Transformée de scattering en spirale temps-chroma-octave

Vincent LOSTANLEN, Stéphane MALLAT

Département d'Informatique, École normale supérieure
45 rue d'Ulm, Paris
vincent.lostanlen@ens.fr

Résumé – Dans le cadre de la représentation temps-fréquence des sons harmoniques, on montre l'intérêt de remplacer l'axe fréquentiel par une spirale faisant un tour à chaque octave. On montre que les déformations mélodiques et harmoniques du modèle source-filtre sont naturellement séparées dans ce modèle. On construit une transformée de scattering multivariable tirant parti de ces résultats.

Abstract – In the framework of time-frequency representation of harmonic sounds, we motivate the idea of replacing the frequency axis by a spiral which makes a full turn at each octave. We show that melodic and harmonic deformations of the source-filter model are naturally disentangled in this model. We capitalize on these results to build a multivariable scattering transform.

1 Introduction

Un défi majeur de la classification automatique de sons repose sur une modélisation efficace de leur structure transitoire sur des échelles temporelles aussi longues que possible. De par sa bonne localisation temps-fréquence et sa faculté de régularisation des signaux modulés, un opérateur non-linéaire tel que le module de la transformée en ondelettes est un premier pas naturel dans ce sens. Cependant, celui-ci est incapable de capturer, par simple intégration temporelle, des éléments acoustiques plus riches tels que les variations de fréquence fondamentale (*chirps*) ou de profil formantique (coarticulations, attaques instrumentales). Or, si le cas des *chirps* et de la variabilité harmonique ont été abordés indépendamment ([Flandrin], [Peeters et al.]), il n'existe pas d'approche systématique qui rende compte de la dynamique jointe de ces deux facteurs.

Dans cet article, nous introduisons une nouvelle représentation des sons, construite à partir du module de la transformée en ondelettes, visant explicitement à caractériser les changements de hauteur et de timbre. Dans une première partie, nous montrons comment enrouler l'axe fréquentiel en une spirale des hauteurs de sorte qu'un tour complet correspond à une transposition d'une octave, afin de séparer hauteur tonale (chroma) et hauteur spectrale (octave). Nous démontrons l'intérêt de cette approche à travers une formulation transitoire du modèle source-filtre. Par la suite, nous définissons un opérateur unitaire et multi-échelles sur la spirale obtenue, construit comme une cascade de trois transformées en ondelettes à valeurs complexes. Nous illustrons le comportement de cet opérateur sur un signal de parole.

Ce travail est financé par la bourse ERC InvariantClass 32095. Le code source des expériences et figures est en libre accès à l'adresse www.github.com/lostanlen/scattering.m.

2 Du temps-fréquence au temps-chroma-octave

Les paradoxes de hauteur synthétisés par Shepard et Risset [Risset] montrent que la perception de la hauteur n'est pas réductible au continuum grave-aigu des fréquences physiques. En effet, en sommant des sinusoïdes de fréquence $2^n f_0$ avec $n \in \mathbb{Z}$, on obtient une note que l'on peut nommer sur une gamme musicale, bien qu'elle ne puisse être localisée dans le grave ou l'aigu. Dès lors, en faisant progressivement croître f_0 jusqu'à $2f_0$, on peut construire un *glissando* qui semble monter indéfiniment lorsqu'il est répété. Par ailleurs, en filtrant les sinusoïdes dans une bande dont la fréquence centrale diminue, on peut donner l'impression de passer du registre aigu au registre grave tout en restant sur la même note. La composition de ces deux opérations produit un son paradoxal qui monte localement tout en descendant globalement. Cette expérience met en lumière le fait qu'il y a en fait deux attributs perceptifs associés à la fréquence. Le premier, appelé hauteur tonale ou *chroma*, encode la mélodie musicale et l'intonation de la voix ; le second, appelé hauteur spectrale ou simplement *octave*, est une composante essentielle du timbre.

