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 sons, y compris lorsque celle-ci est sujette à des variations d'amplitude, de hauteur et de timbre. Cette idée nous conduira à réorganiser l'axe fréquentiel en spirale, de sorte que des partiels sur des octaves consécutives se trouvent alignés.

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.

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 valeurs complexes, de fréquence centrale réduite 1 et de largeur de bande $1/Q$. L'analyse en ondelettes consiste à dilater la transformée de Fourier $\hat{\psi}(\omega)$ de $\psi(t)$ par des facteurs de résolution $\lambda_1 > 0$:

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

La variable λ_1 est homogène à une fréquence en Hertz. Ainsi, chaque ondelette $\psi_{\lambda_1}(t)$ est un filtre passe-bande de fréquence centrale λ_1, de largeur de bande λ_1/Q et de support temporel Q/λ_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 à la fois sparse et stable [2, 3]. On choisit $Q = 16$ dans les figures de cet article.

On appelle scalogramme le module de la transformée en ondelettes résultant de la construction du banc de filtres ψ_{λ_1}, indexé par le logarithme en base 2 de la fréquence. Le scalogramme de la figure 3 illustre la richesse de l'information géométrique transitoire révélée par cette opération, information qu'il s'agit de capturer sur une échelle d'environ $200\,\mathrm{ms}$ par l'application d'une seconde transformée.

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

Puisque les mêmes exigences de stabilité aux déformations se reportent au scalogramme x_1, on peut envisager d'itérer l'étape précédente, c'est-à-dire d'appliquer une nouvelle transformée

en ondelettes sur x_1 lui-même ; c'est ainsi qu'Andén et Mallat [2] ont introduit la transformée de scattering pour les sons, comme le « scalogramme du scalogramme ». Pour toute fréquence λ_1, il s'agit de transformer $x_1(t, \log_2 \lambda_1)$ par une convolution avec des ondelettes $\psi_{\lambda_2}(t)$ le long de la variable de temps t :

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |x_1(., \log_2 \lambda_1) * \psi_{\lambda_2}(t)|. \quad (3)$$

L'index λ_2 est la fréquence centrale relativement à t. L'échelle de cette ondelette, qui donne la taille de son support, est donc $|\lambda_2|^{-1}$.

Le tenseur obtenu x_2, qui quantifie explicitement les modulations d'amplitude du signal pour des périodes atteignant 200 ms, est notablement plus performant sur des problèmes de classification de genre musical ou de reconnaissance de phonèmes [2] que les représentations issues du scalogramme moyenné, telles que les descripteurs cepstraux (MFCC).

2.2 Transformée jointe temps-fréquence

Il faut toutefois remarquer que la transformée de scattering définie à l'équation précédente n'est pas stable aux mouvements de hauteur au-delà de Q^{-1}, soit un seizième d'octave. Pour stabiliser la représentation x_2 relativement à ces variations, on construit une ondelette bidimensionnelle $\psi_{\lambda_2}(t, \log_2 \lambda_1)$, fonction du temps et de la log-fréquence [1]. Cette ondelette est indexée par $\lambda_2 = (\alpha, \beta)$, où α est une variable de fréquence relativement a un déplacement en temps, et β est une variable cepstrale de « quéfrence », c'est-à-dire de fréquence sur les déplacements en log-fréquence :

$$\psi_{\lambda_2}(t, \log_2 \lambda_1) = \psi_\alpha(t)\, \psi_\beta(\log_2 \lambda_1). \quad (4)$$

La quéfrence β peut prendre des valeurs positives ou négatives, ce qui permet de détecter des changements de hauteur montants ou descendants. Le support temporel de ψ_{λ_2} est proportionnel à $|\alpha|^{-1}$, tandis que son support log-fréquentiel est proportionel à $|\beta|^{-1}$. Afin de préserver cette information de signe pour la paire $\lambda_2 = (\alpha, \beta)$, on choisit de noter

$$\log_2 \lambda_2 = (\log_2 \alpha, \log_2 |\beta|, \mathrm{signe}(\beta)) \quad (5)$$

l'indice associé à l'ondelette bidimensionnelle $\psi_{\lambda_2}(t, \log_2 \lambda_1)$. La transformée de scattering jointe en temps-fréquence x_2 se redéfinit comme une convolution bidimensionelle dans le plan $(t, \log_2 \lambda_1)$ avec cette nouvelle ondelette [1] :

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |x_1 * \psi_{\lambda_2}(t, \log_2 \lambda_1)|. \quad (6)$$

Le modèle joint temps-fréquence correspond à la transformée « corticale » introduite par Shamma [3] afin de formaliser ses découvertes en neurologie de l'audition.

