\|_2^2}. \tag{18}$$

La Figure 2 représente cet indicateur en fonction de la durée des sources, pour chacune des 5 pénalités introduites. On remarque que la solution de (1) n'est visiblement pas unique même en régime surdéterminé, ce qui indique que le système est mal conditionné. En pointillé, on a représenté les résultats obtenus avec une descente de gradient pour la minimisation de $\mathcal{L}$ sans pénalité. Cette approche naïve fournit les plus mauvais résultats, elle converge vers le minimum local le plus proche de l'initialisation. La minimisation associée P_4, qui équivaut à une simple pseudo-inversion du système linéaire, est également endessous des autres pénalités.

L'apport des pénalités P_1, P_2, P_3 et P_5 est nettement visible jusqu'à $T = 1300$ ms, donc bien au-delà du seuil de sous-détermination T_c. Les pénalités en norme ℓ_1, associées aux distributions P_1 et P_3 obtiennent les meilleurs résultats. De plus, la modélisation de la variation d'amplitude

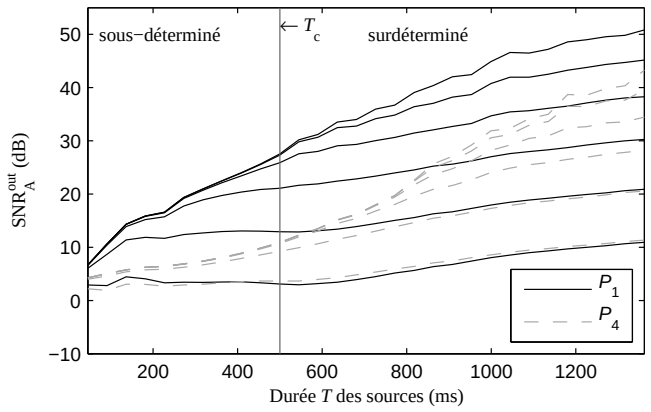

FIG. 3: Estimation du filtre pour six niveaux de bruitage des observations en fonction de la longueur des sources.

des filtres proposée pour P_1 améliore les résultats par rapport à la pénalité ℓ_1 classique associée à P_3.

5.2 Robustesse au bruit

On reprend les mêmes expériences en ajoutant un bruit additif au mélange convolutif

$$\mathbf{X} = \mathbf{A} \star \mathbf{S} + \mathbf{W}. \tag{19}$$

Pour chaque pénalité et pour chaque durée T, six expériences ont été menées pour un rapport signal-à-bruit d'entrée de $30, 40, 50, 60, 70$ et 80 dB. On voit Figure 3 que cette perturbation dégrade la performance mais la pénalité de norme ℓ_1 mise à l'échelle correspondant à P_1 permet d'obtenir une meilleure estimation que le pseudo-inverse correspondant au modèle P_4. Pour T assez grand les deux pénalités donnent le même résultat.

Comme dans la section précédente, les résultats sont moyennés sur un jeu de dix triplets de sources sonores. Les courbes correspondant aux distributions P_1 et P_4 sont confondues pour le niveau 30 dB. Pour un rapport signal-à-bruit d'entrée supérieur à 40 dB, qui correspond à des conditions d'enregistrement réalisables, on obtient avec P_1 un SNR sur les filtres supérieur à 15 dB.

6 Perspectives, conclusion

Pour le problème considéré, l'*a priori* introduit sous la forme d'une pénalité convexe bien choisie permet d'obtenir une meilleure estimation des filtres qu'une simple déconvolution par pseudo-inverse, notamment dans des cas où les sources sont courtes. C'est utile si on veut faire l'hypothèse que le filtre est invariant seulement sur des petits segments de l'enregistrement.

Par ailleurs, on sait que la séparation de sources fonctionne mieux si elle est informée par la matrice de filtres [7] : ces travaux ouvrent donc la voie à l'estimation simultanée des sources et des filtres en vue d'une nouvelle méthode de séparation.

7 Remerciements

Les auteurs remercient les projets EU FET-Open FP7-ICT-225913-SMALL et ANR-08-EMER-006 ECHANGE.

References

[1] `www.irisa.fr/metiss/members/abenicho/filtercs/`.

[2] A. Aissa-El-Bey, K.A. Abed-Meraim, and Y. Grenier. Blind separation of underdetermined convolutive mixtures using their time-frequency representation. *IEEE Transactions on Audio Speech and Language Processing*, 15(5):1540, 2007.

[3] J.B. Allen and A. Berkeley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, Apr. 1979.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2006.

[6] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. AES 108th Convention*, pages 18–22, 2000.

[7] 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*, 18(7):1818–1829, 2010.

[8] H. Kuttruff. *Room Acoustics*. Spon Press, New York, 4rd edition edition, 2000.

[9] Y. Lin, J. Chen, Y. Kim, and D.D. Lee. Blind channel identification for speech dereverberation using ℓ_1-norm sparse learning. In *Advances in Neural Information Processing Systems 20*, pages 921–928. MIT Press, 2007.

[10] S. Makino, T.W. Lee, and H. Sawada. *Blind speech separation*. Springer, 2007.

[11] J.J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

[12] 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*, pages 571–578, 2010.