Afin de prendre en compte ces propriétés, on choisit de remplacer l'axe rectilinéaire des hauteurs par un schéma en spirale, à raison d'un tour par octave : un changement de chroma est alors analogue à une rotation, tandis qu'un changement d'octave est une dilatation par rapport au centre de la spirale. Cette idée a été confirmée par de nombreuses expériences en psychoacoustique, mais aussi en neurologie de l'audition [Warren].

Soit $\psi(t)$ un filtre passe-bande de fréquence centrale $\eta > 0$ et dont le facteur de qualité Q est de l'ordre de 10. On suppose que ψ est une fonction analytique, c'est-à-dire que $\hat{\psi}(\omega) = 0$ pour $\omega \leq 0$. L'analyse en ondelettes consiste à dilater $\psi(t)$ par

FIGURE 1 – À gauche, un scalogramme du mot anglais *lion*. À droite, le spiralogramme correspondant : le temps est la coordonnée longitudinale, le chroma la coordonnée angulaire, l'octave la coordonnée radiale.

différents facteurs d'échelle 2^γ ; les filtres obtenus s'écrivent

$$\psi_\gamma(t) = 2^\gamma \psi(2^\gamma t), \text{ c'est-à-dire } \widehat{\psi_\gamma}(\omega) = \hat{\psi}(2^{-\gamma}\omega). \quad (1)$$

Ainsi, chaque $\psi_\gamma(t)$ est un filtre passe-bande de fréquence centrale $2^{-\gamma}\eta$ et de facteur de qualité Q. Par ailleurs, en notant τ le support typique de $\psi(t)$, chaque $\psi_\gamma(t)$ a un support $2^\gamma\tau$.

Soit $J \in \mathbb{N}^*$; en discrétisant les valeurs de γ dans l'intervalle $[0; J[$ par pas de N^{-1} avec $N \geq Q$ entier, on construit un banc de filtres couvrant les fréquences de $2^{-J}\eta$ à η, soit J octaves exactement. Afin de constuire une transformée qui conserve l'énergie de tout signal réel, il faut ajouter au banc $\{\psi_\gamma\}_{\gamma<J}$ un filtre passe-bas $\phi(t)$ à valeurs réelles, défini dans le domaine de Fourier par : $\forall \omega \geq 0,\ \hat{\phi}(\pm\omega) = \sqrt{1 - \frac{1}{2}\sum_{\gamma<J} |\widehat{\psi_\gamma}(\omega)^2|^2}$. Pourvu que $\psi(t)$ soit judicieusement normalisé, on peut vérifier que les filtres construits couvrent l'intervalle de fréquences $[0;\ \eta]$ de façon quasiment égale :

$$\forall \omega \in [0;\ \eta],\ 1 - \varepsilon \leq |\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_{\gamma<J} |\widehat{\psi_\gamma}(\omega)|^2 \leq 1. \quad (2)$$

Soit W l'opérateur associant à tout signal réel $x(t)$ le vecteur de signaux

$$Wx(t) = \begin{pmatrix} x * \psi_\gamma(t) \\ x * \phi(t) \end{pmatrix}_{\gamma<J}. \quad (3)$$

En appliquant la formule de Plancherel [Mallat], on peut vérifier que l'équation 2 est équivalente à $(1-\varepsilon)\|x\|_2^2 \leq \|Wx\|_2^2 \leq \|x\|_2^2$ pour $\varepsilon \ll 1$ fixé ; ce que l'on propose de noter $\|Wx\|_2 \lesssim \|x\|_2$. En outre, l'inégalité triangulaire permet de vérifier que W est contractant : on a $\|Wx - Wy\|_2 \leq \|x - y\|_2$ pour tous signaux $x(t)$ et $y(t)$, ce qui contribue à lutter contre la « malédiction de la dimensionnalité » en classification.

