Analyse local convolutive en vecteurs indépendantes pour la séparation aveugle de sources dans le scénario réverbérant

Fangchen FENG[1], Matthieu KOWALSKI[2]

[1]APC, Univ. Paris Diderot, CNRS/IN2P3, CEA/Irfu, Obs. de Paris, Sorbonne Paris Cité, France

[2]Laboratoire des signaux et systèmes, Univ Paris-Sud, CNRS, CentraleSupélec, France

fangchen.feng@apc.in2p3.fr, matthieu.kowalski@u-psud.fr

Résumé − Dans cet article, nous proposons une nouvelle formulation pour le problème de séparation aveugle de sources avec des mélanges convolutifs. Nous exploitons d'abord le lien entre les méthodes IVA (Independent Vector Analysis) et SCA (Sparse Component Analysis) à travers la parcimonie structurée, puis proposons un nouveau cadre combinant l'approximation convolutive à bande étroite et le Windowed-Group-Lasso. L'approche proposée évite le problème de permutation en fréquence en prenant en compte directement l'information inter-bandes lors de la séparation.

Abstract − In this paper, we propose a new formulation for blind source separation problem with convolutive mixtures. We first exploit the link between the IVA and the SCA methods through the structured sparsity and then propose a new framework combining the convolutive narrowband approximation and the Windowed-Group-Lasso. Being able to avoid the permutation alignment and to take advantage of the cross-band information during the separation, the proposed approach outperforms the existing methods through numerical evaluations.

1 Introduction

Un modèle direct réaliste pour la séparation aveugle de source (BSS – Blind Source Separation) est donné par le modèle de mélange convolutif afin de prendre en compte les effets de réverbération :

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t) * s_n(t) + n_m(t), \tag{1}$$

où s_n est la n-ième source et x_m est le m-ième mélange. N et M représentent respectivement le nombre de sources et de microphones. $a_{mn}(t)$ est la réponse impulsionnelle de la pièce (RIR) de la n-ième source au m-ième microphone. $n_m(t)$ est un bruit blanc gaussien additif au microphone m.

Le BSS avec des mélanges convolutifs est traité habituellement dans chaque bande fréquenctielle à l'aide de la STFT (Short-Time-Fourier-Transform) et repose sur l'hypothèse de bande étroite [1]. Le mélange convolutif peut alors être approché par des mélanges instantanés (approximation multiplicative) dans chaque bande fréquentielle, de telle sorte que les méthodes de type ICA (Independent Component Analysis) [2], IVA [3], SCA [4] ou NMF (Nonnegative Matrix Factorisation) [5, 6] puissent être appliquées. Récemment, il a été montré dans [7, 8] qu'une approximation convolutive en bande étroite convient mieux au modèle de mélange que l'approximation multiplicative dans le cadre de SCA. Cette approximation convolutive en bande étroite s'écrit :

$$\tilde{\mathbf{x}}(f,\tau) = \sum_{n=1}^{N}\sum_{k=1}^{K} \tilde{\mathbf{h}}_n(f,k)\tilde{s}_n(f,\tau-k), \tag{2}$$

où $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_M]^T$ représente les coefficients STFT des observations. $\tilde{s}_n$ est le coefficient STFT de s_n. $\tilde{\mathbf{h}}_n = \left[\tilde{h}_{1n}, \ldots, \tilde{h}_{Mn}\right]^T$ est le vecteur qui contient les réponses impulsionnelles dans le domaine temps-fréquence associé à la n-ième source. K est la longueur du noyau de convolution dans le domaine temps-fréquence. Une formulation avec les notations tensorielles peut être écrite en concaténant les échantillons temporels et fréquentiels comme suit :

$$\mathsf{X} = \mathsf{H} \star_f \mathsf{S}, \tag{3}$$

où $\mathsf{X} \in \mathbb{C}^{M\times L_F\times L_T}$, $\mathsf{S} \in \mathbb{C}^{N\times L_F\times L_T}$ et $\mathsf{H} \in \mathbb{C}^{M\times N\times K\times L_F}$. L_F et L_T sont respectivement le nombre d'échantillons temporels et fréquentiels dans le domaine temps-fréquence. $\star_f$ désigne le mélange convolutif fréquence par fréquence [7].

