De la faiblesse de rang en temps-fréquence

Ronan HAMON[1], Caroline CHAUX[2], Valentin EMIYA[1]

[1]Aix Marseille Univ, CNRS, Centrale Marseille, LIF, Marseille, France

[2]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

prenom.nom@univ-amu.fr

Résumé – L'hypothèse de faible rang des spectrogrammes a été largement utilisée ces dernières années pour l'analyse des signaux audio. Dans cet article, nous nous intéressons à la représentation temps-fréquence de signaux, et plus particulièrement au rang de la matrice de coefficients issus d'une transformée de Fourier à court terme (TFCT). Nous montrons que ce rang dépend de la convention adoptée pour la TFCT et que le comportement du rang se rapproche de celui observé pour le spectrogramme. Nous appliquons ces résultats à un problème de restauration de données manquantes dans le domaine temps-fréquence. Les coefficients complexes sont reconstruits au travers d'un algorithme proximal prenant en compte l'a priori de faible rang de la matrice des coefficients.

Abstract – The low-rank assumption for spectrogram has been widely used recently for the analysis of audio signals. In this paper, we are interested in time-frequency representations, and more particularly in the rank of the matrix of coefficients obtained after a Short-Time Fourier Transform (STFT). We show that the rank depends on the adopted convention to compute the STFT, and that its behaviour is similar to the one of spectrograms. We apply this result to an audio restoration problem, where missing data are present in the time-frequency plane. The missing complex coefficients are retrieved using a proximal algorithm minimizing the nuclear norm of the STFT matrix, accounting for the low-rank assumption.

1 Introduction

De nombreuses questions liées à l'analyse et au traitement de signaux audio peuvent s'exprimer sous la forme d'un problème de minimisation d'une fonctionnelle, construite en fonction du problème considéré et des connaissances a priori disponibles sur la solution cible. Parmi les différents a priori qu'il est possible de considérer, la contrainte de faible rang est particulièrement utilisée, notamment dans sa version relaxée grâce à l'utilisation de la norme nucléaire ou via une factorisation matricielle [4].

Cette contrainte s'avère pertinente pour le traitement et l'analyse de spectrogrammes issus de signaux audio, par exemple pour des problèmes de séparation de source à l'aide de factorisation en matrices non-négatives [5]. Ces spectrogrammes capturent en effet des structures horizontales caractéristiques du contenu musical pouvant bien se représenter par une approximation de rang faible, comme le montre la figure 1, affichant un spectrogramme d'un son de référence contenant plusieurs occurrences de huit notes générant autour d'une cinquantaine de pics fréquentiels.

La modélisation de spectrogramme par des approximations de faible rang pose néanmoins plusieurs problèmes liés à la reconstruction des phases associés aux modules de chaque coefficients et à la non-linéarité du passage au module des coefficients complexes. La modélisation des coefficients complexes plutôt que de leur module représente ainsi un enjeu important mais généralement compliqué à mettre en œuvre.

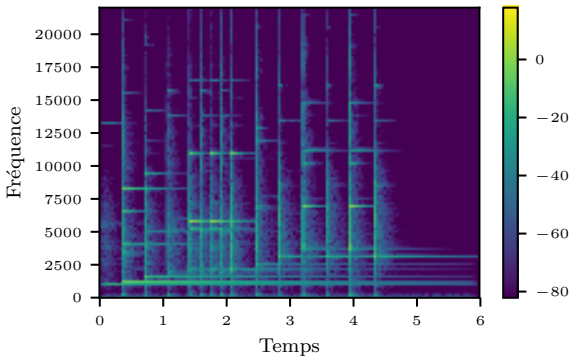

FIGURE 1 – Spectrogramme de *Glockenspiel* (échantillonné à 44.1 KHz ; fenêtre de Hann, taille 2048, recouvrement 75%).

L'objectif de cette contribution est d'étudier dans quelle mesure une décomposition de rang faible fait sens pour des matrices de coefficients temps-fréquence à valeurs complexes. Nous commençons par mettre en évidence dans la Section 2 des propriétés sur le rang des matrices TFCT de signaux synthétiques simples. Dans la Section 3, nous abordons ensuite le cas de l'approximation de rang faible de la TFCT de sons réels à travers plusieurs angles – approximation et prédiction – donnant lieu à plusieurs problèmes d'optimisation accompagnés d'algorithmes de résolution. La Section 4 remet en perspective les résultats obtenus avec les questions posées initialement.