On appelle *scalogramme* $U_1x\,(t,\gamma) = |x * \psi_\gamma|(t)$ le module de la transformée en ondelettes du signal $x(t)$. Afin de faire apparaître la structure en spirale de la variable d'échelle γ, on décompose $\gamma = j + \frac{\chi}{N}$, où les entiers $j < J$ et $\chi < N$ sont respectivement la variable d'octave et de chroma. Sur la figure 1, on a représenté le scalogramme d'un signal de parole et le « spiralogramme » correspondant, que l'on continue de noter $U_1x(t, \chi, j)$.

3 Déformations du modèle source-filtre

Soient $e(t) = \sum_{k=1}^K \exp(\mathrm{i}kf_0t)$ un signal harmonique « source » et $t \mapsto \alpha(t)$ un difféomorphisme ; on définit $e_\alpha(t) = (e \circ \alpha)(t)$ la source déformée. De même, on part d'un « filtre » $h(t)$ et d'un difféomorphisme $t \mapsto \beta(t)$ pour définir $h_\beta(t) = (h \circ \beta)(t)$. Le modèle source-filtre déformé est le signal $x(t) = [e_\alpha * h_\beta](t)$.

Lemme 1. *Pour tout γ tel que*

(a) $\|\ddot{\beta}/\dot{\beta}\|_\infty \ll 2^{-\gamma}\eta/Q$ *(filtre lentement variable) et*

(b) $\|\dot{\hat{h}}/\hat{h}\|_\infty \|1/\dot{\beta}\|_\infty \ll 2^\gamma Q/\eta$ *(réponse fréquentielle régulière)*

on a

$$[h_\beta * \psi_\gamma](t) \approx H\left(\gamma + \log_2 \dot{\beta}(t)\right) \psi_\gamma\left(\frac{\beta(t)}{\dot{\beta}(t)}\right) \quad (4)$$

où $H(\gamma) = \hat{h}(2^{-\gamma}\eta)$.

Démonstration. Grâce à la première hypothèse, on développe $\beta(t-u) \approx \beta(t) - \dot{\beta}(t) \times u$ sur le support de $\psi_\gamma(t)$. Le changement de variable $u' = \dot{\beta}(t) \times u$ conduit à

$$[h_\beta * \psi_\gamma](t) = \int_{\mathbb{R}} h(\beta(t) - u')\psi_\gamma\left(\frac{u'}{\dot{\beta}(t)}\right) \frac{\mathrm{d}u'}{\dot{\beta}(t)}. \quad (5)$$

L'ondelette ψ_γ vérifiant $\psi_\gamma(\dot{\beta}(t)^{-1}u') = \dot{\beta}(t)\psi_{\gamma+\log_2\dot{\beta}(t)}(u')$, on peut convertir le facteur de dilatation $\dot{\beta}(t)$ en une transposition fréquentielle. D'où $[h_\beta * \psi_\gamma](t) = [h * \psi_{\gamma+\log_2\dot{\beta}(t)}](t)$, ce qui s'écrit comme un produit dans le domaine de Fourier :

$$[h_\beta * \psi_\gamma](t) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{h}(\omega)\hat{\psi}_{\gamma+\log_2\dot{\beta}(t)}(\omega)\exp(\mathrm{i}\omega\beta(t))\,\frac{\mathrm{d}u'}{\dot{\beta}(t)}. \quad (6)$$

Grâce à la seconde hypothèse, on approxime $\hat{h}(\omega)$ par la constante $H(\gamma) = \hat{h}(2^{-\gamma}\eta)$ sur le support de $\hat{\psi}_{\gamma+\log_2\dot{\beta}(t)}$. Dès lors, l'intégrale ci-dessus peut être vue comme la transformée de Fourier inverse de $\hat{\psi}_{\gamma+\log_2\dot{\beta}(t)}(\omega)$ évaluée en $\beta(t)$. On conclut en revenant à l'ondelette ψ_γ avec l'équation $\dot{\beta}(t)^{-1}\psi_{\gamma+\log_2\dot{\beta}(t)}(\beta(t)) = \psi_\gamma(\beta(t)/\dot{\beta}(t))$. □

