Méthode Structurée de décomposition en matrices non-négatives appliquée à la séparation de sources audio

Clément LAROCHE[1,2], Matthieu KOWALSKI[2,3], Hélène PAPADOPOULOS[2], Gaël RICHARD[1]

[1]Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, France

[2]Univ Paris-Sud-CNRS-CentraleSupelec, L2S, France

[3]Parietal project-team, INRIA, CEA-Saclay, France

prénom.nom@telecom-paristech.fr, prénom.nom@lss.supelec.fr

Résumé – Dans cet article, nous proposons une méthode structurée de décomposition en matrices non-négatives visant à utiliser la structure multi-couche des signaux audio. Les signaux audio peuvent être vus comme une superposition de deux couches : la couche tonale (modélisée par des sommes de sinusoïdes évoluant lentement en fréquence et en temps) et la couche transitoire (les sons percussifs, évènements de courtes durées étalés en fréquence). Notre méthode décompose une partie du signal en composantes orthogonales parcimonieuses, bien adaptées pour l'extraction tonale tandis que la partie transitoire est représentée par des bases de décomposition classiques. Les résultats de séparation de sources obtenus sur des signaux réels de musique ont montré que notre approche obtient des résultats similaires à ceux de l'état de l'art.

Abstract – In this paper, we propose a new unconstrained nonnegative matrix factorization method designed to utilize the multilayer structure of audio signals to improve the quality of the source separation. The tonal layer is sparse in frequency and temporally stable, while the transient layer is composed of short term broadband sounds. Our method has a part well suited for tonal extraction which decomposes the signals in sparse orthogonal components, while the transient part is represented by a regular nonnegative matrix factorization decomposition. Experiments on real music data in a source separation context show that such decomposition is suitable for audio signal. Compared with three state-of-the-art harmonic/percussive decomposition algorithms, the proposed method shows competitive performances.

1 Introduction

Introduite par Lee & Seung [1], la factorisation en matrices non-négatives (Non-Negative Matrix Factorization : NMF) est utilisée dans de nombreux domaines. Dans le cas du traitement audio, la NMF a été utilisée pour de la transcription automatique et la séparation de sources [2]. Le but de la NMF est d'approximer une matrice de données $V \in \mathbb{R}_+^{n\times m}$ de la façon suivante, $V \approx \tilde{V} = WH$, où $W \in \mathbb{R}_+^{n\times k}$ et $H \in \mathbb{R}_+^{k\times m}$ et où k est le rang de factorisation généralement choisi tel que $k(n+m) << nm$. Comme la matrice V est redondante, le produit WH représente la version compressée de V, où W constitue le dictionnaire de la décomposition et où H est la matrice d'activation. En pratique, il n'est pas garanti que la décomposition ait un sens physique. Pour résoudre ce problème, deux techniques sont généralement utilisées. La première consiste à utiliser de l'information à priori extraite de la partition ou de fichiers MIDI pour effectuer une NMF supervisée [2]. Une autre stratégie consiste à utiliser des contraintes spécifiques aux caractéristiques du signal à traiter. Par exemple, il est montré dans [3] que forcer la régularité temporelle de la matrice d'activation fait converger la NMF vers un résultat ayant plus de sens physique. De la même façon, dans [4], Canadas & al. ont utilisé quatre contraintes spécifiques dans la décomposition NMF pour séparer les instruments harmoniques des instruments percussifs. Dans ce cas, les quatre hyper-paramètres doivent être optimisés et la valeur optimale de chacun dépend bien souvent du signal traité.

D'autres méthodes cherchent à mettre en avant les propriétés mathématiques des matrices décomposées. La NMF projective (PNMF) et la NMF orthogonale (ONMF) imposent l'orthogonalité entre les colonnes du dictionnaire W. La PNMF est utilisée en traitement d'image [5]. En pratique, la PNMF s'est révélée très efficace pour la classification de données et offre une décomposition bien plus parcimonieuse que la NMF. Ces propriétés intrinsèques sont particulièrement intéressantes pour la séparation de sources audio comme montré dans l'article [4]. Contrairement à la NMF contrainte, la parcimonie et l'orthogonalité sont obtenues de façon sous-jacente, ce qui évite ainsi la phase d'optimisation fastidieuse des hyper-paramètres. Ces approches n'ont cependant pas la flexibilité suffisante pour représenter correctement une scène audio complexe composée de multiples sources.

