Transformée en scattering sur la spirale temps-chroma-octave

Vincent LOSTANLEN, Stéphane MALLAT

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

Résumé – On introduit une représentation en scattering pour l'analyse et la classification des sons. Elle est localement invariante par translation, stable par déformation en temps et en fréquence, et elle capture les structures harmoniques. Cette représentation en scattering peut s'interpréter comme un réseau de neurones convolutif, calculé en cascadant une transformée en ondelettes dans le temps, et le long d'une spirale harmonique. Nous étudions son application pour l'analyse des déformations du modèle source-filtre.

Abstract – We introduce a scattering representation for the analysis and classification of sounds. It is locally translation-invariant, stable to deformations in time and frequency, and has the ability to capture harmonic structures. The scattering representation can be interpreted as a convolutional neural network which cascades a wavelet transform in time and along a harmonic spiral. We study its application for the analysis of the deformations of the source-filter model.

1 Introduction

La variabilité des signaux acoustiques naturels peut se modéliser comme une action de déformation localisée en temps et en fréquence. Ainsi, la classification de sons repose essentiellement sur la construction de représentations qui demeurent stables à ces déformations, tout en offrant une bonne discriminabilité entre signaux de classes différentes. En cascadant convolutions locales et non-linéarités, les représentations en réseaux de neurones parviennent à combiner ces deux qualités ; mais elles sont entièrement adaptées aux données, et requièrent par conséquent une vaste base d'entraînement pour atteindre des performances de classification élevée. Dans cet article, nous proposons une représentation en cascade, dite transformée de scattering, dont l'architecture est similaire à un réseau de neurones, mais sans besoin d'optimiser les unités de convolution. On tire parti de la géométrie naturelle des sons pour construire une description stable aux déformations, mais qui préserve l'information discriminante autant que possible. Un enjeu fondamental de cette approche réside dans la préservation de la structure harmonique des partiels, y compris lorsque celle-ci est sujette à des variations d'amplitude, de hauteur et de timbre. En enroulant l'axe log-fréquentiel en une spirale, de sorte que des partiels sur des octaves consécutives se trouvent alignés, on fait apparaître la régularité de l'enveloppe spectrale sur un axe radial. La représentation en scattering consiste alors en une cascade de trois décompositions en ondelettes selon les trois variables introduites : temps, chroma et octave.

Cette idée nous conduira à réorganiser l'axe fréquentiel en spirale,

2 Transformées sur le scalogramme

2.1 Scalogramme et scattering temporel

On commence par construire une transformée en ondelettes couvrant les fréquences audibles. Soit $\psi(t)$ un filtre passe-bande à de fréquence centrale réduite 1 et de largeur de bande $1/Q$. On dilate la transformée de Fourier $\hat{\psi}(\omega)$ de $\psi(t)$ par des facteurs de résolution $\lambda_1 = 2^{j_1+\chi_1}$ où $j_1 \in \mathbb{Z}$ et $\chi_1 \in \{1 \dots Q\}$:

$$\widehat{\psi}_{\lambda_1}(\omega) = \widehat{\psi}({\lambda_1}^{-1}\omega), \text{ soit } \psi_{\lambda_1}(t) = \lambda_1 \psi(\lambda_1 t). \quad (1)$$

Chaque $\psi_{\lambda_1}(t)$ est un filtre passe-bande de fréquence centrale λ_1, de largeur de bande λ_1/Q et de support temporel $2Q/\lambda_1$. Son facteur de qualité, défini comme le rapport de la fréquence centrale sur la largeur de bande, reste égal à Q. On construit donc un banc de filtres à Q constant, capable de produire une représentation temps-fréquence stable et parcimonieuse [**?**, **?**]. On choisit $Q = 16$ dans les figures de cet article.