Théorème 1. *Soit, pour un partiel fixé k, l'échelle correspondante $\gamma = \log_2(kf_0\eta^{-1})$. Si les conditions suivantes sont vérifiées :*

(a) $\|\ddot{\beta}/\dot{\beta}\|_\infty \ll 2^{-\gamma}\eta/Q$ *(filtre lentement variable),*

(b) $\|\dot{\hat{h}}/\hat{h}\|_\infty \|1/\dot{\beta}\|_\infty \ll 2^\gamma Q/\eta$ *(réponse fréquentielle régulière),*

(c) $\|\ddot{\alpha}/\dot{\alpha}\|_\infty \ll 2^{-\gamma}\eta/Q$ *(source lentement variable) et*

(d) $k < Q/2$ *(partiel de rang faible),*

alors le module de la transformée en ondelettes du modèle source-filtre déformé

$$|e_\alpha * h_\beta * \psi_\gamma|(t) \approx E(\gamma + \log_2\dot{\alpha}(t))H(\gamma + \log_2\dot{\beta}(t)) \quad (7)$$

est localement séparable en une réponse de source $E(\gamma) = |\widehat{\psi_\gamma}(kf_0)|$ et une réponse de filtre $H(\gamma) = \hat{h}(2^{-\gamma}\eta) = \hat{h}(kf_0)$, chacune en mouvement rigide sur l'axe γ, celui-ci étant régi par le signal $\log_2\dot{\alpha}(t)$ (resp. $\log_2\dot{\beta}(t)$).

Démonstration. On part des hypothèses (a) et (b) pour affirmer le lemme précédent :

$$[e_\alpha * h_\beta * \psi_\gamma]\,(t) = H\left(\gamma + \log_2\dot{\beta}(t)\right) \times \int_{\mathbb{R}} e_\alpha(t-u)\psi_\gamma\left(\frac{\beta(u)}{\dot{\beta}(u)}\right)\,\mathrm{d}u. \quad (8)$$

Comme dans la preuve du lemme, on pose $u' = \dot\alpha(t) \times (\frac{\beta(t)}{\dot\beta(t)} + u - t)$, on développe et simplifie $\frac{\beta(u)}{\dot\beta(u)} \approx \frac{u'}{\dot\alpha(t)}$, et l'on convertit la dilatation temporelle en transposition fréquentielle avec l'équation $\dot\alpha(t)^{-1}\psi_\gamma(\dot\alpha(t)^{-1}u') = \psi_{\gamma+\log_2\dot\alpha(t)}(u')$:

$$\begin{aligned}&\int_{\mathbb{R}} e_\alpha(t-u)\psi_\gamma\left(\tfrac{\beta(u)}{\dot\beta(u)}\right)\mathrm{d}u\\ &= \int_{\mathbb{R}} e_\alpha\left(\tfrac{\beta(t)}{\dot\beta(t)} - \tfrac{u'}{\dot\alpha(t)}\right)\psi_{\gamma+\log_2\dot\alpha(t)}(u')\,\mathrm{d}u'\end{aligned} \tag{9}$$

Avec l'hypothèse (3), on linéarise le difféomorphisme α autour de $\frac{\beta(t)}{\dot\beta(t)}$, ce qui permet de voir l'intégrale ci-dessus comme la convolution $[e * \psi_{\gamma+\log_2\dot\alpha(t)}]$ évaluée en $\alpha(\frac{\beta(t)}{\dot\beta(t)})$. Puisque le banc de filtres a un facteur de qualité constant Q, la largeur de bande à la fréquence $kf_0\dot\alpha(t)$ est $kf_0\dot\alpha(t)Q^{-1}$. L'hypothèse (4) peut se réécrire $(k+1)f_0\dot\alpha(t) > kf_0 + \frac{kf_0}{2Q}$; autrement dit, le $(k+1)^{\text{ème}}$ partiel est hors de la bande passante de $\psi_{\gamma+\log_2\dot\alpha}$. Plus généralement, les partiels $k' \neq k$ ont une contribution négligeable à la transformée en ondelettes de $e(t)$. En l'absence d'interférences, le module $|e*\psi_{\gamma+\log_2\dot\alpha(t)}|(t)$ se résume au seul terme $E(\gamma+\log_2\dot\alpha(t))$ où l'on a défini $E(\gamma) = |\widehat{\psi_\gamma}(kf_0)|$. □

