Séparation de sources harmoniques/percussives utilisant des estimateurs locaux de modulation linéaire AM-FM

Dominique FOURER

Laboratoire IBISC (EA 4526)
Université d'Évry / Paris-Saclay, 40 rue du Pelvoux, 91000 Evry-Courcouronnes, France
dominique@fourer.fr

Résumé – Ce travail présente une nouvelle méthode de séparation de sources harmoniques/percussives (HPSS) basée sur des estimateurs locaux de modulation d'amplitude et de fréquence dédiés initialement au *synchrosqueezing*. Nous montrons ici que nos estimateurs permettent de discriminer entre les signaux harmoniques et percussifs contenus dans un signal de mélange musical. Notre méthode fonctionne dans le domaine temps-fréquence et affecte chaque point à une source en fonction de son taux de modulation supposé plus important pour les signaux percussifs que pour les signaux harmoniques. Notre approche fournit un cadre mathématique simple et élégant au problème HPSS qui permet d'obtenir des résultats compétitifs en comparaison avec des techniques de l'état de l'art appliquées sur un jeu de données de musique.

Abstract – This paper introduces a new harmonic/percussive source separation (HPSS) method based on recently introduced local AM-FM estimators which were originally designed for synchrosqueezing techniques. Here, we show that these estimators can also be used to discriminate the harmonic part from the percussive part of a musical audio mixture. Our method blindly operates in the time-frequency plane and assigns each point to a source according to its local modulation rate that is expected to be higher for percussive sounds than for harmonic components. Our technique offers a simple and elegant mathematical formulation of the HPSS problem that can provide competitive results in comparison to state-of-the-art methods when comparatively evaluated on a music dataset.

1 Introduction

La séparation de sources est un problème compliqué qui vise à retrouver les sources isolées d'un signal (*i.e.* les instruments isolés) qui composent un mélange observé [1]. Certaines applications d'extraction d'informations musicales préfèrent traiter le problème de Séparation de Sources Harmoniques/Percussives (HPSS) plutôt que d'estimer toutes les sources du mélange. En effet, la partie harmonique permet d'estimer des informations liées aux hauteurs musicales (*i.e.* tonalité, notes, accords) tandis que la partie percussive contient des informations liées au rythme utiles à la transcription [2]. Cet article traite le cas monaural instantané qui correspond au cas où un seul signal de mélange x est disponible et peut être modélisé par :

$$x(t) = s_h(t) + s_p(t) \tag{1}$$

avec s_h et s_p respectivement les parties parties harmoniques et percussives à estimer. Les méthodes HPSS existantes [3, 4, 5, 6] exploitent l'anisotropie d'une Représentation Temps-Fréquence (RTF) qui caractérisent respectivement les signaux harmoniques et percussifs. Cette intuition est validée par des approches plus récentes telles que les réseaux de neurones profonds pour lesquels les premières couches entraînées mettent en évidence les lignes horizontales ou verticales présentes dans un spectrogramme qui caractérisent les sources harmoniques et percussives [7]. Cependant, la plupart des approches HPSS existantes ne tiennent pas compte de la signification physique des paramètres des signaux analysés telles que l'amplitude et la fréquence instantanée ainsi que leur taux de modulation linéaire. Cela conduit parfois à des erreurs d'estimation et de distinction correcte entre les différentes sources. Ainsi, notre étude propose une nouvelle méthode HPSS reposant sur des estimateurs du taux de modulation linéaire de fréquence et d'amplitude (paramètres AM/FM) utilisés initialement dans le cadre du synchrosqueezing développé dans [8, 9]. Cet article est organisé comme suit. Dans la Section 2, nous présentons plusieurs estimateurs locaux de modulation linéaire opérant dans le plan temps-fréquence. Dans la Section 3, nous introduisons notre méthode HPSS basée sur les estimateurs définis précédemment. Enfin, nous présentons nos résultats d'évaluation comparative impliquant plusieurs méthodes HPSS de l'état de l'art dans la Section 4 avant de conclure avec la Section 5.