En raison de l'ambiguïté due à l'indétermination de permutation, la séparation de chaque fréquence est suivie d'une étape d'alignement afin de regrouper les composantes fréquentielles provenant de la même source. Une telle étape d'alignement ne se produit pas dans les méthodes à base d'IVA ou de NMF car ces méthodes tiennent compte des informations inter-bandes lors de la séparation.

Ce travail est soutenu par le projet ANR wavegraph.

Dans cet article, nous établissons d'abord un lien entre l'IVA et le SCA à travers la parcimonie structurée dans la section 2, puis proposons dans la section 3 une approche basée sur la parcimonie structurée et l'approximation convolutive qui explore les informations inter-bandes et évite le problème de permutation. Les évaluations numériques sont effectuées dans la section 4 et nous concluons le document dans la section 5.

2 IVA et la parcimonie structurée

2.1 De l'IVA au Group-Lasso

L'IVA est une généralisation de l'ICA qui évite le problème de permutation en exploitant la corrélation d'ordre supérieur entre les bandes fréquentielles. Dans IVA, les coefficients des sources sont modélisés comme une variable vectorielle multivariée composée de toutes les bandes fréquentielles $\mathsf{S}_{n,\tau} \in \mathbb{C}^{L_F}$. Une distribution multivariée supergaussienne sphérique est utilisée comme a priori sur les coefficients temps-fréquence des sources $p(\mathsf{S}_{n,\tau})$. Avec une distribution laplacienne sphérique [3] l'a priori s'écrit :

$$p(\mathsf{S}_{n,\tau}) = \rho \exp\left(-\sqrt{\sum_{f=1}^{L_F} |\frac{\mathsf{S}_{n,f,\tau}}{r_{n,\tau}}|^2}\right), \tag{4}$$

où ρ est un terme de normalisation et $r_{n,\tau}$ est la variance uniforme sur les bandes fréquentielles, ce qui correspond au spectre de puissance de chaque source. Cette distribution symétrique sphérique des sources assure une corrélation d'ordre supérieur entre les bandes fréquentielles [3].

Si on suppose que $r_{n,\tau} = r$, alors la log-vraisemblance négative de la distribution a priori (4) devient une pénalité de type Group-Lasso [9] :

$$-\log p(\mathsf{S}_{n,\tau}) = G(\mathsf{S}) \propto \sum_{n,\tau} \sqrt{\sum_{f=1}^{L_F} |\mathsf{S}_{n,f,\tau}|^2} = \sum_n \|\mathsf{S}_n\|_{2,1}. \tag{5}$$

Par conséquent, un cadre d'optimisation de type "SCA" couplé avec l'approximation convolutive donne :

$$\min_{\mathsf{H},\mathsf{S}} \frac{1}{2}\|\mathsf{X} - \mathsf{H} \star_f \mathsf{S}\|_F^2 + \lambda \sum_n \|\mathsf{S}_n\|_{2,1} + p(\mathsf{H}), \tag{6}$$

où l'attache aux données nous permet de prendre en compte l'erreur d'approximation. $p(\mathsf{H})$ est une fonction indicatrice de sorte que $\sqrt{\sum_{m,k} |\mathsf{H}_{m,n,k,f}|^2} = 1$ pour éviter les solutions triviales causées par l'indétermination d'échelle. λ est un hyperparamètre qui balance entre l'attache aux données et le terme de pénalité $\sum_n \|\mathsf{S}_n\|_{2,1}$.

2.2 Minimisation

Le problème d'optimisation (6) étant non-différentiable, il peut être résolu avec une méthode itérative basée sur l'opérateur proximal, défini ci-après.