2.3 Transformée sur la spirale

Bien que le modèle précédent soit efficace pour modéliser la variabilité de la hauteur en fonction du temps, il reste agnostique à la structure harmonique du signal, omniprésente dans les sons naturels. L'évolution temporelle de cette structure recèle de l'information sur les formants en parole, ou sur les attaques instrumentales en musique par exemple. On peut mesurer cette évolution en comparant des partiels voisins sur des échelles allant de une à quatre octaves. 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.

$$\psi_{\lambda_2}(t, \log \lambda_1, \lfloor \log \lambda_1 \rfloor) = \psi_\alpha(t)\, \psi_\beta(\log \lambda_1)\, \psi_\gamma(\lfloor \log \lambda_1 \rfloor). \quad (7)$$

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 $\lfloor \log_2 \lambda_1 \rfloor$:

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1 * \psi_{\lambda_2}(t, \log \lambda_1, \lfloor \log \lambda_1 \rfloor)|. \quad (8)$$

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 [4] et a été validée par des imageries fonctionnelles du cortex auditif [5].

3 Déformations du modèle source-filtre

Soit $\sum_n \delta(t - 2\pi\xi^{-1}n)$ 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 $x(t) = [e_\theta * h_\eta](t)$. Dans cette section, on note η la fréquence centrale en Hertz de l'ondelette $\psi(t)$, de sorte que la résolution λ_1 est maintenant une grandeur sans dimension.

Lemme. *Pour tout λ_1 tel que*

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

(b) $\|\dot{\hat{h}}/\hat{h}\|_\infty \|1/\dot{\eta}\|_\infty \ll Q\eta^{-1}/\lambda_1$ *(profil spectral régulier),*

la transformée en ondelettes $[h_\eta * \psi_{\lambda_1}]$ se factorise en

$$[h_\eta * \psi_{\lambda_1}](t) \approx H(\log_2 \lambda_1 - \log_2 \dot\eta(t))\psi_{\lambda_1}\left(\frac{\eta(t)}{\dot\eta(t)}\right) \quad (9)$$

où l'on a défini la réponse du filtre $H(\log_2 \lambda_1) = \lambda_1 \hat h(\lambda_1 \eta)$ *sur un axe log-fréquentiel.*

Démonstration. Grâce à la première hypothèse, on développe $\eta(t-u) \approx \eta(t) - \dot\eta(t) \times u$ sur le support de $\psi_{\lambda_1}(t)$. Le changement de variable $u' = \dot\eta(t) \times u$ conduit à

$$[h_\eta * \psi_{\lambda_1}](t) = \int_{\mathbb{R}} h(\eta(t) - u')\psi_{\lambda_1}\left(\frac{u'}{\dot\eta(t)}\right)\frac{\mathrm{d}u'}{\dot\eta(t)}. \quad (10)$$

L'ondelette ψ_{λ_1} vérifiant $\psi_{\lambda_1}(\dot\eta(t)^{-1}u') = \dot\eta(t)\psi_{\dot\eta(t)^{-1}\lambda_1}(u')$, on peut convertir le facteur de dilatation $\dot\eta(t)$ en une transposition fréquentielle. D'où $[h_\eta * \psi_{\lambda_1}](t) = [h * \psi_{\dot\eta(t)^{-1}\lambda_1}](t)$, ce qui s'écrit comme un produit dans le domaine de Fourier :

$$[h_\eta * \psi_{\lambda_1}](t) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat h(\omega)\hat\psi_{\dot\eta(t)^{-1}\lambda_1}(\omega)\exp(\mathrm{i}\omega\eta(t))\,\mathrm{d}u'. \quad (11)$$

