Séparation de sources audio en milieu réverbérant : Factorisation en matrices non-négatives et représentation temporelle du mélange convolutif

Simon Leglaive*, Roland Badeau, Gaël Richard

LTCI, Télécom ParisTech, Université Paris-Saclay
46 rue Barrault, 75013, Paris, France
prénom.nom@telecom-paristech.fr

Résumé – Cet article traite du problème de séparation de sources audio sous-déterminé pour les mélanges réverbérants multicanaux. Nous visons une application semi-aveugle où les filtres de mélange sont connus. La méthode proposée consiste à travailler directement avec les signaux temporels du mélange. Cette approche permet de représenter de façon exacte le processus de mélange convolutif, elle est donc adaptée pour la séparation de mélanges fortement réverbérants. Les signaux sources sont quant à eux représentés dans le domaine de la transformée en cosinus discrète modifiée, en utilisant un modèle gaussien basé sur la factorisation en matrices non-négatives. L'inférence des sources repose sur un algorithme espérance-maximisation variationnel. Nous montrons expérimentalement l'intérêt d'utiliser conjointement une représentation temporelle du mélange convolutif et un modèle de source basé sur la factorisation en matrices non-négatives.

Abstract – This paper addresses the problem of multichannel audio source separation in under-determined reverberant mixtures. We target a semi-blind scenario assuming that the mixing filters are known. The proposed method consists in working directly with the time-domain mixture signals. This approach makes it possible to accurately represent the convolutive mixing process, it is therefore suitable for the separation of highly reverberant mixtures. The source signals are represented in the modified discrete cosine transform domain with a Gaussian model based on non-negative matrix factorization (NMF). Source inference is based on a variational expectation-maximization algorithm. We experimentally show the advantage of using a time-domain representation of the convolutive mixture and a source model based on NMF.

1 Introduction

La séparation de sources audio multicanale vise à extraire un ensemble de signaux sources à partir d'un mélange capté par plusieurs microphones. Cet article se focalise sur les mélanges sous-déterminés (plus de sources que de microphones) et réverbérants (également appelés convolutifs). Nous nous placerons dans un cas semi-aveugle en supposant la connaissance des filtres de mélange.

Dans un contexte sous-déterminé il est commun de travailler avec une représentation temps-fréquence (TF) des signaux sources. Cela permet en effet d'exploiter au travers de modèles la structure bien particulière des signaux audio dans le plan TF. L'analyse en composantes parcimonieuses [1] et les approches par modélisation de la variance [2] sont deux courants importants en séparation de sources. Pour la seconde catégorie, des techniques de factorisation en matrices non-négatives (NMF en anglais) sont fréquemment employées afin de représenter les caractéristiques TF des sources [3, 4, 5, 6, 7].

*Ce travail a été en partie financé par l'Agence Nationale de la Recherche, dans le cadre du projet EDISON 3D (ANR-13-CORD-0008-02)

Par conséquent, de nombreuses méthodes modélisent également le mélange dans un domaine TF. Plus précisément il est commun de considérer que le mélange convolutif devient instantané dans chaque canal fréquentiel de la transformée de Fourier à court-terme (TFCT) [8, 9]. Cette approximation n'est cependant valide que pour des mélanges faiblement réverbérants.

Nous proposons donc dans cet article une méthode de séparation basée sur l'observation des signaux temporels en sortie des microphones, permettant ainsi une représentation exacte du caractère convolutif du mélange. Nous gardons néanmoins une modélisation TF des sources, basée sur la transformée en cosinus discrète modifiée (MDCT en anglais). Les coefficients MDCT d'une source sont en effet modélisés comme des variables aléatoires gaussiennes centrées dont la variance s'exprime par un modèle NMF. Notre approche est comparée expérimentalement à deux méthodes de l'état de l'art. La première s'appuie également sur un modèle de source gaussien basé sur la NMF mais le mélange convolutif est considéré comme étant instantané dans le domaine de la TFCT [10]. La seconde méthode repose quant à elle sur une représentation exacte

du mélange convolutif dans le domaine temporel, mais le modèle de source exploite uniquement la parcimonie des coefficients TF [11]. Nous montrons expérimentalement que l'approche proposée dans cet article conduit à de meilleures performances. Ces résultats illustrent donc l'intérêt d'utiliser conjointement une représentation temporelle du mélange convolutif et un modèle de source basé sur la NMF.