Définition 1. *Soit ψ une fonction convexe semi-continue inférieurement. L'opérateur de proximal associé est donné par :*

$$\text{prox}_{\psi}(\mathbf{z}) = \underset{\mathbf{u}}{\text{argmin}} \frac{1}{2}\|\mathbf{z} - \mathbf{u}\|_2^2 + \psi(\mathbf{u}). \tag{7}$$

Proposition 1. *L'opérateur proximal associé à la norme mixte $\text{prox}_{\lambda\|\cdot\|_{2,1}}$ est donné par un seuillage élément par élément comme suit :*

$$\hat{\mathsf{S}}_{n,f,\tau} = \mathsf{S}_{n,f,\tau}\left(1 - \frac{\lambda}{\|\mathsf{S}_{n,\tau}\|_2}\right)^+, \tag{8}$$

où $(z)^+ = \max(0, z)$.

Proposition 2. *L'opérateur proximal associé à la fonction indicatrice $p(\mathsf{H})$ est réduit à un opérateur de projection $\mathcal{P}(\mathsf{H})$:*

$$\hat{\mathsf{H}}_{m,n,k,f} = \frac{\mathsf{H}_{m,n,k,f}}{\sqrt{\sum_{m,k} |\mathsf{H}_{m,n,k,f}|^2}}. \tag{9}$$

Grâce à ces opérateurs, le problème (6) peut être optimisé à l'aide d'une méthode de minimisation alternée de type PALM [10].

3 Analyse locale convolutive en vecteurs indépendants (LCIVA)

Dans [11] il est mentionné que l'IVA ne convient pas aux signaux multi-composantes tels que les signaux de musiques. Afin d'éviter cet inconvénient, nous proposons de remplacer l'opérateur du Groupe-Lasso (8) par l'opérateur du Windowed-Group-Lasso (WGL) pour exploiter les structures locales. L'opérateur WGL $\mathbb{S}_{\lambda}^{\text{WGL}}$ est défini comme suit [9] :

$$\hat{\mathsf{S}} = \mathbb{S}_{\lambda}^{\text{WGL}}(\mathsf{S}) \text{ tel que pour tout } n, f, \tau$$
$$\hat{\mathsf{S}}_{n,f,\tau} = \mathsf{S}_{n,f,\tau}\left(1 - \frac{\lambda}{\sqrt{\sum\limits_{(f',\tau')\in\mathcal{N}_{n,f,\tau}} |\mathsf{S}_{n,f',\tau'}|^2}}\right)^+, \tag{10}$$

où $\mathcal{N}_{n,f,\tau}$ est un voisinage prédéfini de l'index temps-fréquence (f, τ) pour la n-ième source. Nous choisissons ici un voisinage impliquant un groupement en fréquence pour chaque index temporel τ. Ainsi, lorsque la taille du voisinage est égale à 100% du nombre de bandes fréquentielles, le WGL se réduit à l'opérateur proximal du Group-Lasso et nous retrouvons alors la solution du problème (6).

Construit sur l'algorithme PALM utilisé pour résoudre (6), l'algorithme proposé pour l'analyse locale convolutive en vecteurs indépendants (LCIVA – Local Convolutive Independent Vector Analysis) est donné dans l'algorithme 1 où L_A et L_S sont les constantes de Lipschitz qui peuvent être calculées avec Power Iteration [7].

