Estimation de paramètres de resynthèse de sons d'instruments de musique avec des outils de morphologie mathématique

Gonzalo Romero-García, Carlos Agón, Isabelle Bloch

ESTIMATION DE PARAMÈTRES DE RESYNTHÈSE DE SONS D'INSTRUMENTS DE MUSIQUE AVEC DES OUTILS DE MORPHOLOGIE MATHÉMATIQUE

Gonzalo Romero-García, Carlos Agón
STMS, IRCAM, Sorbonne Université,
CNRS, Ministère de la Culture, 75004 Paris
{romero, agonc}@ircam.fr

Isabelle Bloch
Sorbonne Université, CNRS, LIP6, Paris
isabelle.bloch@sorbonne-universite.fr

RÉSUMÉ

Les sons représentés par des spectrogrammes peuvent être considérés comme des images dont la dimension horizontale correspond au temps, et la verticale à la fréquence. Dans cet article, nous proposons d'explorer quelques outils algébriques de la morphologie mathématique, théorie largement développée en analyse et interprétation d'images, pour estimer des paramètres de resynthèse d'un son d'instrument de musique en le modélisant comme une partie harmonique à laquelle est ajouté un bruit blanc filtré. En particulier, nous montrons que des transformations non linéaires visant à détecter des structures saillantes permettent de déduire les paramètres des composantes harmoniques d'un signal. Avec les mêmes outils, nous estimons les paramètres de filtrage pour resynthétiser la partie non-harmonique.

1. INTRODUCTION

La synthèse de sons d'instruments de musique est un problème qui a surgi aux débuts du traitement numérique du signal audio. Les reproducteurs de fichiers MIDI utilisaient des modèles très simples formés par l'addition de sinusoïdes harmoniques. Les amplitudes de ces sinusoïdes étaient estimées au moyen de la transformée de Fourier, mais ne pouvaient pas rendre compte d'une évolution temporelle. Des modèles plus complexes ont ensuite été développés, permettant l'évolution temporelle du son, et reposant le plus souvent sur la transformée de Fourier à court terme, plus connue par son sigle en anglais STFT [1]. C'est le cas, par exemple, dans [7], [1] ou encore [17].

Dans cet article, nous proposons de resynthétiser un signal d'un son d'instrument de musique à partir d'un échantillon de celui-ci. Pour ce faire, nous utilisons le modèle proposé dans [18] en modélisant le signal comme une partie générée par synthèse additive (la partie harmonique) à laquelle s'ajoute une partie générée de manière stochastique (la partie non-harmonique). Dans notre cas, la partie stochastique s'obtient par filtrage d'un bruit blanc. Ce modèle est présenté dans la section 2.

1 . *Short-Time Fourier Transform.*

Il s'agit alors d'estimer des paramètres dans notre échantillon qui nous permettent de synthétiser par ce modèle un son similaire. Pour retrouver ces paramètres, nous utilisons une représentation du signal intermédiaire sous forme de spectrogramme de STFT. Cette représentation, présentée en section 3, peut être vue comme une image du son en échelle de gris et fait apparaître les différents partiels qui composent la partie harmonique, ainsi que les distributions fréquentielles de la partie non-harmonique.

La nouveauté proposée dans cet article est l'estimation des amplitudes des partiels et des filtres à l'aide d'outils de morphologie mathématique. Cette théorie, développée dans la deuxième partie du XX$^{\text{ème}}$ siècle [12, 16], est très utilisée pour traiter, analyser et interpréter des images. Dans la section 4, nous présentons quelques outils que nous utilisons pour trouver les paramètres de synthèse.

En particulier, nous exploitons la structure des spectrogrammes où la partie harmonique apparaît sous forme de lignes horizontales ; grâce à des opérations morphologiques, nous détectons les temps, fréquences et amplitudes associés à chaque partiel que nous utilisons ensuite pour la synthèse additive. D'autres opérations morphologiques nous permettent de filtrer ces parties saillantes pour pouvoir utiliser le résultat comme filtre d'un bruit blanc. Ces processus sont expliqués dans la section 5.

Pour évaluer l'efficacité de notre méthode, nous choisissons un cas favorable (un son de synthétique généré par le modèle expliqué en section 2) et nous montrons que nous pouvons estimer ses paramètres de resynthèse avec une très bonne précision (section 5). Ensuite, dans la section 6, nous appliquons le même procédé à des sons d'instruments de musique issus d'une base de données.

2. MODÈLE DE SIGNAL

L'idée de cet article est de reconstruire un signal donné en estimant des paramètres de synthèse à partir de son spectrogramme. Pour ce faire, nous devons d'abord expliquer comment nous construisons le signal et quels sont les paramètres à estimer. Dans cette section, nous exposons un modèle de synthèse composé d'une partie additive, correspondante à la partie harmonique, et d'une partie soustractive, correspondante à la partie non-harmonique.

Pour vérifier l'efficacité de notre méthode, nous allons construire un signal selon ce modèle qui nous servira de test ; nous attendons de notre méthode qu'elle puisse retrouver les paramètres de synthèse d'un tel signal.

2.1. Partie harmonique

La partie harmonique est créée par synthèse additive de sinusoïdes dont les fréquences et les amplitudes varient dans le temps : comme expliqué dans la première section de [1], nous prenons $N \in \mathbb{N}$ partiels sinusoïdaux de durée $T > 0$ dont la fréquence $\omega_n(t)$ et l'amplitude $a_n(t)$ varient en fonction du temps $t \in [0, T]$. Chaque partiel s_n est donné par la formule

$$s_n(t) = a_n(t) \sin\left(\theta_n(t)\right) , \tag{1}$$

où $\theta_n(t) = 2\pi \int_0^t \omega_n(\tau)\, d\tau$, $t \in [0, T]$, $n \in \{0, ...N-1\}$.

Dans le cas de notre signal de référence, nous prenons des fréquences constantes et multiples d'une fréquence fondamentale f_0 :

$$\forall t \in [0, T],\ \omega_n(t) = f_n := (n+1) f_0 . \tag{2}$$

La fonction de phase s'écrit alors :

$$\theta_n(t) = 2\pi \int_0^t (n+1) f_0\, d\tau = 2\pi (n+1) f_0 t . \tag{3}$$

Les amplitudes de notre signal de référence $a_n(t)$ sont des exponentielles décroissantes :

$$\forall t \in [0, T],\ a_n(t) = A_n e^{-2\pi \delta_n t} \tag{4}$$

où les A_n sont les amplitudes initiales et δ_n les facteurs de décroissance [2] . De plus, nous prenons $A_n = A_0 \frac{1}{(n+1)^2}$ comme amplitudes initiales, avec $A_0 > 0$ et des facteurs de décroissance linéaires en fréquence, *i.e.* $\delta_n = \delta f_n = \delta(n+1) f_0$, où $\delta > 0$ est un facteur qui contrôle la décroissance globale.