On appelle scalogramme de $x(t)$ le module de la transformée en ondelettes $(x * \psi_{\lambda_1})$, indexé par le logarithme en base 2 de la fréquence acoustique λ_1 :

$$x_1(t, \log_2 \lambda_1) = |x * \psi_{\lambda_1}|(t). \quad (2)$$

La transformée à Q constant (CQT) $S_1 x$ correspond à un filtrage passe-bas de x_1 par une fenêtre ϕ_T de support T :

$$S_1 x(t, \log_2 \lambda_1) = x_1 \overset{t}{*} \phi_T = |x \overset{t}{*} \psi_{\lambda_1}| * \phi_T. \quad (3)$$

Ce travail est financé par la bourse ERC InvariantClass 320959. Le code source des expériences et figures est en libre accès a l'adresse www.github.com/lostanlen/scattering.m.

$S_1 x$ est ainsi rendu invariant à toute translation inférieure à T. Toutefois, lors de ce filtrage passe-bas, les modulations d'amplitude dans x_1 de fréquence supérieure à $1/T$ sont détruites. Afin de les restaurer, Andén et Mallat [**?**] ont introduit la transformée de scattering comme le scalogramme du scalogramme :

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |x_1 \overset{t}{*} \psi_{\lambda_2}| = \left| |x \overset{t}{*} \psi_{\lambda_1}| \overset{t}{*} \psi_{\lambda_2} \right|. \quad (4)$$

Les ondelettes $\psi_{\lambda_2}(t)$ ont un facteur de qualité égal à 1, mais nous choisissons de conserver la notation ψ par souci de simplicité. Chaque ondelette $\psi_{\lambda_2}(t)$ a pour fréquence centrale λ_2 et pour support temporel $2/\lambda_2$. Comme dans l'équation (**??**), le filtrage de x_2 par $\phi_T(t)$ crée une représentation $S_2 x$ invariante à la translation jusqu'à T :

$$S_2 x(t, \log_2 \lambda_1, \log_2 \lambda_2) = x_2 \overset{t}{*} \phi = \left| |x \overset{t}{*} \psi_{\lambda_1}| \overset{t}{*} \psi_{\lambda_2} \right| \overset{t}{*} \phi_T. \quad (5)$$

2.2 Transformée jointe temps-échelle

La transformée de scattering définie à l'équation (**??**) décompose chaque bande de fréquence λ_1 indépendamment, et ne peut donc pas capturer la cohérence de structures sonores temps-fréquence, telle que les variations de hauteur. Pour y remédier, Andén [**?**] a redéfini les ondelettes ψ_{λ_2}'s comme des fonctions du temps et de la log-fréquence, indexées par la paire $\lambda_2 = (\alpha, \beta)$, où α est une fréquence de modulation en Hertz et β est une fréquence sur les déplacements en log-fréquence :

$$\psi_{\lambda_2}(t, \log_2 \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log_2 \lambda_1). \quad (6)$$

La variable β est mesurée en cycles par octave ; elle peut prendre des valeurs positives ou négatives, ce qui permet de représenter des changements de hauteur montants ou descendants. Le support temporel de ψ_{λ_2} est maintenant $2/\alpha$, tandis que son support log-fréquentiel est $2/\beta$. On note $\overset{\chi_1}{*}$ les convolutions selon l'axe log-fréquentiel, c'est-à-dire le long des chromas. La transformée en scattering est étendu au cadre « joint » temps-échelle en remplaçant ψ_{λ_2} par $(\psi_\alpha \times \psi_\beta)$ dans l'équation (**??**) :

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |x_1 * \psi_{\lambda_2}(t, \log_2 \lambda_1)|. \quad (7)$$

Le modèle joint temps-fréquence correspond à la transformée « corticale » introduite par Shamma [**?**] afin de formaliser ses découvertes en neurologie de l'audition.