This work was supported by ANR JCJC program MAD ANR-14-CE27-0002.

2 Matrices temps-fréquence de rang K

Notations. Les signaux discrets et les vecteurs (respectivement matrices) sont désignés par des lettres minuscules (respectivement majuscules) en gras. Les matrices peuvent être de dimensions infinies. Des parenthèses sont utilisées pour indexer les coefficients des vecteurs, signaux et matrices. On note $[\![K]\!] = \{0, \dots, K-1\}$ pour $K \in \mathbb{N}$. On désigne par $\mathbf{1}_K$ le vecteur de taille K constitué de 1, et par $\mathbf{1}_{\mathbb{Z}}$ le signal constant égal à 1. Le produit de Hadamard est noté $\odot$.

Matrices temps-fréquence. Considérons des signaux complexes $\mathbf{s} = [\mathbf{s}(m)]_{m\in\mathbb{Z}} \in l^2(\mathbb{Z})$. Soit $\mathbf{h} = [\mathbf{h}(m)]_{m\in\mathbb{Z}} \in l^2(\mathbb{Z})$ une fenêtre, généralement localisée autour de $m = 0$, de transformée de Fourier discrète $\widehat{\mathbf{h}}(\nu) = \sum_{m\in\mathbb{Z}} \mathbf{h}(m)\, e^{-2i\pi\nu m}$, $\nu \in [0, 1[$. Nous rappelons les deux définitions possibles de la transformée de Fourier à court terme (TFCT) définie pour K fréquences discrètes $\nu_k = \frac{k}{K}, k \in [\![K]\!]$ et à des instants $m_n = nh, n \in \mathbb{Z}$ où $h \in \mathbb{N}^*$ est un pas temporel arbitraire.

Définition 2.1 (TFCT, convention passe-bande)**.** En convention passe-bande, la TFCT de $\mathbf{s} \in l^2(\mathbb{Z})$ est définie pour $k \in [\![K]\!]$ et $n \in \mathbb{Z}$ par

$$\mathbf{S}_{\text{BP}}(k, n) = \sum_{m\in\mathbb{Z}} \mathbf{s}(m_n + m)\, \mathbf{h}(m)\, e^{-2i\pi\nu_k m} \qquad (1)$$

Définition 2.2 (TFCT, convention passe-bas)**.** En convention passe-bas, la TFCT de $\mathbf{s} \in l^2(\mathbb{Z})$ est définie pour $k \in [\![K]\!]$ et $n \in \mathbb{Z}$ par

$$\mathbf{S}_{\text{LP}}(k, n) = \sum_{m} \mathbf{s}(m)\, \mathbf{h}(m - m_n)\, e^{-2i\pi\nu_k m} \qquad (2)$$

Proposition 2.1. *Pour* $\mathbf{s} \in l^2(\mathbb{Z})$*, on a*

$$\forall k \in [\![K]\!], n \in \mathbb{Z},\ \mathbf{S}_{LP}(k, n) = \mathbf{S}_{BP}(k, n) \times e^{-2i\pi\nu_k m_n} \quad (3)$$

Elles ont ainsi le même module. Le spectrogramme, constitué du module des coefficients de la TFCT, est donc défini de manière unique.

L'étude d'une sinusoïde, d'un dirac et de la somme de plusieurs sinusoïdes permettent de mettre en évidence les différences importantes entre les matrices temps-fréquence en terme d'approximation de rang faible.

Cas d'une sinusoïde. En partant de la définition (2.1), on montre facilement la proposition 2.2 suivante.

Proposition 2.2. *Soit* $\mathbf{s}(m) = e^{2i\pi f_0 m}$ *où* $f_0 \in [0, 1[$*. Alors*

$$\mathbf{S}_{BP} = \left[\widehat{\mathbf{h}}(\nu_k - f_0)\right]_{k\in[\![K]\!]} \times \left[e^{2i\pi f_0 m_n}\right]^T_{n\in\mathbb{Z}} \qquad (4)$$

$$\textit{et } |\mathbf{S}_{BP}| = \left[\left|\widehat{\mathbf{h}}(\nu_k - f_0)\right|\right]_{k\in[\![K]\!]} \times \mathbf{1}^T_{\mathbb{Z}}. \qquad (5)$$

En convention passe-bande, la TFCT et le spectrogramme d'une sinusoïde pure de fréquence f_0 quelconque sont donc de rang 1. La TFCT se décompose en particulier comme le produit tensoriel entre le spectre $\left[\widehat{\mathbf{h}}(\nu_k - f_0)\right]_{k\in[\![K]\!]}$ et une modulation $\left[e^{2i\pi f_0 m_n}\right]_{n\in\mathbb{Z}}$. Il s'agit ici d'un des rares cas où TFCT et spectrogramme coïncident en terme de rang.