Algorithm 1: LCIVA

Initialisation : $\mathcal{S} \in \mathbb{C}^{N \times L_F \times L_T}$, $\mathsf{H} \in \mathbb{C}^{M \times N \times K \times L_F}$;
repeat
  Mettre à jour les coefficients des sources par le descente de gradient suivi par l'opérateur de proximal (10) :
    — Mettre à jour la constante de Lipschitz L_S ;
    — $\mathsf{S} \leftarrow \mathsf{S} - \left(\mathsf{H}^H \star_f (\mathsf{H} \star_f \mathsf{S} - \mathsf{X})\right)/L_S$;
    — $\mathsf{S} \leftarrow \mathrm{prox}_{\lambda}^{\mathrm{WGL}}(\mathsf{S})$;
  Mettre à jour les noyaux de mélange par le descente de gradient suivi par la projection de normalisation (9) :
    — Mettre à jour la constante de Lipschitz L_A ;
    — $\mathsf{H} \leftarrow \mathsf{H} - \left((\mathsf{H} \star_f \mathsf{S} - \mathsf{X}) \star_f \mathsf{S}^H\right)/L_A$;
    — $\mathsf{H} \leftarrow \mathcal{P}(\mathsf{H})$;
until *convergence*;

4 Expériences

Nous prenons la configuration stéréo ($M = 2$) qui est un scénario général pour les signaux de musique [11]. La réverbération est $\mathrm{RT}_{60} = 250$ ms et la distance entre les microphones est 4 cm. Les RIRs sont simulés avec [12]. Nous prenons les sources en trois catégories (musique, voix et batterie) et les regroupons en quatre combinaisons (musique+musique, musique+voix, musique+batterie et parole+batterie). Nous prenons ici le nombre de sources égal à 2 ($N = 2$) pour révéler facilement les performances de différentes catégories. Pour chaque combinaison, nous prenons quatre groupes de sources de [13]. Chaque source est tronquée à 8 s. La fréquence d'échantillonnage est $14,7$ kHz et la longueur de la fenêtre STFT est $69,7$ ms avec un décalage de $17,4$ ms (75% de recouvrement).

La performance de la LCIVA proposée est comparée avec les méthodes C-PALM [7], Full-rank (Full) [14], Bin-wise (Bin) [4], IVA (auxIVA) [15] et MNMF [5]. Pour LCIVA et C-PALM, la longueur du noyau est $K = 6$. L'hyperparamètre λ est choisi de telle sorte que le niveau de parcimonie des sources estimées soit d'environ 2% qui correspond à la situation sans bruit. La figure 1 montre les résultats de l'évaluation avec les critères [16]. LCIVA est évalué avec trois tailles de voisinage différentes.

Étant une méthode basée sur l'approximation convolutive, en termes de SDR et de SIR, C-PALM offre les meilleures performances pour musique+voix, surpasse les méthodes existantes (Full, Bin, auxIVA et MNMF) pour voix + batterie, mais présente une performance médiocre pour musique+musique. Ces observations montrent que C-PALM est particulièrement intéressant pour les signaux de parole comme remarqué dans [7]. LCIVA a les meilleures performances en termes de SDR, SIR et ISR lorsqu'une batterie est dans le mélange. Sachant qu'il existe généralement de fortes corrélations entre les composantes fréquentielles des signaux de batterie, ces remarques montrent les avantages du LCIVA grâce aux informations inter-bandes. Cette conclusion est également soutenue par la remarque que LCIVA (avec un voisinage de 78%) a les meilleures performances en termes de SDR et SIR pour musique+musique.

Le niveau de parcimonie est le pourcentage d'éléments non nuls dans les données.

5 Conclusion

Pour le problème de BSS avec des mélanges convolutifs, nous montrons que la distribution a priori des sources dans la méthode IVA peut être réécrite en terme de pénalité avec une norme mixte, et peut ensuite être associée à l'approximation convolutive dans le cadre d'optimisation. Nous généralisons ensuite l'algorithme de type C-PALM obtenu avec l'opérateur du Windowed-Group-Lasso pour exploiter les informations inter-bandes locales. Les évaluations numériques montrent les avantages de LCIVA grâce à l'approximation convolutive et aux informations inter-bandes.

Références

[1] Walter Kellermann and Herbert Buchner, "Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. IEEE, 2003, vol. 2, pp. 1278–1282.