2.3 Transformée sur la spirale

La transformée jointe temps-échelle décrit la variabilité temporelle de hauteur sans recourir à une segmentation préalable. Cependant, elle est agnostique à la structure harmonique des sons voisés. L'évolution de cette structure recèle de l'information sur les formants en parole, ou sur les attaques instrumentales en musique par exemple. On peut la mesurer en comparant des partiels voisins sur des échelles en log-fréquence allant de une à quatre octaves, et ce à chroma fixé. Nous proposons donc d'étendre la transformée jointe temps-fréquence afin d'incorporer les déplacements sur les octaves en conjonction avec les déplacements sur les log-fréquences voisines. Conceptuellement, cela revient à enrouler la variable de log-fréquence $\log_2 \lambda_1$ selon la spirale des hauteurs (voir figure 2) : on révèle ainsi la variable d'octave $\lfloor \log_2 \lambda_1 \rfloor$ (partie entière de $\log_2 \lambda_1$) et la variable de chroma $\{\log_2 \lambda_1\}$ (partie fractionnaire de $\log_2 \lambda_1$). Nous proposons donc d'étendre la transformée jointe temps-fréquence afin d'incorporer les déplacements sur les octaves en conjonction avec les déplacements sur les log-fréquences voisines. Pour ce faire, on enroule la variable de log-fréquence $\log_2 \lambda_1$ selon la spirale des hauteurs (voir figure 2) : on révèle ainsi la variable d'octave $\lfloor log_2 \lambda_1 \rfloor$ (partie entière de $\log_2 \lambda_1$) et la variable de chroma $\{log_2 \lambda_1\}$ (partie fractionnaire de $\log_2 \lambda_1$). En suivant le même procédé que dans les deux transformées définies auparavant, on commence par définir une ondelette ψ_{λ_2} comme un produit séparable d'ondelettes sur chacune des variables à transformer. Dans cet article, on a choisi une ondelette gammatone (profil asymétrique) selon le temps, une ondelette de Morlet (profil symétrique) selon les chromas et une ondelette gammatone selon les octaves.

$$\psi_{\lambda_2}(t, \log \lambda_1, \lfloor \log \lambda_1 \rfloor) = \psi_\alpha(t)\, \psi_\beta(\log \lambda_1)\, \psi_\gamma(\lfloor \log \lambda_1 \rfloor). \quad (8)$$

La figure 1 illustre la structure géométrique de l'ondelette en spirale ψ_{λ_2} dans le plan $(t, \log_2 \lambda_1)$, pour différentes valeurs de $\lambda_2 = (\alpha, \beta, \gamma)$. Nous définissons la transformée en spirale comme une convolution séparable entre le scalogramme et ψ_{λ_2}, selon les trois variables de temps t, log-fréquence $\log_2 \lambda_1$ et octave $\lfloor \log_2 \lambda_1 \rfloor$:

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1 * \psi_{\lambda_2}(t, \log \lambda_1, \lfloor \log \lambda_1 \rfloor)|. \quad (9)$$

Il se trouve que l'idée consistant à enrouler les hauteurs en spirale est bien connue en théorie de la musique, ne serait-ce que par la circularité des noms de notes. Elle a notamment été étudiée par Shepard et Risset pour construire des paradoxes de hauteurs [**?**] et a été validée par des imageries fonctionnelles du cortex auditif [**?**].

3 Déformations du modèle source-filtre

Un modèle de production sonore classique consiste en la convolution d'un signal de source glottique $e(t)$ avec un filtre de conduit vocal $h(t)$. Dans cette section, on introduit une variabilité de hauteur et d'enveloppe spectrale par des déformations temporelles de e et h. On montre comment les propriétés d'harmonicité de $e(t)$ et de régularité spectrale de $h(t)$ permettent de séparer et linéariser ces deux vitesses de déformation.

3.1 Résultat principal

Soit $\sum_n \delta(t - 2\pi n)$ un signal harmonique « source » et soit $t \mapsto \theta(t)$ un difféomorphisme du temps ; on définit $e_\theta(t) = (e \circ \theta)(t)$ la source déformée. De même, on compose un « filtre » $h(t)$ et un difféomorphisme $t \mapsto \eta(t)$ pour définir $h_\eta(t) = (h \circ$

$\eta)(t)$. Le modèle source-filtre déformé est le signal $x_{\theta,\eta}(t) = (e_\theta * h_\eta)(t)$.

(a) $\|\ddot{\eta}/\dot{\eta}\|_\infty \ll \lambda_1\eta/Q$ (filtre lentement variable),

(b) $\|\dot{\hat{h}}/\hat{h}\|_\infty \|1/\dot{\eta}\|_\infty \ll Q\eta^{-1}/\lambda_1$ (réponse fréquentielle régulière),

(c) $\|\ddot{\theta}/\dot{\theta}\|_\infty \ll \lambda_1\eta/Q$ (source lentement variable) et

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

$$|x_{\theta,\eta} * \psi_{\lambda_1}|(t) \approx \widehat{\psi}_{\lambda_1}(p\dot{\theta}(t)) \times \hat{h}\left(\frac{\lambda_1}{\dot{\eta}(t)}\right), \qquad (10)$$

où l'entier p est tel que $p\dot{\theta}(t) \approx \lambda_1$.