2 Estimateurs de modulation AM-FM

2.1 Modèle de signal et propriétés

Notre méthode suppose qu'une seule source est active (harmonique ou percussive) en chaque point du plan temps-fréquence et peut être modélisée par une exponentielle complexe de la forme [9] :

$$x(t) = \mathbf{e}^{\lambda_x(t)+j\phi_x(t)}, \quad \text{avec } j^2{=}{-}1, \tag{2}$$

$\lambda_x(t) = l_x + \mu_x t + \nu_x \frac{t^2}{2}$, la log-amplitude et $\phi_x(t) = \varphi_x + \omega_x t + \alpha_x \frac{t^2}{2}$, la phase en fonction du temps t. Ce signal vérifie :

$$\frac{dx}{dt}(t) = \left(\frac{d\lambda_x}{dt}(t) + j\frac{d\phi_x}{dt}(t)\right) x(t) = (q_x t + p_x)x(t) \quad (3)$$

avec $q_x = \nu_x + j\alpha_x$ et $p_x = \mu_x + j\omega_x$. Sa Transformée de Fourier à Court-Terme (TFCT) utilisant une fenêtre d'analyse h est donnée par :

$$F_x^h(t,\omega) = \int_{\mathbb{R}} x(u)h(t-u)^* \, \mathbf{e}^{-j\omega u} \, \mathrm{d}u \quad (4)$$

$$= \mathbf{e}^{-j\omega t} \int_{\mathbb{R}} x(t-u)h(u)^* \, \mathbf{e}^{j\omega u} \mathrm{d}u \quad (5)$$

avec z^* le conjugué de $z \in \mathcal{C}$. Dériver l'Eq. (5) par rapport à t conduit à :

$$\frac{\partial F_x^h}{\partial t}(t,\omega) = \int_{\mathbb{R}} x(u)\frac{\mathrm{d}h}{\mathrm{d}t}(t-u)^* \, \mathbf{e}^{-j\omega u} \, \mathrm{d}u \quad (6)$$

$$= -j\omega F_x^h(t,\omega) + \mathbf{e}^{-j\omega t} \int_{\mathbb{R}} \frac{\mathrm{d}x}{\mathrm{d}t}(t-u)h(u)^* \, \mathbf{e}^{j\omega u} \mathrm{d}u. \quad (7)$$

En substituant $\frac{\mathrm{d}x}{\mathrm{d}t}(t-u)$ par $(q_x\,(t-u) + p_x)\; x(t-u)$, on obtient :

$$F_x^{\mathcal{D}h}(t,\omega) = -q_x F_x^{\mathcal{T}h}(t,\omega) + (q_x t + p_x - j\omega)F_x^h(t,\omega), \quad (8)$$

où $F_x^{\mathcal{D}h}(t,\omega)$ et $F_x^{\mathcal{T}h}(t,\omega)$ sont deux TFCTs utilisant les fenêtres d'analyse $\mathcal{D}h(t) = \frac{\mathrm{d}h}{\mathrm{d}t}(t)$ et $\mathcal{T}h(t) = t\,h(t)$. On déduit l'expression des dérivées d'ordre n ($\forall n \geq 1$) par rapport à t [8] :

$$F_x^{\mathcal{D}^n h}{}_{(t,\omega)} = -q_x F_x^{\mathcal{T}\mathcal{D}^{n-1}h}{}_{(t,\omega)} + (q_x t + p_x - j\omega)F_x^{\mathcal{D}^{n-1}h}{}_{(t,\omega)} \quad (9)$$

et des dérivées par rapport à ω à l'ordre n ($\forall n \geq 1$) [8] :

$$F_x^{\mathcal{T}^{n-1}\mathcal{D}h}(t,\omega) + (n-1)\, F_x^{\mathcal{T}^{n-2}h}(t,\omega) = - q_x\, F_x^{\mathcal{T}^n h}(t,\omega) + (q_x t + p_x - j\omega)\, F_x^{\mathcal{T}^{n-1}h}(t,\omega) \quad (10)$$