Dans ce papier, nous proposons une nouvelle approche de séparation de sources audio qui tire profit de la parcimonie de la décomposition PNMF sans pour autant se limiter à des signaux audio très simples. Plus précisément, nous ajoutons à la décomposition PNMF un certain nombre de composantes NMF classiques qui se sont révélées particulièrement pertinentes pour représenter les sons percussifs et les transitoires

du signal. Cette NMF structurée dénommée (SPNMF) a été testée expérimentalement sur des données réelles pour une application de séparation entre des instruments harmoniques et percussifs. Il est important de noter que notre méthode n'utilise aucune contrainte. L'article est organisé de la façon suivante. Dans la Section 2, nous allons décrire et comparer les aspects théoriques de la PNMF et de la ONMF. La SPNMF est présentée en Section 3. En Section 4 nous décrivons notre protocole expérimental et les résultats obtenus. Nous suggérons des conclusions dans la Section 5.

2 Lien entre PNMF et ONMF

Dans cette section nous allons présenter les trois décomposition sus-citées en mettant en avant leurs similarités et leurs différences. Soit V la matrice de données, le problème NMF se pose de la façon suivante :

$$\min_{W,H\geqslant 0} \|V - WH\|^2, \tag{1}$$

où $\|.\|$ correspond à la norme euclidienne.

Le but de la PNMF est de trouver une matrice de projection non-négatives $P \in \mathbb{R}_+^{n\times n}$ telle que $V \approx \tilde{V} = PV$. Dans [6] Yuan & al. ont proposé de chercher la matrice P de la forme $P = WW^T$ où $W \in \mathbb{R}_+^{n\times k}$ avec $k < n$. Le problème peut s'écrire de la façon suivante :

$$\min_{W\geqslant 0} \|V - WW^TV\|^2 \tag{2}$$

La ONMF consiste à résoudre le problème suivant :

$$\min_{W,H\geqslant 0} \|V - WH\|^2 \qquad \text{s.t} \qquad W^TW = I_k \tag{3}$$

Dans cette approche, l'orthogonalité est forcée au cours de l'optimisation.

Par rapport à la NMF, la PNMF (resp. ONMF) inclue la contrainte $H = W^TV$ (resp. $W^TW = I_k$). Si on suppose que V admet une décomposition ONMF sans erreurs, on peut établir le théorème suivant :

Théorème 1 *Soit $V \in \mathbb{R}_+^{n\times m}$, et soit $W_{onmf}, W_{pnmf} \in \mathbb{R}_+^{n\times k}$ les solutions de la décomposition ONMF et PNMF, avec des notations similaires pour les matrices $H \in \mathbb{R}_+^{k\times m}$. On suppose que $k \leq \min(m,n)$, rang(W) =rang$(H) = k$. Alors, à une matrice d'échelle et de permutation près, on a :*

$$W_{onmf} = W_{pnmf}$$

et

$$H_{onmf} = H_{pnmf} = W_{pnmf}^TV.$$

De façon équivalente, à une matrice d'échelle et de permutation près :

$$W_{pnmf}^TW_{pnmf} = I \quad \textit{and} \quad H_{pnmf} = W_{pnmf}^TV.$$

En pratique l'hypothèse $V = WH$ n'est plus vérifiée dès que $k < \min(n,m)$ d'ou l'intéret d'introduire la PNMF et ONMF. Yuan & Oja dans [6] indiquent que l'orthogonalité de W est nécessaire pour obtenir un vrai projecteur. Or, la matrice W obtenue par l'algorithme de PNMF est "presque orthogonale" (i.e., $\|W^TW - I_k\|^2$ est petit). La PNMF et la ONMF donnent donc des résultats similaires, cette remarque nous a encouragé à construire notre méthode autour de la PNMF.