En combinant ces paramètres, nous obtenons un signal de référence qui s'écrit :

$$s(t) = \sum_{n=0}^{N-1} s_n(t) \tag{5}$$

$$= \sum_{n=0}^{N-1} A_n e^{-2\pi \delta_n t} \sin\left(\theta_n(t)\right) \tag{6}$$

$$= \sum_{n=0}^{N-1} \frac{A_0}{(n+1)^2} e^{-2\pi \delta (n+1) f_0 t} \sin\left(2\pi (n+1) f_0 t\right) . \tag{7}$$

2.2. Partie non-harmonique

La partie non-harmonique du signal est modélisée de manière soustractive par filtrage d'un bruit blanc. Par la suite, nous utilisons des procédés classiques tels que le filtrage, le fenêtrage ou la recomposition d'un signal par la méthode OLA [3] [6]. Pour une présentation de ces outils, nous proposons comme références [13, 14].

Pour générer notre signal de référence, nous procédons comme suit :

1. Génération d'un bruit blanc b donné par une fonction $t \mapsto b(t)$.
2. Décomposition du bruit en trames $b_n(t)$ de longueur $L > 0$ et d'espacement $H > 0$, *i.e.* :

$$b_n(t) = b(t + nH),\ t \in [0, L] . \tag{8}$$

3. Fenêtrage de ces trames par une fenêtre de Hann $w(t) = \sin^2\left(\frac{\pi n}{N}\right)$.
4. Filtrage des trames par un filtre linéaire qui varie dans le temps Θ.
5. Recomposition du signal par la méthode OLA.

Pour notre signal de référence, le filtre que nous utilisons est un filtre passe-bande dont l'amplitude décroît exponentiellement en temps. La réponse en fréquence du filtre au temps $\tau \in [0, T]$ est donnée par :

$$\Theta_\tau(\omega) = \chi_{[f_{c_1}, f_{c_2}]}(\omega) e^{-2\pi \tau \eta} , \tag{9}$$

où $\chi_{[f_{c_1}, f_{c_2}]}$ est la fonction indicatrice de l'intervalle $[f_{c_1}, f_{c_2}]$, f_{c_1} et f_{c_2} sont les fréquences de coupure et $\eta > 0$ est une constante qui détermine la rapidité de la décroissance de l'amplitude du filtre en temps.

2.3. Génération de notre signal de référence

Pour construire notre signal de référence, nous générons d'une part la partie harmonique et d'autre part la partie non-harmonique et nous les additionnons. Le bruit blanc est généré en choisissant des nombres au hasard entre -1 et 1 de manière uniforme. Les paramètres que nous utilisons pour la partie additive et pour filtrer le bruit sont exposés dans la table 1.

Partie harmonique	
N	16
f_0	220 Hz
A_0	0.1
δ	5×10^{-4}

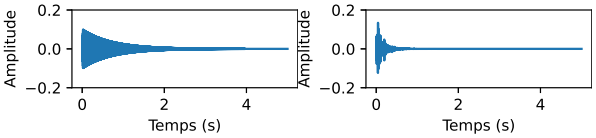
Partie non-harmonique	
L	0.1 s
H	0.01 s
f_{c_1}	100 Hz
f_{c_2}	300 Hz
η	1 Hz

Table 1. Paramètres utilisés pour la synthèse des parties harmonique et non-harmonique de notre signal de référence.

La figure 1 montre le signal de référence. Sa représentation est temporelle et ne nous donne pas beaucoup d'informations sur la composition du signal hormis le fait qu'il a une amplitude décroissante. Dans la section suivante, nous présentons la notion de spectrogramme de STFT qui nous permettra d'avoir une représentation du signal sous forme d'une image où sa composition sera plus évidente.

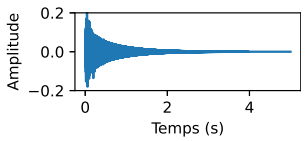
2 . On multiplie les facteurs de décroissance par 2π pour qu'ils soient exprimés en Hz.

3 . *Overlap-add method*.



(a) Partie harmonique

(b) Partie non-harmonique



(c) Somme des parties harmonique et non-harmonique

Figure 1. Signal de référence généré par addition d'une partie harmonique et une partie non-harmonique.

3. SPECTROGRAMMES

Dans cette section, nous présentons les spectrogrammes, considérés comme « images » d'un son. Ce sont sur ces représentations que nous appliquerons les outils morphologiques. Le spectrogramme que nous utilisons est celui de la STFT. Une description détaillée de ces transformations et des fondements mathématiques sur lesquels elles reposent peut être trouvée dans [8].

3.1. Cas continu

La STFT peut être défine dans le cas continu comme suit.

Définition 1 *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction bornée de classe C^∞. Soit $w : \mathbb{R} \to \mathbb{R}^+$ une fonction de classe C^∞ à support compact. La STFT de f avec fenêtre w est définie par*

$$\begin{array}{rccl} \mathrm{STFT}_w[f] : & \mathbb{R}\times\mathbb{R} & \to & \mathbb{C} \\ & (\tau,\omega) & \mapsto & \int_{\mathbb{R}} f(t)w(t-\tau)e^{-\jmath 2\pi\omega t}\,dt\,. \end{array} \tag{10}$$

La définition de la STFT peut être établie dans plusieurs contextes d'analyse fonctionnelle ; ici, nous utilisons des fonctions infiniment dérivables et une fenêtre à support compact pour garantir l'existence de l'intégrale. Néanmoins, il est fréquent de se placer dans le cadre des fonctions de Schwartz et même de la théorie des distributions. Ces considérations sont explorées dans [8].

Une fois la STFT établie, son spectrogramme se définit comme suit.

Définition 2 *Soient f et w des fonctions comme dans la définition 1. Le* spectrogramme *de STFT de f avec fenêtre w est défini par*

$$\begin{array}{rccl} \mathrm{SPEC}_w[f] : & \mathbb{R}\times\mathbb{R} & \to & \mathbb{R}^+ \\ & (\tau,\omega) & \mapsto & |\mathrm{STFT}_w[f](\tau,\omega)|^2\,. \end{array} \tag{11}$$

Ainsi, un spectrogramme représente l'information d'amplitude de la STFT au carré. On parle souvent de la valeur d'un spectrogramme de f au point (τ,ω) comme de la densité de puissance de f au temps τ et à la fréquence ω. Cette densité de puissance est souvent exprimée en dB par la formule

$$\mathrm{SPEC}^{\mathrm{dB}} = 10\log_{10}(\mathrm{SPEC})\,. \tag{12}$$

Si l'on considère les spectrogrammes en échelle logarithmique, notre ensemble d'arrivée n'est plus $\mathbb{R}^+$ mais $\mathbb{R}\cup\{-\infty\}$; de plus, si l'on considère que $|f|$ est bornée par 1 (ce qui est le cas dans les signaux audio) et que l'intégrale de w est égale à 1, un résultat établi dans [8] affirme que $\mathrm{SPEC}_w[f]$ est borné par 1 et donc que $\mathrm{SPEC}_w^{\mathrm{dB}}[f]$ est borné par 0.