On peut calculer explicitement $E(\gamma) = \sum_{k=1}^{K}\delta(\gamma - \log_2(kf_0\eta^{-1}))$, ce qui montre que $E(\gamma+\Delta_j) = E(\gamma)$ pour tout $\Delta_j \in \mathbb{N}$. Par ailleurs, pourvu que la déviation maximale Δ_ν du chirp soit petite devant les variations typiques de H, il est possible d'écrire $H(\gamma+\Delta\nu) \approx H(\gamma)$. Ce résultat suggère qu'il est possible de séparer les fonctions $\log_2\dot\alpha(t)$ et $\log_2\dot\beta(t)$ à partir du spiralogramme $U_1x(t,\nu,j)$, en décomposant leur trajectoire temporelle selon les variables t, γ et j.

4 Transformées en ondelettes sur la spirale

L'idée d'effectuer une nouvelle transformée en ondelettes sur $U_1x(t,\gamma)$ est due à [Andén]. Avec le même procédé qu'à la section 1, on construit un second banc de filtres $\psi^{\text{temps}}_{\gamma_2}(t) = 2^{\gamma_2}\psi^{\text{temps}}(2^{\gamma_2}t)$ pour différents $\gamma_2 < J$ à partir d'une ondelette ψ^{temps} de fréquence centrale η et de facteur de qualité Q_2 ; le filtre passe-bas correspondant est noté $\phi^{\text{temps}}(t)$. On définit le spectre de modulation d'amplitude $Y_2^{\text{temps}}x(t,\gamma,\gamma_2) = [U_1x \overset{t}{*} \psi^{\text{temps}}_{\gamma_2}](t,\gamma)$ et le spectre acoustique moyen $S_1x(t,\gamma) = [U_1x \overset{t}{*} \phi](t)$, de sorte que l'énergie du scalogramme est conservée : $\|S_1x\|_2^2 + \|Y_2^{\text{temps}}x\|_2^2 \lesssim \|U_1x\|_2^2$. Bien que cette seconde transformée apporte de l'information sur la dynamique temporelle de chaque canal fréquentiel, son module $U_2x = |Y_2x|$ ne suffit pas à caractériser les chirps, car il ne préserve pas leur cohérence temporelle à travers des chromas voisins. Pour remédier à ce problème, on peut penser à effectuer une transformée selon la variable γ avant d'appliquer le module : cette idée se rapproche du modèle *Spectro-Temporal Receptive Fields* ou STRF [Patil]. Les coefficients qui en résultent sont bien adaptés aux modulations d'amplitude et de fréquence ; mais, dans le cas d'un signal harmonique, ils ne rendent pas compte de la

FIGURE 2 –

FIGURE 3 –

régularité de U_1x à travers des octaves voisines, à chroma fixé. Pour tirer parti de cette régularité, nous proposons de transformer les variables t, γ et j avant d'appliquer le module.