2.2 Estimation des paramètres

En utilisant les Eqs. (8) et (9), on construit $\forall (t,\omega) \in \mathbb{R}^2$ un système linéaire dont les inconnues sont q_x et $\Psi_x = q_x t + p_x$:

$$\begin{pmatrix} F_x^{\mathcal{D}^{n-1}h} & -F_x^{\mathcal{T}\mathcal{D}^{n-1}h} \\ F_x^h & -F_x^{\mathcal{T}h} \end{pmatrix} \begin{pmatrix} \Psi_x \\ q_x \end{pmatrix} = \begin{pmatrix} F_x^{\mathcal{D}^n h} + j\omega F_x^{\mathcal{D}^{n-1}h} \\ F_x^{\mathcal{D}h} + j\omega F_x^h \end{pmatrix} \quad (11)$$

Quand (11) est inversible (*i.e.* $|F_x^h(t,\omega)|^2 > 0$), on obtient les estimateurs suivants notés (tn) :

$$\hat{q}_x^{(tn)}(t,\omega) = \frac{F_x^{\mathcal{D}^n h}F_x^h - F_x^{\mathcal{D}^{n-1}h}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}h}F_x^{\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}\mathcal{D}^{n-1}h}F_x^h} \quad (12)$$

$$\hat{\Psi}_x^{(tn)}(t,\omega) = \frac{F_x^{\mathcal{D}h}F_x^{\mathcal{T}\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}h}F_x^{\mathcal{D}^n h}}{F_x^{\mathcal{T}\mathcal{D}^{n-1}h}F_{x\,(t,\omega)}^h - F_x^{\mathcal{T}h}F_x^{\mathcal{D}^{n-1}h}} + j\omega \quad (13)$$

D'après l'Eq. (2), on exprime la modulation de la log-amplitude $\dot{\lambda}_x(t) = \frac{d\lambda_x}{dt}(t) = \mu_x + \nu_x t$ et la fréquence instantanées $\dot{\phi}_x(t) = \frac{d\phi_x}{dt}(t) = \omega_x + \alpha_x t$ qui peuvent être retrouvés sachant que $\Psi_x(t) = \dot{\lambda}_x(t) + j\dot{\phi}_x(t) = q_x t + p_x$ en utilisant $\hat{q}_x^{(tn)}$ ou n'importe quel estimateur $\hat{q}_x$ [8]. Ainsi, nous obtenons les estimateurs des paramètres du modèle de l'Eq. (2) :

$$\hat{\nu}_x(t,\omega) = \mathrm{Re}\Big(\hat{q}_x(t,\omega)\Big), \qquad \hat{\alpha}_x(t,\omega) = \mathrm{Im}\Big(\hat{q}_x(t,\omega)\Big) \quad (14)$$

$$\hat{\dot{\lambda}}_x(t,\omega) = \mathrm{Re}\left(\hat{\Psi}_x(t,\omega)\right), \quad \hat{\dot{\phi}}_x(t,\omega) = \mathrm{Im}\left(\hat{\Psi}_x(t,\omega)\right) \quad (15)$$

Un nombre infini d'estimateurs $\hat{q}_x(t,\omega)$, peut être obtenu par les Eqs. (8) et (9). Ces derniers reposant sur les dérivées d'ordre n par rapport au temps sont notés (tn). De même, d'autres estimateurs utilisant les dérivées d'ordre n par rapport à ω notés (ωn), sont obtenus par l'Eq. (10) pour $n = 1$ et pour $n \geq 2$, en utilisant $\hat{q}_x^{(\omega n)}(t,\omega) =$

$$\frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + {}_{(n-1)}F_x^{\mathcal{T}^{n-2}h})F_x^h - F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h}F_x^h}. \quad (16)$$

3 Séparation de sources

À présent, on considère les versions à temps discret des expressions précédentes obtenues par approximation rectangulaire $F_x^h[k,m] \approx F_x^h(\frac{k}{F_s}, 2\pi\frac{mF_s}{M})$ avec les indices de temps $k \in \mathbb{Z}$ et de fréquence $m \in [-M/2+1; M/2]$, F_s et M étant respectivement la fréquence d'échantillonnage et le nombre d'indices de fréquences calculés.