En utilisant le fait que la matrice $\left[e^{-2i\pi\nu_k m_n}\right]_{k\in[\![K]\!],n\in\mathbb{Z}}$ est de rang plein dans la relation (3), on montre qu'en convention passe-bas, la TFCT $\mathbf{S}_{\text{LP}}$ d'une sinusoïde est de rang plein K. On voit ici une différence importante entre les deux conventions de la TFCT : le choix de la convention est alors déterminant si l'on souhaite un modèle de rang faible.

Cas d'un dirac. On montre de façon similaire que la TFCT en convention passe-bas et le spectrogramme d'un dirac sont de rang un, et que sa TFCT en convention passe-bande est de rang plein. À travers les cas de la sinusoïde et du dirac, tous deux avec une TFCT de rang 1 dans une convention différente, on remarque que le choix d'une convention pour obtenir une TFCT de rang faible dépend du contenu du signal et qu'un choix erroné met facilement en défaut l'hypothèse de rang faible, alors que le spectrogramme est insensible à ce choix.

Cas de plusieurs sinusoïdes. Une différence majeure entre TFCT et spectrogramme apparait lorsque l'on somme plusieurs sinusoïdes.

Proposition 2.3. *Soit* $P \in \mathbb{N}^*$ *et* $\mathbf{s}(m) = \sum_{p\in[\![P]\!]} \alpha_p e^{2i\pi f_p m}$ *où* $\forall p \in [\![P]\!], (f_p, \alpha_p) \in [0, 1[\times \mathbb{C}^*$ *et* $f_p \neq f_q$ *pour* $p \neq q$*. Alors*

$$\mathbf{S}_{BP} = \sum_{p\in[\![P]\!]} \alpha_p \left[\widehat{\mathbf{h}}(\nu_k - f_p)\right]_{k\in[\![K]\!]} \times \left[e^{2i\pi f_p m_n}\right]^T_{n\in\mathbb{Z}}. \quad (6)$$

Cette expression sous la forme d'une somme de matrices de rang 1 permet de voir que $\mathbf{S}_{\text{BP}}$ est en général de rang égal à P, les vecteurs colonnes dans (6) formant une famille libre. Quant au rang du spectrogramme, on peut distinguer deux cas. Le spectrogramme est approximativement de rang 1 si les fréquences $\{f_p\}_{p\in[\![P]\!]}$ sont suffisamment distantes deux à deux : les lobes principaux des vecteurs $\left[\widehat{\mathbf{h}}(\nu_k - f_p)\right]_{k\in[\![K]\!]}$ sont en effet distincts pour des valeurs de p différentes :

$$|\mathbf{S}_{\text{BP}}(k, n)|^2 = \sum_{p,q} \alpha_p \alpha_q^* \widehat{\mathbf{h}}(\nu_k - f_p)\, \widehat{\mathbf{h}}^*(\nu_k - f_q)\, e^{2i\pi(f_p - f_q)m_n}$$
$$\approx \sum_{p\in[\![P]\!]} \left|\alpha_p \widehat{\mathbf{h}}(\nu_k - f_p)\right|^2$$

Cette approximation n'est pas possible si certaines composantes ont des fréquences proches à cause de la création de battements, impliquant un spectrogramme qui n'est pas de rang faible.