Des précisions sont à faire pour le domaine de la STFT et des spectrogrammes : la première dimension, correspondant au temps, est infinie en théorie mais finie en pratique puisque les signaux d'entrée sont finis. Pour la dimension fréquentielle, même si elle est définie pour tout $\mathbb{R}$ dans la pratique, lorsqu'il s'agit de signaux réels, on ne garde que la partie positive, puisque la partie négative est symétrique hermitienne. De plus, pour un signal échantillonné à une fréquence $F_e \in \mathbb{N}$, nous ne gardons que les fréquences jusqu'à $\frac{F_e}{2}$, rendant la dimension fréquentielle finie.

Ces considérations nous mènent à établir le cadre discret utilisé lors des calculs.

3.2. Cas discret

Pour pouvoir calculer des spectrogrammes dans la pratique, l'entrée est, non pas une fonction continue et infiniment dérivable, mais plutôt un signal échantillonné. La formule de la STFT discrète est donnée dans la définition suivante.

Définition 3 *Soit $(\mathbf{f}[l])_{l=0}^{L}$ un signal de taille $L+1$ échantillonné à une fréquence F_e. Soit $(\mathbf{w}[n])_{n=0}^{N}$ une fenêtre de taille $N+1$. Soit $N_{\mathrm{FFT}} \in \mathbb{N}$, $N_{\mathrm{FFT}} \geq N$ et $H \in \mathbb{N}$. Pour tout $m = 0 : \lfloor \frac{L}{H} \rfloor$, $k = 0 : \lfloor \frac{N_{\mathrm{FFT}}}{2} \rfloor$, on définit la STFT de $\mathbf{f}$ avec fenêtre $\mathbf{w}$ par*

$$\mathrm{STFT}[\mathbf{f}][m,k] = \sum_{n=0}^{N} \mathbf{f}\left[mH + n - \lfloor \frac{N}{2} \rfloor\right] \mathbf{w}[n] e^{-\jmath 2\pi n \frac{k}{N_{\mathrm{FFT}}}} \tag{13}$$

La coordonnée m de la STFT correspond au temps $\frac{mH}{F_e}$ et la coordonnée k correspond à la fréquence $\omega_k = F_e \frac{k}{N_{\mathrm{FFT}}}$. On complète $\mathbf{f}$ par des zéros aux bords pour que la formule soit bien définie.

La table 2 donne les paramètres choisis pour le calcul des spectrogrammes. La fenêtre choisie est la fenêtre de Hann et est normalisée pour que sa somme fasse 1.

Les figures 2 et 3 montrent les spectrogrammes des parties harmonique et non-harmonique de notre signal de référence commençant après 1 s de silence et d'une durée $T = 4$ s, synthétisées avec les paramètres donnés dans la table 1.

F_e	44 100 Hz
N	4410
N_{FFT}	8820
H	441
$\mathbf{w}[n]$	$\sin^2\left(\frac{\pi n}{N}\right)$

Table 2. Paramètres utilisés pour la STFT.

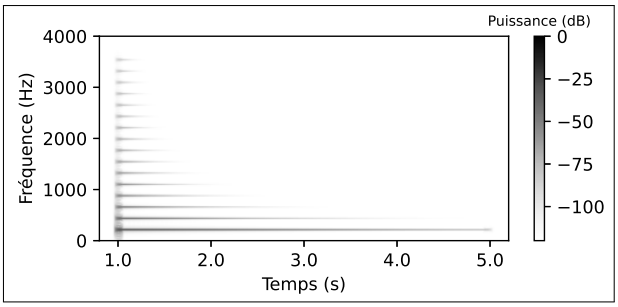


Figure 2. Spectrogramme de la partie harmonique du signal de référence.

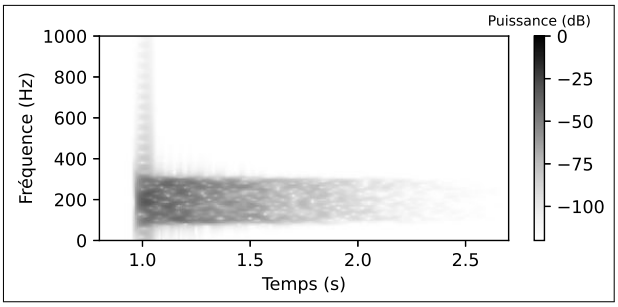


Figure 3. Spectrogramme de la partie non-harmonique du signal de référence.

4. MORPHOLOGIE MATHÉMATIQUE

Nous proposons maintenant d'utiliser quelques outils de la morphologie mathématique pour analyser les spectrogrammes et déduire les paramètres de synthèse des parties harmonique et non-harmonique.

Nous nous plaçons dans le cadre de la morphologie mathématique déterministe définie sur des treillis complets [4]. Nous utilisons ici les opérations de base (dilatation et érosion), le résultat de leurs compositions (ouverture et fermeture) et des opérateurs dérivés (résidus d'ouverture et squelette). Toutes les définitions que nous présentons peuvent être trouvées par exemple dans [3, 9, 15].

4.1. Dilatation et érosion

Définition 4 *Soient* $(L_1, \leq_1)$ *et* $(L_2, \leq_2)$ *deux treillis complets, et* $\vee_i, \wedge_i (i = 1, 2)$ *les supremum et infimum associés. Un opérateur* $\delta : L_1 \to L_2$ *est une* dilatation *si*

$$\forall X_1 \subseteq L_1\,,\ \bigvee\nolimits_2 \delta(X_1) = \delta\left(\bigvee\nolimits_1 X_1\right). \tag{14}$$

Un opérateur $\varepsilon : L_2 \to L_1$ *est une* érosion *si*

$$\forall X_2 \subseteq L_2\,,\ \bigwedge\nolimits_1 \varepsilon(X_2) = \varepsilon\left(\bigwedge\nolimits_2 X_2\right). \tag{15}$$

Ces définitions implicites peuvent prendre des formes particulières explicites, en particulier avec la notion d'*élément structurant*. Nous nous plaçons ici dans le treillis des fonctions dont l'ensemble d'arrivée est un treillis complet.

Proposition 1 *Soit E un ensemble. Soit* $(T, \leq)$ *un treillis complet. Alors, la paire* $(T^E, \leq)$ *où* T^E *est l'ensemble des fonctions* $f : E \to T$ *et* $\leq$ *est la relation d'ordre partiel donnée par*

$$\forall f, g \in T^E, f \leq g \Leftrightarrow \forall p \in E, f(p) \leq g(p) \tag{16}$$

est un treillis complet.