3.1 Analyse discriminante de la modulation

La séparation est effectuée par discrimination du plan temps-fréquence en utilisant une hypothèse d'orthogonalité entre les sources comme dans [10]. Pour cela, nous utilisons les estimations de modulation locale du mélange x décrit par les paramètres AM $\hat{\dot{\lambda}}_x[k,m]$, FM $\hat{\alpha}_x[k,m]$ et leur combinaison AM-FM $C_x[k,m] = \sqrt{\hat{\dot{\lambda}}_x[k,m]^2 + \hat{\alpha}_x[k,m]^2}$. Ces descripteurs notés $G_x[k,m] \in \{|\hat{\dot{\lambda}}_x[k,m]|, |\hat{\dot{\phi}}_x[k,m]|, C_x[k,m]\}$ sont calculés à partir du mélange observé $x[k]$ avec les Eqs. (15) et (14) pour lesquelles $\hat{q}_x[k,m]$ est calculé en utilisant au choix l'un des estimateurs décrit en Section 2 (*i.e.* (tn) ou (ωn)). Pour assigner indépendamment chaque point temps-fréquence $[k,m]$ à une source, on considère un voisinage pondéré par l'énergie du signal autour du point considéré :

$$\mathcal{Q}_{x[k,m]} = \left\{ \frac{G_{x[k',m']}|F_{x[k',m']}|^2}{\sum_{k'}\sum_{m'}|F_{x[k',m']}|^2} \;\middle|\; \begin{matrix} \forall k' \in [k-\Delta_k; k+\Delta_k] \\ \forall m' \in [m-\Delta_m; m+\Delta_m] \end{matrix} \right\} \quad (17)$$

Afin de trouver la meilleure projection linéaires des coefficients $\mathcal{Q}_x[k,m]$ permettant de maximiser la distance entre les classes harmoniques et percussives, nous utilisons une Analyse Linéaire Discriminante (ALD) [11]. Pour cela, nous calculons une matrice de covariance inter-classe B puis une matrice de covariance intra-classe W à partir des coefficients $\mathcal{Q}_x[k,m]$

estimés sur des données d'entraînement (sources séparées disponibles). Par la suite, l'ALD fournit une solution en utilisant les vecteurs propres de la matrice $D = (B+W)^{-1}B$, obtenue par pseudo-inverse.

3.2 HPSS

3.2.1 Apprentissage

Notre méthode reposant sur l'ALD nécessite des données d'entraînement pour obtenir la meilleure combinaison linéaire des descripteurs utilisés pour le calcul du masque temps-fréquence de séparation des sources harmoniques/percussives. L'apprentissage de la fonction discriminante nécessite de connaître le vrai masque de séparation noté $M_h^{(true)}[k,m]$ et $M_p^{(true)}[k,m]$ qui doit satisfaire notre hypothèse d'orthogonalité. Ce masque est déduit des spectrogrammes de chaque source isolée (disponible uniquement pour les données d'entraînement) :

$$M_h^{(true)}[k,m] = \begin{cases} 1 & \text{si } |F_{s_h}^h[k,m]|^2 > |F_{s_p}^h[k,m]|^2 \\ 0 & \text{sinon} \end{cases}, \quad (18)$$

et $M_p^{(true)}[k,m] = 1 - M_h^{(true)}[k,m]$. Ainsi, la TFCT du mélange x est utilisée pour calculer les vecteurs propres de l'ALD des coefficients $\mathcal{Q}_x[k,m]$. Ainsi, un nombre arbitraire de mélanges audios peut être utilisé pour l'entraînement permettant d'apprendre les centroïdes des sources (*i.e.* μ_h ou μ_p) à partir des projections dans l'espace discriminant des coefficients estimés à partir du mélange obtenu.