Le modèle est introduit en section 2. L'inférence variationnelle est décrite en section 3. Nous présentons les résultats expérimentaux en section 4 avant de conclure en section 5.

2 Modèle

Soit $s_j(t) \in \mathbb{R}$, $t = 0, ..., L_s - 1$, $j = 1, ..., J$, le signal source j et $a_{ij}(t) \in \mathbb{R}$, $t = 0, ..., L_a$, $i = 1, ..., I$, le filtre de mélange entre la source j et le microphone i. Soit $T = L_s + L_a - 1$. Le signal $x_i(t) \in \mathbb{R}$ capté par le microphone i s'exprime pour $t = 0, ..., T-1$ par :

$$x_i(t) = \sum_{j=1}^{J} y_{ij}(t) + b_i(t), \tag{1}$$

où $y_{ij}(t) = [a_{ij} \star s_j](t)$ est la j-ème source image associée au microphone i, avec $\star$ l'opérateur de convolution discrète, et $b_i(t)$ est un bruit blanc gaussien :

$$b_i(t) \sim \mathcal{N}(0, \sigma_i^2). \tag{2}$$

De façon similaire à l'approche proposée dans [12], chaque signal $s_j(t)$ est représenté par un ensemble de coefficients temps-fréquence de synthèse $\{s_{j,fn} \in \mathbb{R}\}_{(f,n)\in\mathcal{B}}$ avec $\mathcal{B} = \{0, ..., F-1\} \times \{0, ..., N-1\}$:

$$s_j(t) = \sum_{(f,n)\in\mathcal{B}} s_{j,fn}\psi_{fn}(t). \tag{3}$$

Nous choisissons de travailler avec une représentation des sources dans le domaine de la MDCT [13]. L'atome $\psi_{fn}(t)$ est donc fixé en conséquence. La MDCT possède l'avantage d'être une transformation à échantillonnage critique. Contrairement à la TFCT qui est un transformation redondante, la MDCT permet d'avoir le même nombre de coefficients dans le domaine transformé que d'échantillons dans le domaine temporel. Ce choix permet donc de limiter la charge de calcul. D'après cette représentation, une source image s'écrit avec $g_{ij,fn}(t) = [a_{ij} \star \psi_{fn}](t)$:

$$y_{ij}(t) = [a_{ij} \star s_j](t) = \sum_{(f,n)\in\mathcal{B}} s_{j,fn}\, g_{ij,fn}(t). \tag{4}$$

Chaque coefficient temps-fréquence de synthèse d'une source est ensuite modélisé comme une variable aléatoire suivant une distribution gaussienne centrée dont la variance s'exprime par un modèle NMF [3] :

$$s_{j,fn} \sim \mathcal{N}(0, v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}), \tag{5}$$

avec $\mathbf{W}_j \in \mathbb{R}_+^{F\times K_j}$, $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N}$ et K_j est le rang de la factorisation.

3 Inférence variationnelle

Soit $\mathbf{x} = \{x_i(t)\}_{i,t}$ l'ensemble des données observées, $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$ l'ensemble des variables latentes et $\boldsymbol{\theta} = \big\{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{\sigma_i^2\}_i\big\}$ les paramètres du modèle. Nous rappelons que les filtres de mélange $\{a_{ij}(t)\}_{i,j,t}$ sont supposés connus. Nous souhaitons inférer les variables latentes d'après leur moyenne a posteriori, en s'appuyant sur une estimation des paramètres au sens du maximum de vraisemblance :

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star}[\mathbf{s}], \quad \text{avec} \quad \boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}). \tag{6}$$