Les opérations de supremum et infimum sont données par : $\forall \{f_i\}_{i\in I} \subseteq T^E$ *(où I est un ensemble d'indices quelconque),*

$$\bigvee_{i\in I} f_i : E \to T, \quad p \mapsto \bigvee_{i\in I} f_i(p) \tag{17}$$

$$\bigwedge_{i\in I} f_i : E \to T, \quad p \mapsto \bigwedge_{i\in I} f_i(p) \tag{18}$$

Nous demandons en outre que $(E, +)$ soit un groupe additif de sorte que l'on puisse parler de translation d'une fonction ; la translation de $f \in T^E$ par l'élément $h \in E$ est définie par

$$T_h f : E \to T, \quad p \mapsto f(p + h)\,. \tag{19}$$

En ce qui concerne T, deux choix sont privilégiés : le premier est le treillis binaire $(\{0, 1\}, \leq)$, ce qui donne lieu à la morphologie binaire. Dans ce cas, les fonctions $\{0, 1\}^E$ peuvent être identifiées aux sous-ensembles de E par la fonction indicatrice, la translation de fonctions devient la translation de sous-ensembles et la relation $\leq$ entre fonctions devient l'inclusion $\subseteq$ entre sous-ensembles. Nous parlons alors du treillis $(\mathcal{P}(E), \subseteq)$. Le deuxième choix est d'utiliser un ensemble numérique comme $\mathbb{R}$, ce qui donne lieu à la morphologie en échelle de gris. Ce dernier choix pose le problème que $(\mathbb{R}, \leq)$ n'est pas un treillis complet, car il n'y a pas d'élément maximal ou minimal. Il est alors habituel de travailler avec l'extension $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

Les propositions suivantes montrent des définitions d'érosion et dilatation dans les cas binaire et en échelle de gris.

Proposition 2 *Soit* $(\mathcal{P}(E), \subseteq)$ *le treillis complet des sous-ensembles d'un groupe additif* $(E, +)$. *Soit* $B \in \mathcal{P}(E)$. *Alors, les opérateurs*

$$\begin{array}{rccc} \delta_B : & \mathcal{P}(E) & \to & \mathcal{P}(E) \\ & A & \mapsto & \{a + b \in E : a \in A, b \in B\} \end{array} \tag{20}$$

$$\begin{array}{rccc} \varepsilon_B : & \mathcal{P}(E) & \to & \mathcal{P}(E) \\ & A & \mapsto & \{p \in E : T_p B \subseteq A\} \end{array} \tag{21}$$

4 . Un treillis est un ensemble partiellement ordonné où toutes les paires d'éléments ont un supremum noté $\vee$ et un infimum noté $\wedge$. Un treillis complet est un treillis dont tous les sous-ensembles ont un supremum et un infimum.

sont respectivement une dilatation et une érosion. Le sous-ensemble $B \subseteq E$ est appelé élément structurant. On parle alors de dilatation et d'érosion binaires avec élément structurant *B.*

Proposition 3 *Soit $(\overline{\mathbb{R}}^E, \leq)$ le treillis complet des fonctions d'un groupe additif $(E, +)$ dans $\overline{\mathbb{R}}$. Soit $\sigma \in \overline{\mathbb{R}}^E$. Alors, les opérateurs*

$$\begin{aligned} \delta_\sigma : \overline{\mathbb{R}}^E &\to \overline{\mathbb{R}}^E \\ f &\mapsto \delta_\sigma[f] : E \to \overline{\mathbb{R}} \\ & \qquad p \mapsto \sup_{h \in E} (f(h) + \sigma(p - h)) \end{aligned} \tag{22}$$

$$\begin{aligned} \varepsilon_\sigma : \overline{\mathbb{R}}^E &\to \overline{\mathbb{R}}^E \\ f &\mapsto \varepsilon_\sigma[f] : E \to \overline{\mathbb{R}} \\ & \qquad p \mapsto \inf_{h \in E} (f(h) - \sigma(h - p)) \end{aligned} \tag{23}$$

sont respectivement une dilatation et une érosion. Par convention, $\infty - \infty$ vaut ∞ pour la dilatation et $-\infty$ pour l'érosion. Ici, l'élément structurant est la fonction σ (appelée aussi fonction structurante*). On parle alors de dilatation et d'érosion en échelle de gris avec élément structurant σ.*

Un cas particulier très fréquent se présente lorsque l'élément structurant est égal à zéro dans un sous-ensemble A de E et $-\infty$ ailleurs [5] ; dans ce cas, la dilatation (resp. l'érosion) d'une fonction f en un point $p \in E$ revient à prendre le supremum (resp. l'infimum) de f dans l'ensemble translaté de A par $p : T_p A = \{a + p \in E : a \in A\}$.

Ces deux opérateurs permettent d'en construire d'autres, en particulier par composition.

4.2. Fermeture et ouverture

Nous rappelons que, si $(L, \leq)$ est un treillis, un opérateur $\psi : L \to L$ est dit

— croissant si $\forall X, Y \in L, X \leq Y \Rightarrow \psi(X) \leq \psi(Y)$,
— extensif si $\forall X \in L, X \leq \psi(X)$,
— anti-extensif si $\forall X \in L, \psi(X) \leq X$,
— idempotent si $\psi^2 = \psi$,

Définition 5 *Soit $(L, \leq)$ un treillis. Soit $\psi : L \to L$ un opérateur. On dit que*

1. *ψ est une* fermeture *si ψ est croissante, extensive et idempotente,*
2. *ψ est une* ouverture *si ψ est croissante, anti-extensive et idempotente.*

De par leurs propriétés de croissance et idempotence, les fermetures et les ouvertures sont des *filtres morphologiques*.

5 . Cela revient à prendre comme élément structurant la fonction indicatrice de A, notée χ_A, en échelle logarithmique.

Comme dans la section précédente, nous construisons des cas particuliers de fermetures et d'ouvertures, par composition d'érosions et de dilatations, définies en particulier à partir d'éléments structurants.

Proposition 4 *Soit $(L, \leq)$ un des deux treillis complets présentés dans la section précédente* [6] *. Soit $\beta \in L$. Alors,*

1. *$\varphi_\beta := \varepsilon_\beta \circ \delta_\beta$ es une fermeture et*
2. *$\gamma_\beta := \delta_\beta \circ \varepsilon_\beta$ es une ouverture.*

4.3. Résidus d'ouverture et squelette

Dans cette section, nous présentons deux opérateurs qui nous seront particulièrement utiles : les résidus d'ouverture et le squelette morphologique.

Les résidus d'ouverture, présentés dans la définition suivante, permettent d'extraire des informations sur la saillance d'une image. Nous présentons cet opérateur pour le cas en échelle de gris.

Définition 6 *Soit $(\overline{\mathbb{R}}^E, \leq)$ le treillis complet présenté dans les sections précédentes. Prendre les résidus de l'ouverture γ d'un élément $f \in \overline{\mathbb{R}}^E$ revient à lui appliquer l'opérateur*

$$\mathbb{1} - \gamma \tag{24}$$

où $\mathbb{1}$ est l'opérateur identité.

Étant donné que les ouvertures sont anti-extensives, on a $\forall f \in \overline{\mathbb{R}}^E$,

$$(\mathbb{1} - \gamma)(f) \geq 0\,. \tag{25}$$

Si nous pensons l'ouverture comme une sorte de régularisation vers le bas, prendre les résidus d'ouverture nous permet de récupérer les éléments saillants qui ont été filtrés.