3.2.2 Séparation

Après l'entraînement du modèle d'ALD décrit plus haut composé des vecteurs propres de l'espace discriminant et des centroïdes de chaque classe, la décision d'affecter un point temps-fréquence à une source est réalisé en projetant ces coefficients $\mathcal{Q}_x[k,m]$ dans l'espace discriminant puis en affectant le point à la source dont la centroïde apprise est la plus proche en terme de distance euclidienne. La procédure peut être résumée comme suit pour un signal d'entrée x :

- Calcul de la TFCT $F_x^h[k,m]$ par l'Eq. (5).
- Pour chaque point temps-fréquence, calcul de $\mathcal{Q}_x[k,m]$ en utilisant Eq. (17).
- Calcul de la projection linéaire $P_{\mathcal{Q}}$ avec la base de vecteurs propres obtenue par ALD.
- Calcul des masques de séparation :

$$M_h[k,m] = \begin{cases} 1 & \text{si } ||P_{\mathcal{Q}}[k,m] - \mu_h|| < ||P_{\mathcal{Q}}[k,m] - \mu_p|| \\ 0 & \text{sinon} \end{cases} \quad (19)$$

  $M_p[k,m] = 1 - M_h[k,m]$.
- Reconstruction de chaque source par inversion de la TFCT :

$$\hat{s}_h = \text{TFCT}^{-1}(F_x^h[k,m]M_h[k,m]) \quad (20)$$

$$\hat{s}_p = \text{TFCT}^{-1}(F_x^h[k,m]M_p[k,m]) \quad (21)$$

avec $||.||$ la norme l_2 et TFCT^{-1} la formule d'inversion de la TFCT permettant de reconstruire un signal temporel.

4 Simulations numériques

4.1 Données

Nous avons utilisé le corpus proposé par E. Cano en 2010 [1] contenant 10 enregistrements professionnels de musique populaire ayant chacun une durée de 25 secondes et dont les parties harmoniques et percussives sont disponibles dans des pistes isolées permettant le calcul des masques utilisé pour l'entraînement de la méthode décrite dans la Section 3.2.1. Chaque extrait est rééchantillonné à $F_s = 22,05$ kHz et le mélange x est créé d'après l'Eq. (1). Nos calculs de TFCT utilisent une fenêtre de Hann d'une durée de 92,9 ms avec un chevauchement de 50% entre trames successives. Chaque point temps-fréquence est caractérisé par $\mathcal{Q}_{x[k,m]}$ calculé avec $\Delta_k = \Delta_m = 1$ pour un total de 9 valeurs par point. L'entraînement de la méthode a été réalisé à partir des 300 000 premiers points temps-fréquence du premier extrait musical. Le même modèle est ensuite utilisé sur le jeu de données complet afin d'obtenir les résultats présentés ci-après.