On peut calculer explicitement la réponse de source dans le cas d'un spectre harmonique :

$$E(\log_2 \lambda_1) = \sum_{k=1}^{K} \delta(\log_2(\lambda_1) - \log_2(k\xi\eta^{-1})). \qquad (11)$$

Soit $n \in \mathbb{N}$; pourvu que $\lambda_1 = k\xi\eta^{-1}$ soit tel que $k < 2^{-n}K$, on retrouve un partiel n octaves exactement au-dessus de la fréquence λ_1 : d'où $E(\log_2 \lambda_1 + n) = E(\log_2 \lambda_1)$. Par ailleurs, les hypothèses (b) et (c) permettent d'écrire $H(\log_2 \lambda_1) \approx H(\log_2 \lambda_1 + \Delta)$ pour toute déviation de chroma Δ relative à $\log_2 \dot{\theta}$. Ce résultat suggère qu'il est possible de séparer les fonctions $\log_2 \dot{\theta}(t)$ et $\log_2 \dot{\eta}(t)$ en décomposant leurs trajectoires sur les couples de variables temps-chroma et temps-octave. La figure 3 présente un exemple de signal de parole dont chaque syllabe peut être modélisée par un source-filtre déformé. On constate que la transformée de scattering en spirale parvient correctement à distinguer les deux syllabes d'après leurs vitesses de déformation respectives.

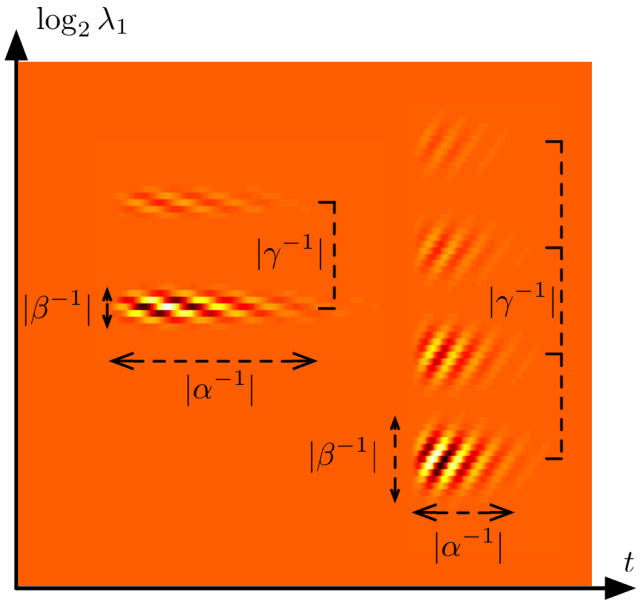

FIGURE 1 – Deux ondelettes en spirale ψ_{λ_2} étalées sur le plan temps-fréquence, présentant des $\lambda_2 = (\alpha, \beta, \gamma)$ différents et une localisation différente sur le scalogramme. À gauche : $\alpha^{-1} = 120\,\text{ms}$, $\beta^{-1} = -0.25\,\text{octave}$, $\gamma^{-1} = +2$ octaves. À droite : $\alpha^{-1} = 60\,\text{ms}$, $\beta^{-1} = +0.5\,\text{octave}$, $\gamma^{-1} = -4$ octaves. On a affiché la partie réelle des coefficients. Le noir correspond à des coefficients positifs et le blanc à des coefficients négatifs.