L'autre opérateur que nous proposons, le squelette morphologique, est présenté pour le cas binaire grâce à une formule due à Lantuéjoul [11]. Nous demandons aussi que l'espace de base E soit $\mathbb{Z}^2$ pour avoir la notion de boule centrée en 0 de rayon $n \in \mathbb{N}$; dans $\mathbb{Z}^2$ nous appelons boule centrée en 0 de rayon n l'ensemble $B_n = \{x \in \mathbb{Z}^2 : ||x|| \leq n\}$, où le choix de la norme détermine sa forme.

Définition 7 *Soit $(\mathcal{P}(E), \subseteq)$ le treillis complet des sous-ensembles de E ordonnés par l'inclusion. Soit $X \in \mathcal{P}(E)$. Le squelette de X s'obtient par la formule*

$$S(X) = \bigcup_{n \in \mathbb{N}} (\varepsilon_{B_n}(X) \setminus \gamma_B(\varepsilon_{B_n}(X))) \tag{26}$$

où ε_{B_n} est l'érosion binaire avec comme élément structurant une boule centrée en 0 de rayon n, et γ est l'ouverture élémentaire, c'est-à-dire où l'élément structurant est la boule centrée en 0 de rayon 1.

6 . $(\mathcal{P}(E), \subseteq)$ pour le cas binaire et $(\overline{\mathbb{R}}^E, \leq)$ dans le cas en échelle de gris.

5. ESTIMATION DES PARAMÈTRES DE SYNTHÈSE

Dans cette section, nous montrons comment les opérateurs morphologiques présentés auparavant peuvent être utilisés pour détecter les amplitudes, temps et fréquences des partiels harmoniques et la réponse en fréquence variant dans le temps du filtre.

Pour illustrer l'efficacité de notre méthode, nous prenons le signal de référence que nous avons construit dans la section 2, dont nous recherchons les paramètres de synthèse. Le spectrogramme du signal de référence est présenté dans la figure 4.

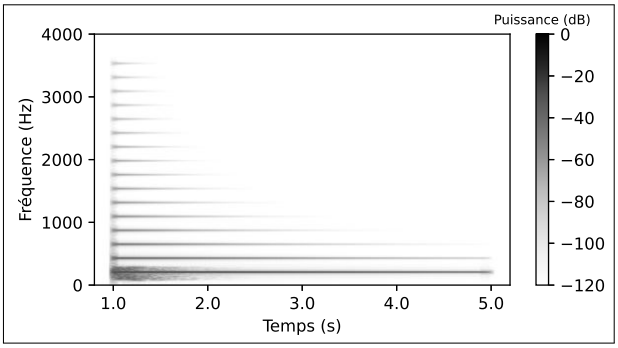


Figure 4. Spectrogramme du signal de référence présenté dans la section 2 et illustré dans la figure 1.

5.1. Estimation des paramètres de la partie harmonique

Pour estimer les paramètres de la partie harmonique, il s'agit d'isoler les lignes horizontales que nous voyons dans la figure 4 en les rétrécissant à des lignes de un pixel d'épaisseur. Pour cela, nous utilisons l'opérateur des résidus d'ouverture ; étant donné que cet opérateur est très sensible au bruit, nous pré-traitons l'image en la filtrant avec une fermeture. Les résidus d'ouverture nous donnent une information de contraste, et nous gardons donc les pixels qui dépassent un certain seuil τ, de sorte à avoir une image binaire. Nous pouvons ensuite calculer le squelette de cette image binaire, et nous obtenons des lignes horizontales d'un pixel d'épaisseur. La donnée des abscisses, ordonnées et intensités de chaque ligne sera considérée comme le résultat de notre processus et nous permettra de synthétiser (après interpolation) un signal qui sera similaire à la partie harmonique du signal d'entrée.

Les éléments structurants présentés par la suite sont définis dans $\mathbb{R} \times \mathbb{R}$ et prennent leurs valeurs dans $\overline{\mathbb{R}}$, par souci de généralité. Pour les calculs, nous réalisons un échantillonnage de ces fonctions en prenant les valeurs aux points de la grille données par le spectrogramme. Souvent, nous prenons des éléments structurants correspondant à des fonctions indicatrices de sous-ensembles A de $\mathbb{R} \times \mathbb{R}$ exprimés en dB : si on appelle σ l'élément structurant, il s'exprime alors comme

$$\sigma = 10 \log_{10}(\chi_A) \,. \tag{27}$$

5.1.1. Fermeture

La première étape consiste à appliquer une fermeture φ_{σ_1} au signal d'entrée, qui permet d'éliminer les variations dues au bruit et à l'étalement spectral. Pour cela, nous prenons comme élément structurant σ_1 la fonction indicatrice du sous-ensemble $A_1 = [0, 0, 05]$ s$\times[0, 50]$ Hz exprimée en dB.

Le résultat de cette opération sur notre signal de référence est montré dans la figure 5.

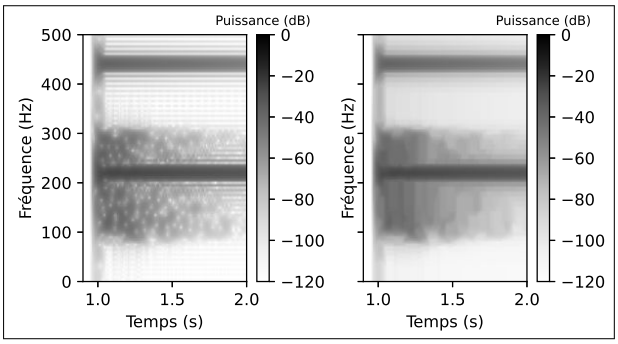


Figure 5. Image d'entrée (à gauche) et le résultat de la fermeture φ_{σ_1} (à droite).

5.1.2. Résidus d'ouverture seuillés

L'étape suivante consiste à prendre les résidus d'ouverture $\mathbb{1} - \gamma_{\sigma_2}$ de la fermeture et à les seuiller à une valeur τ pour obtenir une image binaire. Nous prenons comme élément structurant σ_2 la fonction indicatrice du sous-ensemble $A_2 = \{0\,\text{s}\} \times [0, 25]$ Hz exprimée en dB. Nous utilisons comme seuil $\tau = 3\,\text{dB}$.

Le résultat de cette opération sur notre signal de référence est montré dans la figure 6. Nous voyons comment les lignes horizontales sont isolées du reste.

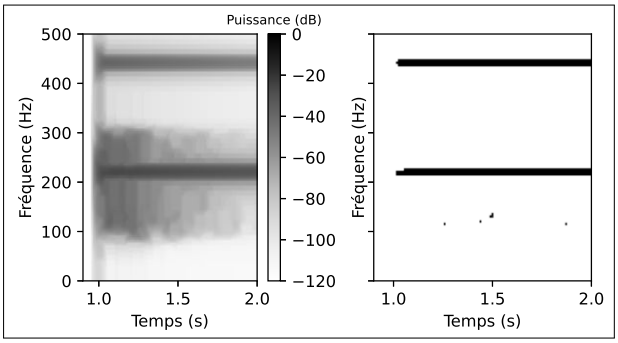


Figure 6. Fermeture (à gauche) et le résultat des résidus d'ouverture $\mathbb{1} - \gamma_{\sigma_2}$ seuillés à 3 dB (à droite).