On part d'une ondelette analytique $\psi^{\text{chroma}}(\gamma)$ de fréquence centrale η^{chroma}, que l'on dilate par des facteurs $2^{\gamma::\gamma}$ — la variable $\gamma :: \gamma$, prononcée « γ selon γ », est construite avec l'opérateur infixe :: de construction de liste chaînée, en analogie avec les langages ML. Il faut remarquer que, puisque le signal $Y_2^{\text{temps}}x$ est complexe, l'intervalle de fréquences à couvrir n'est plus $[0;\eta]$ mais $[-\eta;\eta]$. Les fréquences centrales sont donc de la forme $(\theta :: \gamma)2^{-\gamma::\gamma}\eta^{\text{chroma}}$, où $(\theta :: \gamma) = \pm 1$ est une variable de signe. L'équation 1 devient dans ce cas

$$\psi^{\text{chroma}}_{\gamma::\gamma,\theta::\gamma}(\gamma) = (\theta :: \gamma)2^{\gamma::\gamma}\psi^{\text{chroma}}((\theta :: \gamma)2^{\gamma::\gamma}\gamma). \tag{10}$$

Enfin, on fait de même avec la variable d'octave : un banc de filtres discret $\{\psi^{\text{octave}}_{\gamma::j,\theta::j}\}$ prenant ses valeurs dans les entiers $j < J$ est créé, ainsi que le filtre passe-bas correspondant ϕ^{octave}.

$$\|U_2x\|_2^2 \lesssim \|Y_2^{\text{temps}}x\|_2^2 \tag{11}$$

$$W_2U_1x = \begin{pmatrix} U_2x \\ S_1x \end{pmatrix} \tag{12}$$

$$\|W_2U_1x\|_2 \lesssim \|U_1x\| \lesssim \|x\|. \tag{13}$$

4.1 Bancs de filtres

4.2 Opérateur non-linéaire

$$W_2U_1x(t,\gamma) = \begin{pmatrix} U_1x \overset{t}{*} \psi^{\text{temps}}_{\gamma_2} \overset{\gamma}{*} \psi^{\text{chroma}}_{\gamma::\gamma,\theta::\gamma} \overset{j}{*} \psi^{\text{octave}}_{\gamma::j,\theta::j} \\ U_1x \overset{t}{*} \psi^{\text{temps}}_{\gamma_2} \overset{\gamma}{*} \phi^{\text{chroma}} \overset{j}{*} \psi^{\text{octave}}_{\gamma::j,\theta::j} \\ U_1x \overset{t}{*} \psi^{\text{temps}}_{\gamma_2} \overset{\gamma}{*} \psi^{\text{chroma}}_{\gamma::\gamma,\theta::\gamma} \overset{j}{*} \phi^{\text{octave}} \\ U_1x \overset{t}{*} \psi^{\text{temps}}_{\gamma_2} \overset{\gamma}{*} \phi^{\text{chroma}} \overset{j}{*} \phi^{\text{octave}} \\ U_1x \overset{t}{*} \phi^{\text{temps}} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \gamma_2 < J \\ \gamma :: \gamma < J :: \gamma \\ \theta :: \gamma = \pm 1 \\ \gamma :: j < J :: j \\ \theta :: j = \pm 1 \end{matrix} \tag{14}$$

Références

[1] J. Andén, S. Mallat. Deep Scattering Spectrum. *IEEE Transactions on Signal Processing*, vol. 62, n° 16, p. 4114–4128, 2014.

[2] P. Flandrin. Time-frequency and chirps. In *Proc. SPIE Meeting Wavelet Applications VIII*, vol. 4391, p. 161–175, Orlando (FL), 2001.

[3] S. Mallat. *Une exploration des signaux en ondelettes*. Les Éditions de l'École polytechnique, 2000.

[4] J.-C. Risset. Paradoxes de hauteur. Rapport Ircam 11/78, 1978.

[5] K. Patil, D. Pressnitzer, S. Shamma, M. Elhilali. Music in our ears : the biological bases of musical timbre perception. *PLoS computational biology*, vol. 8, n° 11, 2012.

[6] J. D. Warren, S. Uppenkamp, R. D. Patterson, T. Griffiths. Separating pitch chroma and pitch height in the human brain. *Proceedings of the National Academy of Sciences*, vol. 100, n° 17, p. 10038–10042, 2003.