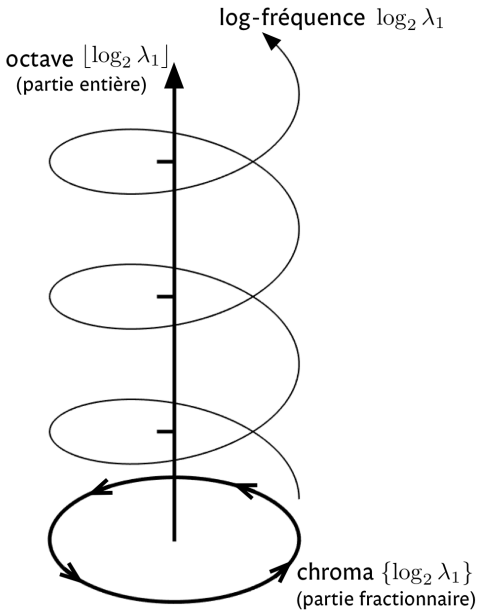

FIGURE 2 – Spirale des log-fréquences

Références

[1] J. Andén. Time and Frequency Scattering for Audio Classification. Thèse de doctorat, École Polytechnique, 2014.

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

[3] 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.

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

[5] 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.

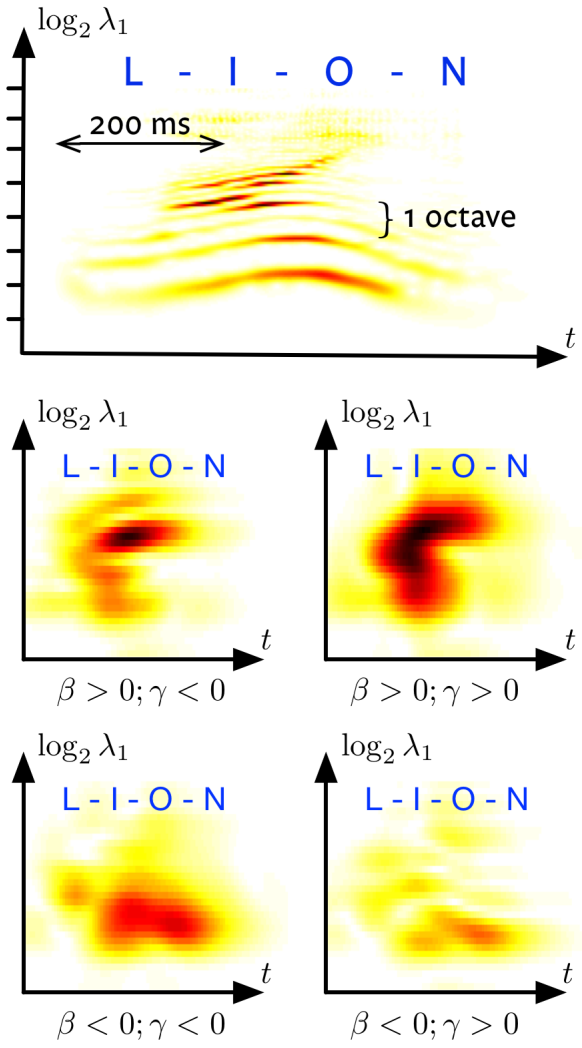

FIGURE 3 – En haut, un scalogramme $x_1(t, \log_2 \lambda_1)$ du mot anglais *lion* (prononcé /ˈlaɪən/). En bas, coefficients de scattering de $x_2(t, \log_2 \lambda_1, \log_2 \lambda_2)$ en fonction du temps t et de la log-fréquence $\log_2 \lambda_1$, pour $\lambda_2 = (\alpha, \beta, \gamma)$ fixé avec $\alpha^{-1} = 120\,\text{ms}$, $\beta^{-1} = \pm 1\,\text{octave}$, $\gamma^{-1} = \pm 4\,\text{octaves}$. On constate que la syllabe /ˈlaɪ/ active en particulier les coefficients tels que $\beta > 0$, $\gamma > 0$ (hauteur montante, timbre montant) tandis que /ɪən/ active les coefficients tels que $\beta < 0$, $\gamma < 0$ (hauteur descendante, timbre descendant). Ces signes sont corrélés avec les sens de déformations du modèle source-filtre : $\ddot{\theta}(t) < 0$ et $\ddot{\eta}(t) < 0$ pour la syllabe /ˈlaɪ/, $\ddot{\theta}(t) > 0$ et $\ddot{\eta}(t) > 0$ pour la syllabe /ɪən/. La clarté est inversement proportionelle à l'amplitude des coefficients.