4.2 Résultats comparatifs

Nous comparons les résultats de séparation de notre méthode avec 2 techniques de l'état de l'art nommées FMF [13] et *JL14* [4] utilisant les réglages recommandés par leurs auteurs. Nous évaluons nos méthodes basées sur les estimateurs $(t2)$ et $(\omega 2)$ utilisés pour calculer les paramètres AM, FM et AM-FM décrits en Section 3.2. Les estimateurs d'ordre supérieurs ($n > 2$) n'ont pas été présentés car étant plus sensibles au bruit, ils fournissent de moins bons résultats que les estimateurs d'ordre 2, confirmant les observations décrites dans [9]. Les résultats pour toute la base de données traitée, exprimés en termes de $\text{RQF}(s,\hat{s}) = 20\log_{10}(\frac{||s||}{||s-\hat{s}||})$, et de SIR (interférences), SAR (artefacts) et SDR (distorsion) fournis par BSSEval [12] sont présentés dans la Fig. 1. Ils montrent que nos méthodes fournissent des résultats compétitifs comme le confirment nos tests d'écoute informels [2]. Les meilleurs résultats en termes de RQF, SIR, SAR (sauf pour la source percussive), et SDR (sauf la source harmonique) sont fournis par notre méthode en comparaison avec les techniques FMF et JL14. Notre approche semble présenter un avantage pour estimer correctement les sources harmoniques, surtout quand les paramètres FM et AM-FM sont utilisés, quelque soit l'estimateur ($(t2)$ ou $(\omega 2)$). Les résultats montrent que le paramètre AM seul n'est pas suffisant pour obtenir une bonne séparation bien qu'il permette dans certain cas une amélioration de l'estimation de la source percussive en termes de SIR et de SAR. Ainsi, une combinaison adéquate de chaque paramètre et des bons estimateurs peut conduire dans tous les cas à une estimation correcte de la source d'intérêt.

1. https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/phase_based_harmonic_percussive_separation.html
2. Résultats en ligne http://fourer.fr/publi/gretsi22/