5.1.3. Squelette morphologique des résidus d'ouverture seuillés

La dernière étape consiste à calculer le squelette morphologique des résidus d'ouverture ; même si les lignes

horizontales des résidus d'ouverture ont déjà isolé les partiels harmoniques, elles ont encore une épaisseur de plusieurs pixels. Nous voulons n'avoir qu'un pixel d'épaisseur et qu'il soit le plus central possible. Pour calculer le squelette, nous utilisons des boules B_n de la pseudo-norme [7] $||(x_1, x_2)|| = |x_2|$, où $(x_1, x_2) \in \mathbb{Z}^2$; ce choix a la particularité qu'elle ne fait rétrécir la forme que dans la direction de la fréquence, propriété qui nous intéresse. Pour la boule centrée en 0 de rayon 1, utilisée dans l'ouverture, nous devons nous restreindre à l'ensemble $\{(0,0),(0,1)\}$ au lieu de l'ensemble $\{(0,-1),(0,0),(0,1)\}$ car sinon nous risquons de faire apparaître des lignes de deux pixels d'épaisseur.

Le squelette morphologique des résidus d'ouverture seuillés est montré dans la figure 7 ; les lignes correspondant aux partiels ont bien été rétrécies jusqu'à une ligne d'un pixel d'épaisseur.

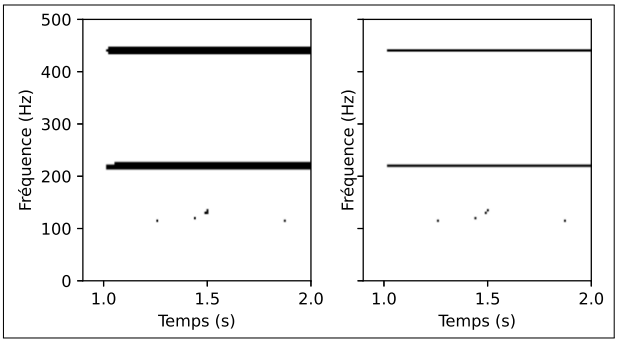


Figure 7. Résidus d'ouverture seuillés (à gauche) et son squelette morphologique (à droite).

5.1.4. Calcul des temps, fréquences et amplitudes des partiels

Finalement, nous nous servons du squelette obtenu pour retrouver les informations des partiels ; pour cela, nous parcourons l'image pixel par pixel et nous parcourons les lignes obtenues en notant à quel temps et à quelle fréquence elles apparaissent. Les amplitudes correspondantes sont les valeurs de l'image fermée [8] .

La figure 8 montre les amplitudes originales correspondant à chacune des fréquences ainsi que le résultat de nos estimations. On constate une parfaite adéquation.

5.2. Estimation des paramètres de la partie non-harmonique

L'estimation des paramètres de la partie non-harmonique est obtenue aussi par des méthodes morphologiques ; nous pré-traitons l'image au moyen d'un fermeture pour limiter l'effet du bruit et de l'étalement spectral (comme dans le cas de la partie harmonique), puis,

7 . Cette définition ne donne pas une norme car elle ne vérifie pas la condition de séparation, mais elle nous sert quand même pour notre cas particulier.

8 . On pourrait argumenter qu'il faudrait prendre celles de l'image originale, mais ce choix est fait pour limiter l'effet du bruit.

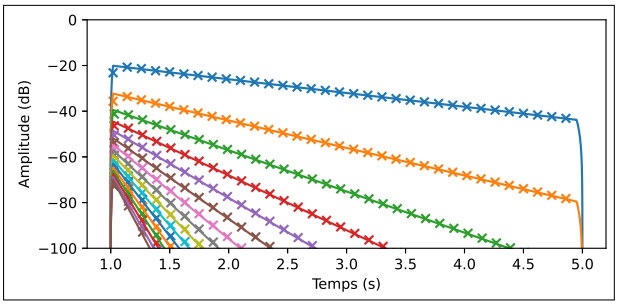


Figure 8. Amplitudes originales (ligne) et nos estimations (croix) ; chaque ligne correspond à une fréquence.

nous appliquons une ouverture pour éliminer les parties saillantes (correspondant aux partiels harmoniques) et garder seulement la partie du spectrogramme correspondant à la partie non-harmonique. Finalement, nous appliquons une érosion pour réduire l'effet de l'étalement spectral.

5.2.1. Fermeture

La fermeture utilisée est la même que dans la section précédente, φ_{σ_1}, et le résultat de son application sur notre image d'entrée est le même que dans la figure 5.

5.2.2. Ouverture

L'ouverture, que nous nommons γ_{σ_3}, a pour élément structurant la fonction indicatrice du sous-ensemble $A_3 = \{0\}$ s $\times$ $[0, 100]$ Hz exprimée en dB. Le résultat sur la fermeture est montré dans la figure 9.

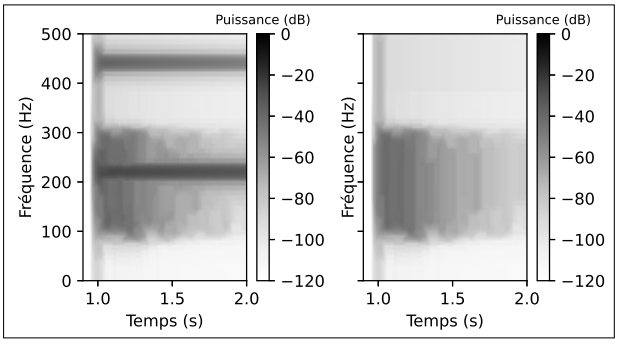


Figure 9. Fermeture (à gauche) et le résultat de l'ouverture γ_{σ_3} (à droite).

5.2.3. Érosion

L'érosion ε_{σ_4} a pour élément structurant une fonction dont la forme est celle de la fenêtre utilisée dans le spectrogramme pour la fréquence 0 et qui vaut zéro ailleurs, en échelle logarithmique, *i.e.* : $\forall (t, \omega) \in \mathbb{R} \times \mathbb{R}$,

$$\sigma_4(t, \omega) = 10 \log_{10}(w(t)\chi_{\{0\}}(\omega)) \,. \qquad (28)$$

Elle sert à concentrer le bruit en temps puisqu'il s'était étalé lors de sa transformation en spectrogramme. Le résultat de cette opération est montré dans la figure 10 ; nous voyons comment l'intensité qui s'étalait autour de l'instant 1 s dans l'ouverture est concentrée après l'application de l'érosion.