La figure 2 illustre ces phénomènes, en considérant la somme de trois sinusoïdes. En convention passe-bas, toutes les valeurs singulières sont élevées : la matrice $\mathbf{S}_{\text{LP}}$ n'est pas bien approximée par une matrice de rang faible. En convention passe-bande, on observe un important saut spectral – d'un facteur 10^9 – entre les trois premières valeurs singulières de $\mathbf{S}_{\text{BP}}$ et les suivantes :

la matrice peut être considérée de rang 3, aux erreurs numériques près. Dans le cas de composantes distantes, le spectrogramme est bien approximé par une matrice de rang 1, avec un saut spectral modéré – d'un facteur 10^4 – entre les deux premières valeurs singulières. En comparant avec les valeurs singulières de $\mathbf{S}_{\text{BP}}$, le spectrogramme n'est qu'approximativement de rang faible puisque les valeurs singulières d'ordre supérieur ont des valeurs non négligeables. On voit apparaître ici les limites de la modélisation de rang faible du spectrogramme, en raison de la non-linéarité du passage au module. Dans le cas de fréquences proches, le spectrogramme n'est pas bien approximé par une matrice de rang faible : les trois premières valeurs singulières sont du même ordre que celles de $\mathbf{S}_{\text{BP}}$ et ne sont pas suivies du moindre saut spectral. La factorisation de $\mathbf{S}_{\text{BP}}$ semble ainsi plus adaptée que la factorisation du spectrogramme pour la modélisation de composantes ayant des fréquences proches.

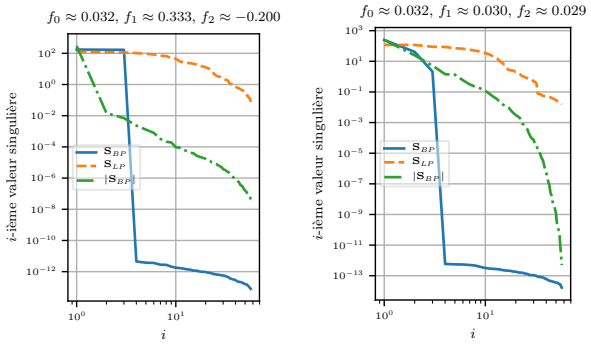

FIGURE 2 – Valeurs singulières de la représentation TFCT et du spectrogramme dans le cas d'une somme de 3 sinusoïdes de fréquences distantes (gauche) ou proches (droite), pour $h = 16$, $K = 512$, un signal de longueur 1024 et une fenêtre $\mathbf{h}$ de Hann de taille 128.

3 TFCT de rang faible

Nous étudions ici expérimentalement la qualité des modèles de rang faible de matrices temps-fréquence d'un son, en termes d'approximation puis de reconstruction de données manquantes.

3.1 Approximation de rang faible

Problème d'optimisation. Le problème (classique) considéré ici consiste à trouver la meilleure approximation $\widetilde{\mathbf{X}} \in \mathbb{C}^{F\times T}$ de rang K d'une matrice $\mathbf{X} \in \mathbb{C}^{F\times T}$, c'est-à-dire à résoudre :

$$\widetilde{\mathbf{X}} = \underset{\mathbf{Y}\in\mathbb{C}^{F\times T}, \text{rg}(\mathbf{Y})\leq K}{\arg\min} \|\mathbf{X}-\mathbf{Y}\|_F^2 \qquad (7)$$

D'après le théorème d'Eckart-Young [3], la solution est donnée par $\widetilde{\mathbf{X}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^*$ où $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ est la décomposition en valeurs singulières de $\mathbf{X}$ et $\widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^*$ est sa troncature associée aux K plus grandes valeurs singulières de $\mathbf{X}$. Dans la suite, ce problème est résolu pour $\mathbf{X} = \mathbf{S}_{\text{BP}}$ et $\mathbf{X} = |\mathbf{S}_{\text{BP}}|$ pour différentes valeurs de K.

Évaluation de l'approximation en faible rang. La qualité de l'approximation est évaluée via la valeur minimale de la fonctionnelle (7), appelée erreur d'approximation. Nous proposons aussi de mesurer le rapport signal à distortion (SDR) défini par $10\log\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x}-\widetilde{\mathbf{x}}\|_2^2}$ avec $\mathbf{x}$ le signal original et $\widetilde{\mathbf{x}}$ le signal reconstruit : dans le cas $\mathbf{X} = \mathbf{S}_{\text{BP}}$, on a $\widetilde{\mathbf{x}} = \text{TFCT}^{-1}(\widetilde{\mathbf{X}})$; dans le cas $\mathbf{X} = |\mathbf{S}_{\text{BP}}|$, les phases originales sont utilisées pour la reconstruction en posant $\widetilde{\mathbf{x}} = \text{TFCT}^{-1}(\widetilde{\mathbf{X}}e^{i\angle\mathbf{X}})$.