D'après le modèle présenté précédemment, nous pouvons montrer que la distribution a posteriori des sources est gaussienne mais paramétrée par une matrice de covariance pleine et de trop grande dimension pour être implémentée en pratique. C'est pourquoi nous nous orientons vers une technique d'inférence variationnelle qui va permettre de contraindre la matrice de covariance a posteriori à être diagonale. Soit $\mathcal{F}$ un ensemble de densités de probabilité sur les variables latentes $\mathbf{s}$. Pour toute densité de probabilité $q \in \mathcal{F}$, l'inférence variationnelle consiste à optimiser un critère appelé énergie variationnelle libre et défini par [14] :

$$\mathcal{L}(q;\boldsymbol{\theta}) = \langle \ln\big(\, p(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) \,/\, q(\mathbf{s})\,\big)\rangle_q, \tag{7}$$

avec $<\cdot>_q$ l'espérance mathématique prise par rapport à la distribution q. Plus précisément nous allons mettre en œuvre un algorithme espérance-maximisation variationnel (VEM) qui consiste à itérer deux étapes jusqu'à convergence : l'étape E où on calcule $q^\star = \arg\max_{q\in\mathcal{F}} \mathcal{L}(q;\boldsymbol{\theta}^\star)$ et l'étape M où les paramètres sont mis à jour suivant $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star;\boldsymbol{\theta})$. En pratique nous nous appuierons sur l'approximation de champ moyen qui contraint la famille variationnelle $\mathcal{F}$ à l'ensemble des densités qui se factorisent sous la forme : $q(\mathbf{s}) = \prod_{j,f,n} q_{jfn}(s_{j,fn})$. Sous cette approximation nous pouvons montrer que la densité de probabilité sur une variable $s \in \mathbf{s}$ qui maximise l'énergie variationnelle libre satisfait :

$$\ln q^\star(s) \stackrel{c}{=} \langle \ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})\rangle_{q(\mathbf{s}\setminus s)}, \tag{8}$$

où $\stackrel{c}{=}$ représente l'égalité à une constante additive près et $\mathbf{s}\setminus s$ représente l'ensemble des variables latentes privé de s.

Estimation des sources. Sous l'approximation de champ moyen, l'estimation de la j-ème source dans le domaine TF est donnée par :

$$\hat{s}_{j,fn} = \langle s_{j,fn}\rangle_q. \tag{9}$$

Le signal source dans le domaine temporel $\hat{s}_j(t)$ est ensuite reconstruit par MDCT inverse et la source image $\hat{y}_{ij}(t)$ est obtenue par convolution avec le filtre de mélange associé au microphone i : $\hat{y}_{ij}(t) = [a_{ij} \star \hat{s}_j](t)$.

Log-vraisemblance des données complètes. A partir des équations (1) à (5), la log-vraisemblance des données complètes $\ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{s};\boldsymbol{\theta}) + \ln p(\mathbf{s};\boldsymbol{\theta})$ s'écrit :

$$\ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) \stackrel{c}{=} -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{1}{\sigma_i^2}\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right] - \frac{1}{2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln(v_{j,fn}) + \frac{s_{j,fn}^2}{v_{j,fn}}\right]. \quad (10)$$

Étape E. D'après les équations (8) et (10) nous pouvons montrer que $q^\star_{jfn}(s_{j,fn}) = N(s_{j,fn};\hat{s}_{j,fn},\gamma_{j,fn})$ avec :

$$\gamma_{j,fn} = \left(\frac{1}{v_{j,fn}} + \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1} g_{ij,fn}^2(t)\right)^{-1}; \quad (11)$$

$$\hat{s}_{j,fn} = \hat{s}_{j,fn} - \gamma_{j,fn} d_{j,fn}\,; \quad (12)$$

$$d_{j,fn} = \frac{\hat{s}_{j,fn}}{v_{j,fn}} - \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1} g_{ij,fn}(t)\left(x_i(t) - \sum_{j'=1}^{J}\hat{y}_{ij'}(t)\right). \quad (13)$$

L'équation (12) est valide si les coefficients $\hat{s}_{j,fn}$ sont mis à jour séquentiellement [1]. Cependant nous pouvons montrer que $d_{j,fn} = \partial(-\mathcal{L}(q^\star;\boldsymbol{\theta}))/(\partial\hat{s}_{j,fn})$ où l'expression de $\mathcal{L}(q^\star;\boldsymbol{\theta})$ sera détaillée au prochain paragraphe. L'équation (12) correspond donc à une mise à jour par descente de coordonnées sur l'opposé de l'énergie variationnelle libre. Nous utiliserons en pratique la méthode du gradient conjugué avec préconditionnement pour optimiser ce même critère [15]. Ce choix nous permet en effet de diminuer le temps de calcul requis par l'étape E.