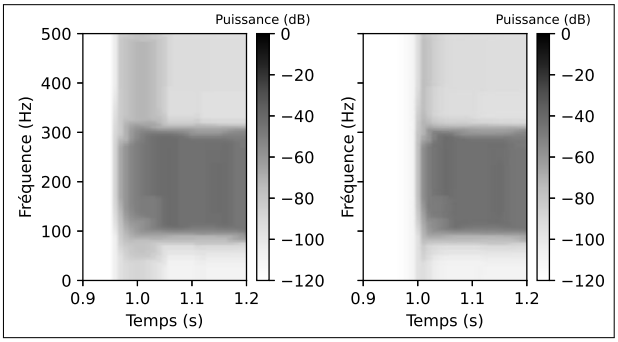


Figure 10. Ouverture (à gauche) et le résultat de l'érosion ε_{σ_4} (à droite).

5.2.4. Filtrage du bruit avec le résultat des opérations

Si nous composons les opérations morphologiques, nous obtenons un résultat $\Theta : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}}$ qui s'écrit

$$\Theta = \varepsilon_{\sigma_4}\left[\gamma_{\sigma_3}\left[\varphi_{\sigma_1}\left[\mathrm{SPEC}^{\mathrm{dB}}[f]\right]\right]\right] . \qquad (29)$$

Il s'agit alors de l'utiliser comme masque pour filtrer notre bruit blanc, et nous procédons comme dans [10] pour utiliser la STFT :

1. Nous générons un bruit blanc.
2. Nous le normalisons pour qu'il ait une puissance maximale [9] de 0 dB.
3. Nous calculons sa STFT.
4. Nous multiplions point par point la STFT du bruit blanc par Θ en échelle linéaire [10].
5. Nous reconstruisons le signal par STFT inverse [11].

Le résultat de ce processus nous donne un signal dont le spectrogramme est montré dans la figure 11 en le comparant à la partie non-harmonique du signal de référence. Nous retrouvons la partie non-harmonique dans notre synthèse, mais aussi du bruit supplémentaire ; celui-ci est dû à l'étalement spectral de la partie harmonique. En revanche, il se situe à une puissance aux alentours de −90 dB, donc sa contribution au signal est relativement petite.

9 . Dans la pratique, si nous normalisons pour que le maximum soit 0 dB, nous aurons une perte de puissance dans le bruit ; pour les applications, nous nous servons de la fermeture du bruit blanc pour normaliser chaque région à son maximum local.

10 . Θ est en échelle logarithmique car les spectrogrammes sont en dB ; Θ en échelle linéaire s'écrit alors $\Theta^{\mathrm{lin}} = 10^{\frac{\Theta}{10}}$.

11 . La STFT inverse nous permet de reconstruire un signal à partir de sa STFT. Pour sa définition, voir [8].

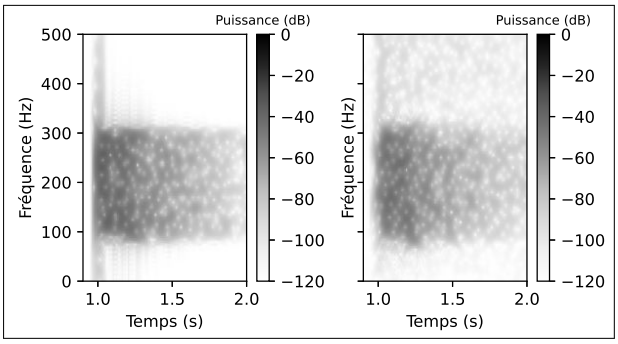


Figure 11. Partie non harmonique du signal de référence (à gauche) le bruit filtré synthétisé par notre méthode (à droite).

6. RÉSULTATS

Jusqu'à présent, nous avons testé notre méthode sur un signal de référence, obtenant un très bon résultat pour la partie harmonique et un résultat correct pour la partie non-harmonique. La méthode a ensuite été appliquée à des sons d'instruments issus la base de données TinySOL [4] qui a été récupérée grâce à la librairie `mirdata` [2].

Nous invitons le lecteur à consulter

`https://github.com/Manza12/JIM-2022.`

pour écouter les résultats. Dans le `README.md`, des instructions sont données pour l'écoute des sons et l'utilisation du code, disponible en libre accès [12].

6.1. Bois

Nous avons testé la méthode sur des notes de flûte, clarinette en si bémol, hautbois, basson et saxo alto. Les résultats sont très convaincants ; la partie harmonique est très bien resynthétisée (sauf, parfois, le transitoire d'attaque) et la partie non-harmonique simule l'effet du souffle. La figure 12 montre une comparaison des spectrogrammes des sons d'entrée et de sortie d'un La 3 de flûte.

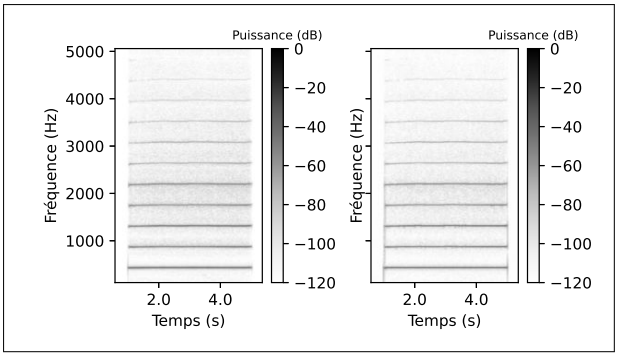


Figure 12. Comparaison du spectrogramme du signal d'entrée (à gauche) et de sortie (à droite) d'un La 3 de flûte.

12 . Licence GNU General Public License v3.0.

6.2. Cuivres

Pour les cuivres, nous avons testé la méthode sur la trompette, le cor, le trombone et le tuba. Les résultats sont moins bons, notamment en ce qui concerne le transitoire d'attaque, très caractéristique des cuivres. La partie harmonique oscille plus que dans le signal original, notamment dans les graves, à cause de la résolution fréquentielle limitée et du fait que le squelette donne des pixels qui ne sont pas toujours alignés. La partie non-harmonique correspondant au souffle est retrouvée comme dans le cas des bois. La figure 13 montre un Do 2 de trombone et le résultat de sortie.

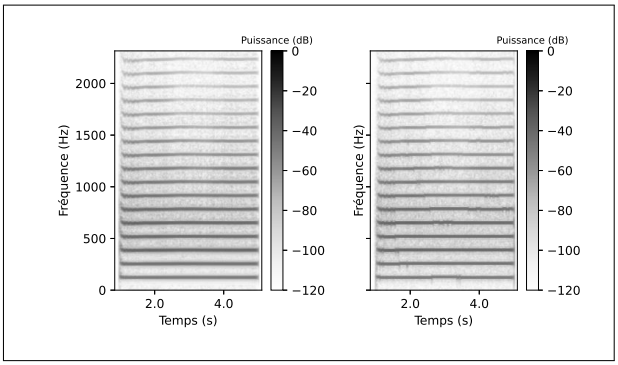


Figure 13. Comparaison du spectrogramme du signal d'entrée (à gauche) et de sortie (à droite) d'un Do 2 de trombone.