3 NMF projective structurée (SPNMF)

Les instruments harmoniques possèdent des spectres parcimonieux alors que les instruments percussifs ont des spectres beaucoup plus plats. Dans le cas de la PNMF, comme les colonnes de W sont orthogonales, lorsque deux sources se recouvrent dans le domaine Temps-Fréquence (TF) une seule fonction de base va représenter le mélange ce qui n'est pas convenable pour une séparation de sources efficace. Pour palier à ce problème, nous proposons d'ajouter un terme de NMF standard à la PNMF. Le rang de la décomposition est augmenté tel que $k = k' + e$ où e est le nombre de composantes NMF supplémentaires. Nous pouvons nous attendre à ce que la plupart des composantes harmoniques soient transcrites par la partie orthogonale alors que les instruments percussifs seront représentés par les termes NMF classiques. Le modèle s'écrit

$$V \approx \tilde{V} = W_1H_1 + W_2H_2, \tag{4}$$

où W_1H_1 est la partie presque orthogonale de rang k' et où W_2H_2 sont e composantes NMF. On pose ensuite la contrainte de la PNMF : $H_1 = W_1^T(V - W_2H_2)$ ssi $W_1^TW_1 = I$. Le problème s'écrit de la façon suivante :

$$\min_{W_1,W_2,H_2\geq 0} ||V - W_1W_1^T(V - W_2H_2) - W_2H_2||^2. \tag{5}$$

Dans notre cas, e est choisi plus petit que k'. Le but est de représenter une grande partie de l'énergie dans la partie orthogonale pour bénéficier de la décomposition parcimonieuse de la PNMF.

L'optimisation de W_1 est faite en utilisant une méthode similaire à celle utilisé dans [6]. Soit F la fonction de coût associée à (5). La méthode consiste à diviser le gradient $\nabla F(W_1)$ en sa partie positive $[\nabla F(W_1)]^+$ et négative $[\nabla F(W_1)]^-$. Ainsi, on obtient

$$\begin{aligned}[\nabla F(W_1)]^+ = [4(VH_2^TW_2^T + W_2H_2V^T) \\ +2W_1W_1^T(VV^T + W_2H_2H_2^TW_2^T) \\ +2(VV^T + W_2H_2H_2^TW_2^T)W_1W_1^T]W_1\end{aligned} \tag{6}$$

et :

$$\begin{aligned}[\nabla F(W_1)]^- = [4VV^T + 4W_2H_2H_2^TW_2^T \\ +2W_1W_1^T(VH_2^TW_2^T + W_2H_2V^T) \\ +2(VH_2^TW_2^T + W_2H_2V^T)W_1W_1^T]W_1.\end{aligned} \tag{7}$$

Nous pouvons maintenant écrire les règles de mise à jour de la façon suivante :

$$W_1 \leftarrow W_1 \otimes \frac{[\nabla F(W_1)]^-}{[\nabla F(W_1)]^+}.$$

On obtient des expressions similaires pour W_2 et H_2 mais elles ont été omises par soucis de concision.

4 Validation expérimentale

4.1 Protocole et description de la base de données

Nous allons comparer la SPNMF avec la PNMF et la NMF. Pour comparer les algorithmes de façon équitable, nous avons utilisé la distance euclidienne pour l'algorithme de NMF. Les trois algorithmes sont initialisés avec les même matrices aléatoires non-négatives $W_{ini} \in \mathbb{R}^{n\times k}$ et $H_{ini} \in \mathbb{R}_+^{k\times m}$. Le rang de factorisation k est le même pour toutes les méthodes. Les k composantes de la décomposition sont extraites par un filtrage de Wiener et sont associées aux sources audio en utilisant la méthode introduite par Virtanen dans [3]. Le Rapport Signal-Bruit (RSB) est calculé entre la j^{eme} composante estimée $\tilde{x_j}$ et la m^{eme} source originale x_m de la façon suivante :

$$\text{RSB}(m,j) = 10\log(\frac{\tilde{x_j}^2}{(\tilde{x_j} - x_m)^2})$$

La composante j est assignée à la source avec laquelle elle obtient le RSB le plus grand. Les résultats sont ensuite comparés par le Rapport Signal-Distorsion (RSD) le Rapport Signal-Artefact (RSA) et le Rapport Signal-Interférence (RSI) en utilisant la toolbox BSS Eval [7].