Énergie variationnelle libre. D'après les équations (7), (10) et les résultats de l'étape E, l'expression de l'énergie variationnelle libre est donnée par :

$$\mathcal{L}(q^\star;\boldsymbol{\theta}) \stackrel{c}{=} -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{e_i(t)}{\sigma_i^2}\right] - \frac{1}{2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln\left(\frac{v_{j,fn}}{\gamma_{j,fn}}\right) + \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{v_{j,fn}}\right], \quad (14)$$

où $e_i(t) = \left\langle \left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2 \right\rangle_{q^\star}$ se développe d'après l'approximation de champ moyen et l'équation (4) comme :

$$e_i(t) = \left(x_i(t) - \sum_{j=1}^{J}\hat{y}_{ij}(t)\right)^2 + \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\gamma_{j,fn} g_{ij,fn}^2(t). \quad (15)$$

1. Nous pouvons mentionner que si l'on injecte (11) et (13) dans (12), les coefficients $\hat{s}_{j,fn}$ disparaissent du membre de droite de (12).

Étape M. L'étape M consiste à maximiser (ou seulement augmenter) l'énergie variationnelle libre définie à l'équation (14) par rapport aux paramètres $\boldsymbol{\theta}$. Annuler la dérivée de ce critère par rapport σ_i^2 conduit à la mise à jour suivante avec $e_i(t)$ défini à l'équation (15) :

$$\sigma_i^2 = \frac{1}{T}\sum_{t=0}^{T-1} e_i(t). \quad (16)$$

Concernant les paramètres de NMF, nous pouvons reconnaître à l'équation (14) la divergence d'Itakura-Saito [3] entre la moyenne a posteriori du spectrogramme de puissance des sources $\langle s_{j,fn}^2\rangle_{q^\star} = \hat{s}_{j,fn}^2 + \gamma_{j,fn}$ et $v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}$ (à une constante additive près par rapport aux paramètres NMF). Par conséquent les paramètres de NMF seront mis à jour par quelques itérations des règles multiplicatives données dans [3]. Cette approche permet d'augmenter l'énergie variationnelle libre [16].

4 Expériences

Les expériences sont réalisées à partir de signaux sources fournis par la base de données MASS [2]. Nous considérons 8 mélanges stéréo échantillonnés à 16 kHz et créés à partir de filtres de mélange simulés avec la boîte à outils Roomsimove [3]. Chaque mélange dure entre 12 et 28 secondes et contient entre 3 et 5 sources spatialement disjointes. Le temps de réverbération de la salle est de 256 ms. Les performances de séparation sont évaluées sur les signaux sources mono reconstruits. Nous utilisons les mesures standard de rapport signal sur distorsion (SDR), interférence (SIR) et artéfact (SAR). Ces mesures s'expriment en décibels (dB) et sont définies dans [17]. Elles sont calculées grâce à la boite à outils BSS Eval [4].

La méthode de séparation proposée dans cet article est comparée à deux approches de l'état de l'art. La première approche considère également un modèle de source gaussien basé sur la NMF mais le mélange convolutif est approché par un mélange instantané dans le domaine de la TFCT [10] [5]. La seconde méthode se base quant à elle sur une représentation exacte du mélange convolutif dans le domaine temporel avec une régularisation ℓ_1 sur les coefficients TF des sources [11]. Cette méthode repose donc sur un modèle basé sur la parcimonie des sources dans le domaine TF.

Pour toutes les méthodes comparées, nous utilisons une fenêtre d'analyse/synthèse TF sinusoïdale d'une longueur de 128 ms. Pour les méthodes basées sur la NMF, le rang de factorisation est arbitrairement fixé à $K_j = 10$ pour toutes les sources. Les filtres de mélange sont supposés connus et fixés pour l'ensemble des méthodes comparées

2. http://mtg.upf.edu/download/datasets/mass
3. http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip
4. http://bass-db.gforge.inria.fr/bss_eval/ (Version 3.0)
5. Les paramètres NMF sont mis à jour par des règles multiplicatives comme proposé ensuite par les auteurs dans [18].