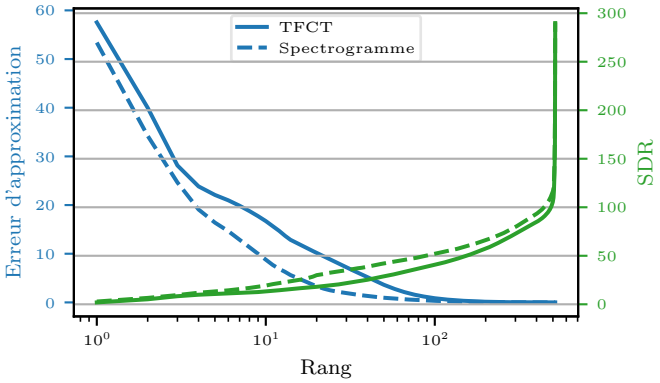

FIGURE 3 – Son de *Glockenspiel* : erreur d'approximation (bleu) et SDR (vert) en fonction du rang de l'approximation de la TFCT (trait continu) et du spectrogramme (trait pointillé).

La figure 3 illustre les résultats sur le son de *Glockenspiel* à différentes valeurs de rang. Pour la TFCT comme pour le spectrogramme, la qualité d'approximation s'améliore rapidement, en particulier pour $K \leq 4$, puis jusque $K = 100$. Les performances semblent meilleures pour le spectrogramme, mettant en évidence la difficulté d'approximer une TFCT par une matrice de rang faible. Les représentations temps-fréquences pour $K = 11$ données sur la figure 4 montrent que le résultat reste très approximatif, comparativement à l'original sur la figure 1.

3.2 Reconstruction de données manquantes

La reconstruction de données manquantes permet d'évaluer la capacité d'une approximation de rang faible à effectuer une tâche de prédiction.

L'extension du problème (7) au cas avec des données manquantes se formalise en considérant un masque binaire $\mathbf{M}$ de la taille de $\mathbf{X}$, qui met les valeurs manquantes à zéro. Le problème de trouver une approximation de faible rang pour $\mathbf{X}$ devient :

$$\widetilde{\mathbf{X}} = \underset{\mathbf{Y}\in\mathbb{C}^{F\times T}, \text{rg}(\mathbf{Y})\leq K}{\arg\min} \|\mathbf{M}\odot(\mathbf{X}-\mathbf{Y})\|_F^2. \qquad (8)$$

Ce problème est difficile du fait du masque $\mathbf{M}$ et peut être relaxé en faisant appel à la norme nucléaire :

$$\min_{\mathbf{Y}\in\mathbb{C}^{F\times T}} \frac{1}{2}\|\mathbf{M}\odot(\mathbf{X}-\mathbf{Y})\|_F^2 + \lambda\|\mathbf{Y}\|_*, \text{ où } \lambda > 0. \qquad (9)$$

Le problème (9) peut se résoudre en utilisant un algorithme de type Implicite-Explicite [2], étant donné que le premier terme est quadratique et donc différentiable et de gradient 1-Lipschitz, et que le deuxième terme est « proximable ». L'opérateur proximal associé à la norme nucléaire consiste en un seuillage des valeurs singulières [1], c'est-à-dire :

$$\text{prox}_{\lambda\|\,.\,\|_*}(\mathbf{Y}) = \mathbf{U}\max(\mathbf{\Sigma} - \lambda\mathbf{I}, \mathbf{0})\mathbf{V}^*, \text{ où } \mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*. \tag{10}$$

permattant d'obtenir l'algorithme 1.

Algorithm 1 Algorithme Implicite-Explicite pour résoudre (9).

1: Choisir $\gamma < 2$ et initialiser l'algorithme avec $\mathbf{Y}^{(0)}$.
2: **for** $\ell = 0, 1, ..., L$ **do** {ℓ-ème itération}
3: $\quad \mathbf{Y}^{(\ell+1)} = \text{prox}_{\gamma\lambda\|\,.\,\|_*}(\mathbf{Y}^{(\ell)} - \gamma\mathbf{M} \odot (\mathbf{X} - \mathbf{Y}^{(\ell)}))$
4: **end for**

Les résultats sont présentés sur la figure 5 pour des données manquantes réparties aléatoirement de manière uniforme. L'erreur d'approximation calculée sur les données observées décroit bien avec le rang. L'erreur de prédiction calculée sur les données manquantes reconstruites présente un minimum autour de $K = 11$. Ceci montre que la contrainte de rang faible a une capacité de régularisation, y compris pour la TFCT, et la valeur optimale du rang obtenue est un signe de la pertinence de l'utilisation de cette contrainte.