La base de données est composée d'extraits de musique stéréophonique. Chaque signal contient des instruments percussifs, des instruments harmoniques, et de la voix. Une partie des données utilisées sont issues de la base SiSec 2010 [8]. Elle est composée de quatre enregistrements de durée comprise entre 14 et 24 s. Pour élargir le nombre de signaux, nous avons ajouté cinq fichiers supplémentaires de la base de données Medley-dB [9]. Le but est d'effectuer une séparation entre les instruments percussifs et harmoniques, aussi de la même manière que [4], nous avons omis la voix. De plus nous avons transformé les signaux stéréo en fichiers monophoniques en effectuant la moyenne des deux canaux. Tout les signaux sont échantillonnés à 44, 1kHz. Nous calculons la Transformée de Fourier à Court Terme (TFCT) avec une fenêtre de Hann de 1024 échantillons et un recouvrement de 50%. La valeur absolue de la TFCT des signaux est notre matrices de données V. Deux tests ont été effectués. Le premier vise à comparer la SPNMF à la NMF et la PNMF sur la base de données complète. Le second correspond à une comparaison entre la SPNMF et trois autres méthodes de l'état de l'art sur les morceaux de la base SiSEC uniquement.

4.2 Comparaison de la SPNMF

Pour le premier test, on choisi $k = 16$, le nombre de composantes NMF classiques pour la SPNMF est $e = 5$. Les résultats sont affichés sur la Figure 1 sous la forme de boites à moustaches. Chacune est composé d'une ligne centrale indiquant la médiane des données, les bords hauts et bas des boites indiquent le 1^{er} et 3^{eme} quartiles, les "moustaches" représentent les valeurs maximales et minimales des données.

Globalement, la SPNMF est plus performante que la NMF du point de vue du RSD et RSI. Le RSA est identique pour les deux méthodes. Les résultats sur des signaux réels indiquent que la SPNMF obtient une meilleure séparation de sources avec moins d'interférences. Quant à elle, la PNMF est moins performante que les autres méthodes. Ce modèle est trop restrictif pour la séparation audio. Les composantes PNMF représentent principalement les instruments harmoniques et seulement une petite partie de l'énergie est contenue dans la partie percussive. Cela explique pourquoi la PNMF obtient des scores élevés pour le RSA et le RSI.

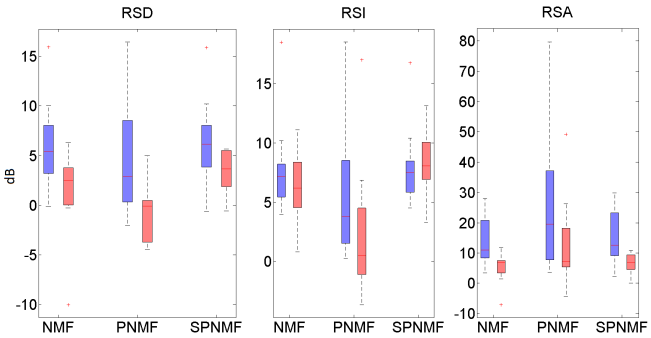

FIGURE 1: RSD, RSI et RSA des sources harmoniques (barre de gauche)/percussives (barre de droite) des 9 signaux test.

4.3 Comparaison à l'état de l'art

Pour le second test, la SPNMF est comparée à trois autres méthodes de décomposition harmonique/percussive de l'état de l'art : HPSS [10], MFS [11] et NMF contrainte [4]. Les trois algorithmes sont non-supervisés.

Le Tableau 1 montre que notre méthode est plus performante que l'état de l'art sur les deux premiers morceaux. Sur ces exemples, les instruments harmoniques ont des transitoires peu marqués. Le troisième morceau est un morceau de rap contenant des sons percussifs proéminents et une partie harmonique de faible amplitude. La SPNMF échoue à séparer les sources qui sont mélangées dans ce que l'algorithme attribue à la partie percussive. La partie harmonique quant à elle ne contient que peu d'énergie et n'obtient pas un RSD et un RSI satisfaisant. Le recouvrement plus important que précédemment dans le domaine TF entre les instruments harmoniques et percussifs peut expliquer cette chute de performance. Sur le dernier fichier, les instruments harmoniques ont des attaques très marquées et comme précédemment, la SPNMF n'obtient pas un bon score. Sur cet exemple, la guitare et la caisse claire sont transcrites par la même composante de la décomposition.

En moyenne, la SPNMF sépare mieux les instruments percussifs que les méthodes MFS et HPSS mais elle obtient de moins bon résultats pour la séparation des instruments harmoniques. Les sources initiales et les sources estimées peuvent être écoutées sur notre site web [1].