TABLE 1 – Résultats de séparation moyens en dB.

	SDR	SIR	SAR
Ozerov et Févotte [10]	1.7	8.5	4.9
Kowalski et al. [11]	5.5	11.7	8.8
Méthode proposée	**6.7**	**12.5**	**9.5**

tandis que tous les autres paramètres sont initialisés de façon aveugle. Comme la réponse impulsionnelle d'un filtre est plus longue que la fenêtre d'analyse TF, elle est tronquée avant de calculer par transformée de Fourier discrète la réponse en fréquence requise par la méthode [10].

Les résultats de séparation moyennés sur l'ensemble des sources de la base sont présentés dans le tableau 1. Nous observons que la méthode proposée conduit aux meilleures performances pour toutes les mesures de la qualité de séparation. Ces résultats montrent donc l'importance de la modélisation exacte du processus de mélange convolutif et l'intérêt d'exploiter la structure particulière des signaux sources dans le domaine TF par l'intermédiaire d'un modèle NMF et non seulement par un modèle de parcimonie. Des exemples audio et le code Matlab pour la méthode proposée sont disponibles en ligne [6].

5 Conclusion

Nous avons proposé dans cet article une méthode de séparation de sources basée sur l'exploitation des signaux temporels en sortie des microphones. Cette approche permet de représenter de façon exacte le processus de mélange convolutif, elle est donc adaptée pour la séparation de mélanges fortement réverbérants. Les résultats expérimentaux ont montré l'intérêt de cette méthode ainsi que l'importance d'exploiter un modèle de source basé sur la NMF. Par la suite nous nous focaliserons sur l'élaboration d'une méthode totalement aveugle exploitant des a priori sur la réponse impulsionnelle des filtres de mélange. Nous pourrons par exemple considérer des a priori similaires à ceux proposés dans [19] afin de promouvoir la parcimonie et la décroissance exponentielle des filtres en temporel.

Références

[1] R. Gribonval et M. Zibulevsky, "Sparse Component Analysis," dans *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, P. Comon et C. Jutten, Eds., 2010, pages 367–420.

[2] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, et M. E. Davies, "Probabilistic modeling paradigms for audio source separation," *Machine Audition : Principles, Algorithms and Systems*, pages 162–185, 2010.

[3] C. Févotte, N. Bertin, et J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pages 793–830, 2009.

[4] T. Virtanen, A. T. Cemgil, et S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," dans *Actes IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pages 1825–1828.

[5] A. Liutkus, D. Fitzgerald, et R. Badeau, "Cauchy nonnegative matrix factorization," dans *Actes IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2015, pages 1–5.

[6] U. Şimşekli, A. Liutkus, et A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pages 2289–2293, 2015.

[7] K. Yoshii, K. Itoyama, et M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," dans *Actes IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pages 51–55.

[8] Y. Avargel et I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pages 337–340, 2007.

[9] L. Parra et C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pages 320–327, 2000.

[10] A. Ozerov et C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pages 550–563, 2010.

[11] M. Kowalski, E. Vincent, et R. Gribonval, "Beyond the narrowband approximation : Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pages 1818–1829, 2010.

[12] C. Févotte et M. Kowalski, "Low-rank time-frequency synthesis," dans *Actes Adv. Neural Inf. Process. Syst.*, 2014, pages 3563–3571.

[13] H. S. Malvar, *Signal Processing with Lapped Transforms*. Artech House, 1992.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] G. H. Golub et C. F. Van Loan, *Matrix computations*. Johns Hopkins University Press, 1996.

[16] C. Févotte et J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural comput.*, vol. 23, no. 9, pages 2421–2456, 2011.

[17] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, et N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010) : Achievements and remaining challenges," *Signal Process.*, vol. 92, pages 1928–1936, 2012.

[18] A. Ozerov, C. Févotte, R. Blouet, et J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," dans *Actes IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pages 257–260.

[19] A. Benichoux, L. S. R. Simon, E. Vincent, et R. Gribonval, "Convex regularizations for the simultaneous recording of room impulse responses," *IEEE Trans. Signal Process.*, vol. 62, no. 8, pages 1976–1986, 2014.

6. perso.telecom-paristech.fr/leglaive/demo-gretsi17.html