[2] Hiroshi Sawada, Shoko Araki, Ryo Mukai, and Shoji Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[3] Taesu Kim, Hagai T Attias, Soo-Young Lee, and Te-Won Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 1, pp. 70–79, 2007.

[4] Hiroshi Sawada, Shoko Araki, and Shoji Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 516–527, 2011.

[5] Hideyuki Sawada, Hirokazu Kameoka, Shunsuke Araki, and Naonori Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 971–982, 2013.

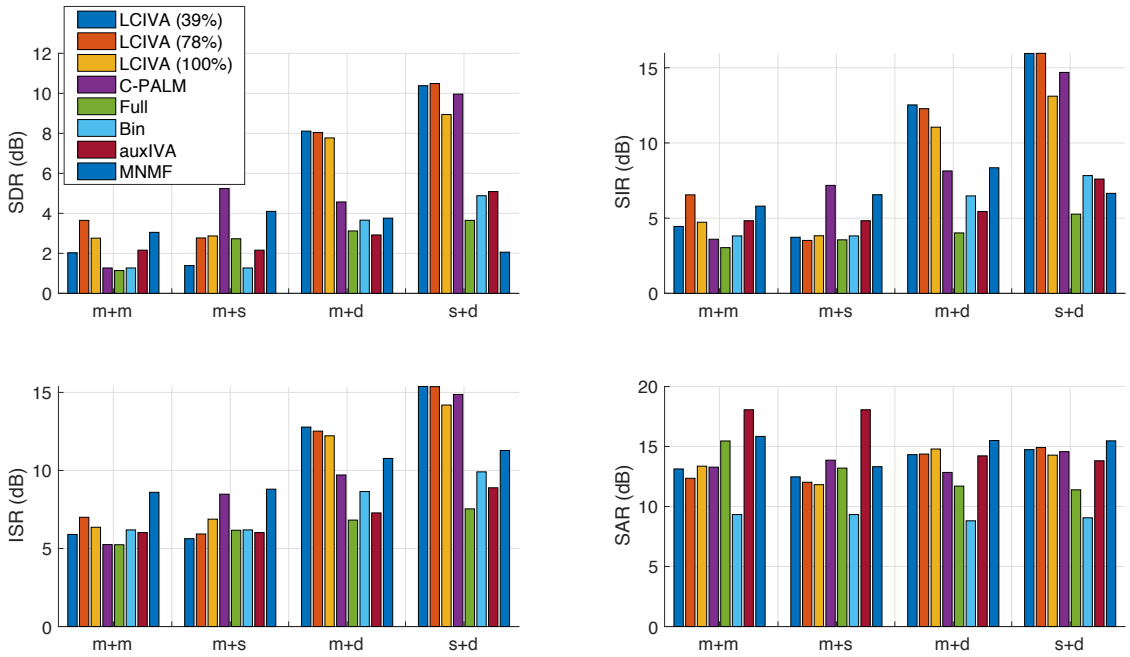

FIGURE 1 – Résultats de séparation avec différentes méthodes pour les quatre combinaisons : musique+musique (m+m), musique+voix (m+v), musique+batterie (m+b) et voix+batteire (v+b).

[6] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, 2010.

[7] Fangchen Feng and Matthieu Kowalski, "Underdetermined reverberant blind source separation : Sparse approaches for multiplicative and convolutive narrowband approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 442–456, 2019.

[8] Xiaofei Li, Laurent Girin, Sharon Gannot, and Radu Horaud, "Multichannel speech separation and enhancement using the convolutive transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

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

[10] Jérôme Bolte, Shoham Sabach, and Marc Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, pp. 1–36, 2013.

[11] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 9, pp. 1622–1637, 2016.

[12] Eric A Lehmann and Anders M Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[13] M Vinyes, "MTG MASS database," `http://www.mtg.upf.edu/static/mass/resources`, 2008.

[14] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1830–1840, 2010.

[15] Nobutaka Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 189–192.

[16] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.