1. http ://perso.telecom-paristech.fr/laroche/Article/GRETSI2015/

TABLE 1: Résultats de la séparation de sources harmoniques/percussives (en dB)

	HPSS [10]			MFS [11]			NMF contrainte [4]			SPNMF		
Séparation percussive	RSD	RSI	RSA	RSD	RSI	RSA	RSD	RSI	RSA	RSD	RSI	RSA
T2_01	2.6	**13.2**	1.1	-0.2	-1.5	**8.4**	4.0	6.5	5.7	**4.3**	11.0	5.6
T2_02	2.4	**10.2**	3.4	3.1	8.0	4.9	**5.2**	8.3	**7.5**	5.0	9.7	7.2
T2_03	2.6	**6.9**	4.0	2.5	2.1	**12.3**	**2.8**	2.6	11.1	1.5	6.6	4.0
T2_04	5.5	**11.5**	6.5	6.2	9.6	8.0	**7.5**	10.3	**10.3**	3.6	6.5	7.7
Moyenne	3.2	**10.5**	3.8	2.9	4.6	8.4	**4.9**	7.0	**8.7**	3.6	8.4	6.1
Séparation harmonique	RSD	RSI	RSA	RSD	RSI	RSA	RSD	RSI	RSA	RSD	RSI	RSA
T2_01	9.8	13.8	11.9	7.1	13.8	11.5	**11.0**	**14.8**	13.9	7.3	7.4	**28.3**
T2_02	4.8	6.3	9.8	5.5	**16.2**	11.6	**7.5**	9.3	12.1	6.5	7.8	**12.8**
T2_03	4.8	8.7	6.3	4.6	**11.0**	8.0	**5.0**	9.1	**8.6**	3.8	6.2	8.4
T2_04	5.6	**11.5**	6.7	6.2	9.3	8.7	**7.5**	10.6	**10.5**	4.7	6.6	10.1
Moyenne	6.3	10.1	8.7	5.9	**12.6**	10.0	**7.8**	11.0	11.3	5.6	7.0	**14.9**

4.4 Discussion

Les résultats de la section 4.2 montrent que la décomposition d'un signal audio avec seulement des bases orthogonales (PNMF) n'est pas satisfaisante. Dans le cas de la SPNMF, la partie tonale est extraite par les bases orthogonales alors que la plupart des instruments percussifs sont transcrits par les composantes NMF classiques. Dans la section 4.3, la SPNMF obtient des résultats similaires à l'état de l'art. La méthode est particulièrement efficace quand les instruments harmoniques ont des attaques peu marquées et que les sons percussifs ne sont pas trop proéminents. Dans ce cas, le signal correspond bien au modèle défini par (5). Cependant, si les instruments harmoniques ont des attaques très présentes, la SPMNF est surclassée par les autres méthodes de l'état de l'art car ils ne sont pas bien modélisés par des bases orthogonales.

5 Conclusion

Nous avons montré que la SPNMF est une décomposition non-contrainte prometteuse. Elle est capable d'extraire les sources avec moins d'interférences et une meilleure qualité générale que la NMF. Sur une tache de séparation d'instruments harmoniques/percussifs, la SPNMF obtient des résultats similaires à l'état de l'art. Elle fournit une décomposition structurée du signal avec la couche tonale principalement extraite par les composantes orthogonales tandis que les transitoires sont représentés par les composantes NMF classiques.

Les futurs travaux sur la SPNMF seront dédiés à la conception de stratégies d'initialisation judicieuses. Par exemple W_2 et H_2 peuvent être forcées à représenter de façon plus efficace les instruments percussifs afin d'obtenir une décomposition complètement non supervisée.

Références

[1] D. Lee and S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] S. Ewert and M. Müller, "Score-informed source separation for music signals," *Multimodal music processing*, vol. 3, pp. 73–94, 2012.

[3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[4] F. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–17, 2014.

[5] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. of IEEE IJCNN*, 2008, pp. 1828–1832.

[6] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," *Image Analysis*, pp. 333–342, 2005.

[7] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1462–1469, 2006.

[8] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong, "The 2010 signal separation evaluation campaign : audio source separation," in *Proc. of LVA/ICA*, 2010, pp. 114–122.

[9] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "Medleydb : A multitrack dataset for annotation-intensive mir research," in *proc. of ISMIR*, 2014.

[10] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. of EUSIPCO*, 2008.

[11] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of DAFX*, 2010.