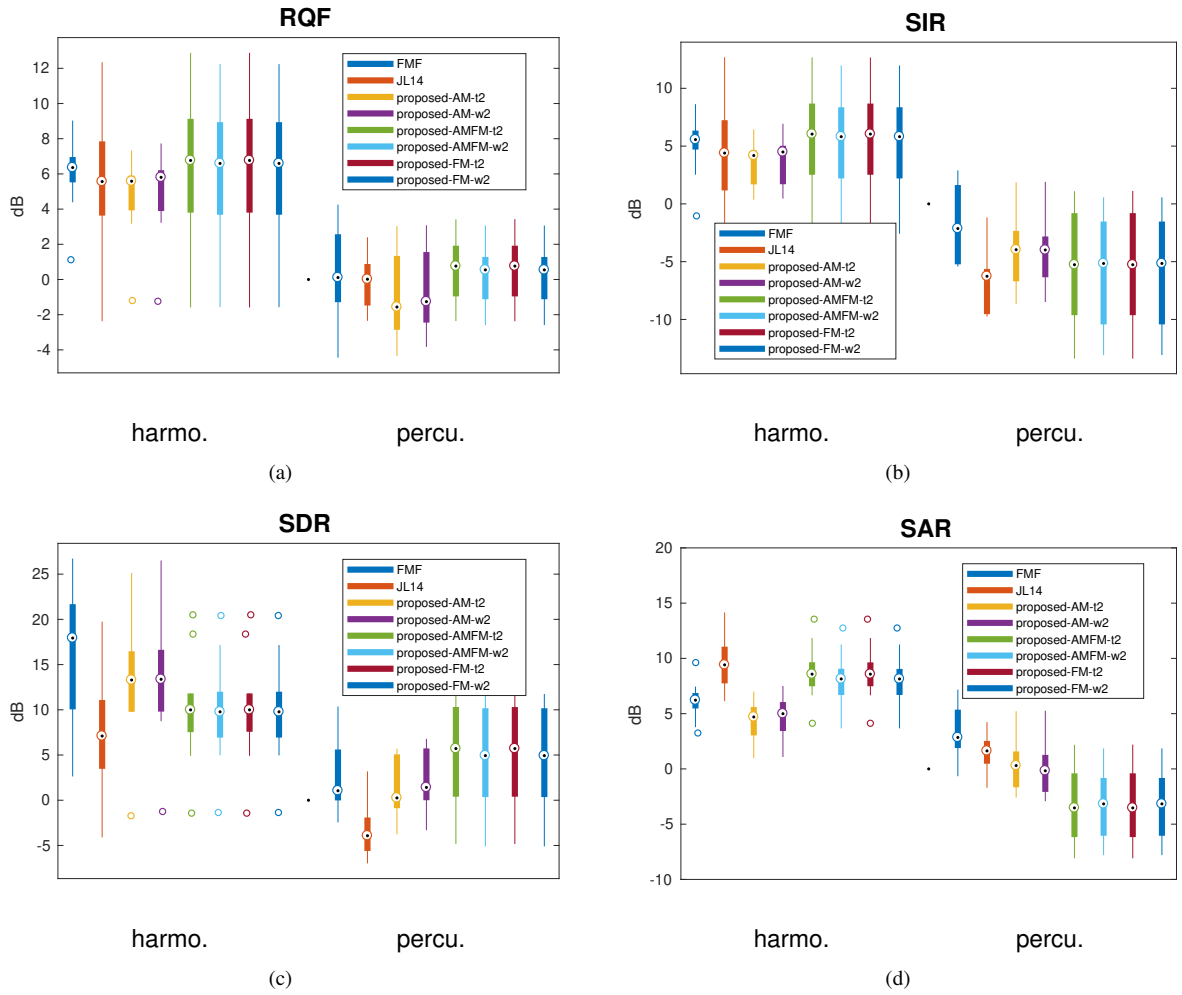

FIGURE 1 – Résultats comparatifs exprimés en termes de RQF et de scores calculés par BssEval[12].

5 Conclusion

Une nouvelle technique HPSS [3] basée sur des estimateurs locaux AM-FM a été présentée. Elle montre que l'analyse spectrale malgré sa sensibilité au bruit demeure une approche efficace pour le traitement audio. Cela est confirmé par des résultats compétitifs avec des techniques de l'état de l'art reposant sur des techniques de traitement d'images réputées plus robustes. Nous pouvons aussi noter que l'hypothèse d'orthogonalité entre les sources est valide pour fournir des résultats acceptables à partir d'un mélange monophonique. Par la suite, nous envisageons d'étendre notre modèle de signal pour considérer le bruit et améliorer l'estimation de la source percussive.

Références

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation : Independent component analysis and applications*. Academic press, 2010.

[2] K. O'Hanlon and M. B. Sandler, "Improved detection of semi-percussive onsets in audio using temporal reassignment," in *Proc. IEEE ICASSP*, 2018, pp. 611–615.

[3] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. Digital Audio Effects Conference (DAFx-10)*, Graz, Austria, Sep. 2010, pp. 10–13.

3. Ce travail est soutenu par le projet ANR ASCETE (ANR-19-CE48-0001)

[4] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," vol. 21, no. 10, pp. 1197–1200, 2014.

[5] E. Cano, M. Plumbley, and C. Dittmar, "Phase-based harmonic/percussive separation," pp. 1628–1632, Sep. 2014.

[6] R. Füg, A. Niedermeier, J. Driedger, S. Disch, and M. Müller, "Harmonic-percussive-residual sound separation using the structure tensor on spectrograms," in *Proc. IEEE ICASSP*, 2016, pp. 445–449.

[7] W. Lim and T. Lee, "Harmonic and percussive source separation using a convolutional auto encoder," in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, Aug. 2017.

[8] D. Fourer, F. Auger, K. Czarnecki, S. Meignen, and P. Flandrin, "Chirp rate and instantaneous frequency estimation : Application to recursive vertical synchrosqueezing," *IEEE Signal Processing Letters*, vol. 24, no. Issue 11, pp. 1724–1728, Nov. 2017.

[9] D. Fourer, F. Auger, and G. Peeters, "Local am/fm parameters estimation : application to sinusoidal modeling and blind audio source separation," *IEEE Signal Processing Letters*, vol. 25, pp. 1600–1604, Oct. 2018.

[10] D. Fourer and G. Peeters, "Fast and adaptive blind audio source separation using recursive levenberg-marquardt synchrosqueezing," in *Proc. IEEE ICASSP*, Calgary, Canada, Apr. 2018.

[11] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, USA : Wiley-Blackwell, 1958.

[12] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[13] D. FitzGerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," in *25th IET Irish Signals Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014)*, Jun. 2014, pp. 35–40.