6.3. Cordes

Le cas des cordes est plus variable. Nous avons testé la méthode sur des sons de contrebasse, violoncelle, alto et violon. Les résultats dépendent de deux facteurs principaux ; le premier est la tessiture : quand les sons sont trop graves, comme dans le cas de la contrebasse, la résolution fréquentielle limitée fait que les partiels se confondent avec le bruit. Il faut alors augmenter la taille de la fenêtre de la STFT et diminuer les largeurs des éléments structurants de la fermeture et de l'ouverture pour qu'ils s'adaptent à des fréquences plus serrées. Les résultats s'améliorent mais restent moins bons que dans les tessitures plus hautes. L'autre facteur est le vibrato ; quand l'instrumentiste vibre, les résidus d'ouverture ont du mal à détecter cette modulation de fréquence rapide et rendent le squelette non connexe [13] , surtout dans les aigus. La figure 14 montre cet effet. Cela provoque des artéfacts très perceptibles. En revanche, quand le son est produit sur une corde à vide, le son est très réussi et on retrouve bien le son de l'archet dans la partie non-harmonique. Il faut noter aussi que les attaques sont bien meilleures que dans le cas des vents, puisque plus progressives.

13 . Notons qu'une squelettisation par amincissement homotopique pourrait en partie résoudre ce problème.

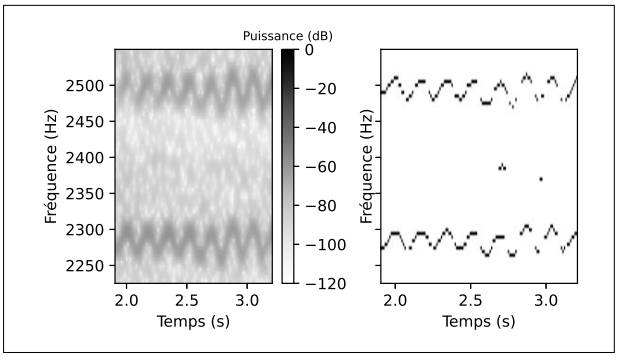


Figure 14. Comparaison du spectrogramme du signal d'entrée (à gauche) et le résultat du squelette (à droite) dans des partiels aigus d'un La bémol 2 de violon vibré.

7. CONCLUSION

Dans cet article, nous avons montré comment les opérateurs morphologiques appliqués à des spectrogrammes de sons d'instruments de musique peuvent nous aider à estimer des paramètres permettant de reconstruire les sons avec une qualité considérable, à la fois pour la partie harmonique et pour la partie non-harmonique.

Parmi ces opérateurs, le squelette des résidus d'ouverture s'avère particulièrement intéressant comme méthode pour identifier les partiels harmoniques car il parvient à réduire les lignes horizontales du spectrogramme à un pixel d'épaisseur. L'ouverture a été aussi très utile pour filtrer ces lignes et retrouver la réponse en fréquence d'un filtre qui varie dans le temps avec laquelle nous filtrons un bruit blanc.

Pour les travaux futurs, nous proposons de privilégier deux voies d'amélioration. En premier lieu, l'attaque est problématique à synthétiser par cette méthode ; une idée serait d'ajouter une composante qui lui soit particulière et voir comment estimer ses paramètres. Étant donné que l'attaque est représentée comme une ligne verticale dans un spectrogramme, nous pourrions utiliser des techniques similaires à celles de la partie harmonique, mais adaptées à la détection des lignes saillantes verticales au lieu des lignes saillantes horizontales.

En second lieu, nous pourrions améliorer la détection des partiels harmoniques en utilisant une autre technique de morphologie mathématique appelée *amincissement* [5] pour calculer un squelette homotope à l'ensemble de départ ; elle pourrait ainsi résoudre le problème de connexité des squelettes des notes vibrées. De plus, cette technique appliquée directement à l'image en niveaux de gris éviterait la binarisation de l'image.

8. REFERENCES

[1] James W. Beauchamp. Analysis and synthesis of musical instrument sounds. In *Analysis, Synthesis, and Perception of Musical Sounds*, pages 1–89. Springer, Urbana, USA, 2007.

[2] Rachel M Bittner, Magdalena Fuentes, David Rubinstein, Andreas Jansson, Keunwoo Choi, and Thor Kell. mirdata : Software for Reproducible Usage of Datasets. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 99–106, Delft, The Netherlands, 2019.

[3] Isabelle Bloch, Henk Heijmans, and Christian Ronse. Mathematical Morphology. In *Handbook of spatial logic*. Springer, Dordrecht, The Netherlands, 2007.

[4] C.E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz. OrchideaSOL : a dataset of extended instrumental techniques for computer-aided orchestration. In *Proceedings of the International Computer Music Conference*, Santiago, Chile, 2020.

[5] Michel Couprie, Nivando Bezerra, and Gilles Bertrand. A Parallel Thinning Algorithm for Grayscale Images. In *Proceedings of the 17th IAPR International Conference on Discrete Geometry for Computer Imagery*, pages 71–82, Seville, Spain, 2013.

[6] Ali Daher, El Houssaïn Baghious, Gilles Burel, and Emanuel Radoi. Overlap-Save and Overlap-Add Filters : Optimal Design and Comparison. *IEEE Transactions on Signal Processing*, 58(6) :3066–3075, 2010.

[7] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. DDSP : Differentiable Digital Signal Processing. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[8] Karlheinz Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston, 2001.

[9] H. J. A. M Heijmans and C Ronse. The algebraic basis of mathematical morphology I. Dilations and erosions. *Computer Vision, Graphics, and Image Processing*, 50(3) :245–295, 1990.

[10] Nian-chyi Huang and J. K. Aggarwal. Time-varying digital signal processing. In *Proceedings of the 19th IEEE Conference on Decision and Control, Including the Symposium on Adaptive Processes*, pages 586–587, Albuquerque, USA, 1980.

[11] Christian Lantuejoul. *La squelettisation et son application aux mesures topologiques des mosaiques polycristallines*. PhD thesis, École des Mines, Paris, 1978.

[12] Georges Matheron. *Random sets and integral geometry*. J. Wiley & Sons, New York, USA, 1975.

[13] Lawrence R. Rabiner and Bernard Gold. *Theory and Application of Digital Signal Processing*. Prentice Hall, Englewood Cliffs, USA, 1975.

[14] Richard A. Roberts and Clifford T. Mullis. *Digital signal processing*. Addison-Wesley, Reading, USA, 1987.

[15] C. Ronse and H. J. A. M. Heijmans. The algebraic basis of mathematical morphology : II. Openings and closings. *CVGIP : Image Understanding*, 54(1) :74–97, 1991.

[16] Jean P. Serra. *Image analysis and mathematical morphology*. Academic Press, Orlando, USA, 1982.

[17] Xavier Serra. Musical Sound Modeling with Sinusoids plus Noise. In *Musical Signal Processing*, pages 91–122. Routledge, New York, USA, 1997.

[18] Xavier Serra and Julius Smith. Spectral modeling synthesis : A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4) :12–24, 1990.