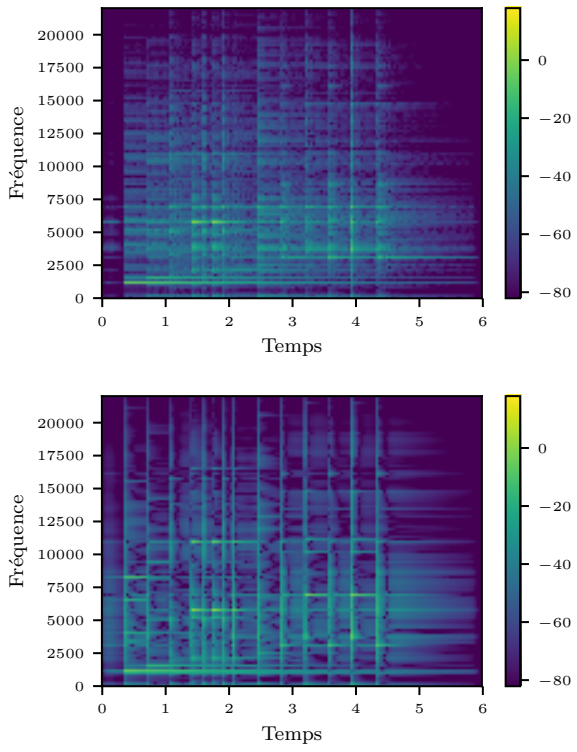

FIGURE 4 – Visualisation des approximations de rang 11 pour la TFCT (haut) et le spectrogramme (bas).

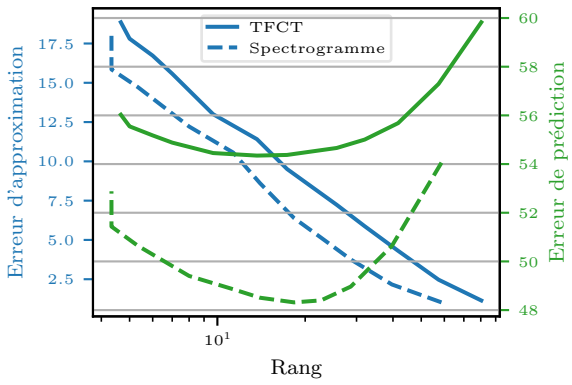

FIGURE 5 – Son de *Glockenspiel* avec 75% de données manquantes : erreurs d'approximation et de prédiction en fonction du rang de la solution (calculé pour chaque valeur de λ testée).

4 Conclusion

Les considérations théoriques et expérimentales présentées montrent des résultats ambivalents quant à l'approximation de rang faible des TFCT. D'une part, plusieurs limites ont été mises en évidence, notamment sur l'importance de la convention définissant la TFCT. Celle-ci dépend ainsi du contenu sonore , et peut affecter la capacité de modéliser des sons de natures variées. Il ressort également qu'un spectrogramme de faible rang n'implique pas une TFCT de rang aussi faible. D'autre part, plusieurs aspects positifs peuvent être relevés : d'une manière générale, une contrainte de rang faible permet de modéliser la TFCT d'un signal constitué de composantes sinusoïdales ; elle permet également de dépasser les capacités de modélisation du spectrogramme lorsque les composantes ont des fréquences proches. Ces travaux offrent donc de premiers résultats encourageants, appelant à approfondir la caractérisation et la compréhension des approximations de rang faible de la TFCT.

Références

[1] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20(4) :1956–1982, March 2010.

[2] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4) :1168–1200, Nov. 2005.

[3] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3) :211–218, September 1936.

[4] B. Mishra, G. Meyer, F. Bach, and R. Sepulchre. Low-rank optimization with trace norm penalty. *SIAM J. Optim.*, 23(4) :2124–2149, 2013.

[5] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. WASPAA*, pages 177–180, New Paltz, NY, USA, October 2003.