Grâce à la seconde hypothèse, on approxime localement $\hat h(\omega)$ par la constante $\hat h(\dot\eta(t)^{-1}\lambda_1\eta)$ sur le support fréquentiel de $\hat\psi_{\dot\eta(t)^{-1}\lambda_1}$. Dès lors, l'intégrale ci-dessus peut être vue comme la transformée de Fourier inverse de $\hat\psi_{\dot\eta(t)^{-1}\lambda_1}(\omega)$ évaluée en $\eta(t)$. On conclut en revenant à l'ondelette ψ_{λ_1} avec l'équation $\dot\eta(t)^{-1}\psi_{\dot\eta(t)^{-1}\lambda_1}(\eta(t)) = \psi_{\lambda_1}(\eta(t)/\dot\eta(t))$. □

Théorème. *Soit* λ_1 *de la forme* $k\xi\eta^{-1}$*, avec* $k \le K$*. Si les conditions suivantes sont vérifiées :*

(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),*

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

$$|e_\theta * h_\eta * \psi_{\lambda_1}|(t) \approx E(\log_2 \lambda_1 - \log_2 \dot\theta(t))H(\log_2 \lambda_1 - \log_2 \dot\eta(t)) \quad (12)$$

est localement séparable en une réponse de source $E(\log_2 \lambda_1) = |\widehat{\psi_{\lambda_1}}(k\xi)|$ *et une réponse de filtre* $H(\log_2 \lambda_1) = \lambda_1 \hat h(\lambda_1\eta)$*, chacune en mouvement rigide sur l'axe log-fréquentiel* $\log_2 \lambda_1$ *; le mouvement de* E *(resp.* H*) étant régi par le signal* $\log_2 \dot\theta(t)$ *(resp.* $\log_2 \dot\eta(t)$*).*

Démonstration. On part des hypothèses (a) et (b) pour affirmer le lemme précédent. Comme dans la preuve du lemme, on pose $u' = \dot\theta(t) \times (\frac{\eta(t)}{\dot\eta(t)} + u - t)$, on développe et simplifie $\frac{\eta(u)}{\dot\eta(u)} \approx \frac{u'}{\dot\theta(t)}$, et l'on convertit la dilatation temporelle en transposition fréquentielle avec l'équation $\dot\theta(t)^{-1}\psi_{\lambda_1}(\dot\theta(t)^{-1}u') = \psi_{\dot\theta(t)^{-1}\lambda_1}(u')$:

$$\begin{aligned} &\textstyle\int_{\mathbb{R}} e_\theta(t-u)\psi_{\lambda_1}\left(\frac{\eta(u)}{\dot\eta(u)}\right)\mathrm{d}u \\ &= \textstyle\int_{\mathbb{R}} e_\theta\left(\frac{\eta(t)}{\dot\eta(t)} - \frac{u'}{\dot\theta(t)}\right)\psi_{\dot\theta(t)^{-1}\lambda_1}(u')\,\mathrm{d}u' \end{aligned} \quad (13)$$

Avec l'hypothèse (c), on linéarise le difféomorphisme θ autour de $\frac{\eta(t)}{\dot\eta(t)}$, ce qui permet de voir l'intégrale ci-dessus comme la convolution $[e * \psi_{\dot\theta(t)^{-1}\lambda_1}]$ évaluée en $\theta(\frac{\eta(t)}{\dot\eta(t)})$. Puisque le banc de filtres a un facteur de qualité constant Q, la largeur de bande à la fréquence $k\xi\dot\theta(t)$ est $k\xi\dot\theta(t)Q^{-1}$. L'hypothèse (d) peut se réécrire $(k+1)\xi\dot\theta(t) > k\xi\dot\theta(t) + \frac{k\xi\dot\theta(t)}{2Q}$; autrement dit, le $(k+1)^{\text{ème}}$ partiel est hors de la bande passante de $\psi_{\dot\theta(t)\lambda_1}$. 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 $|e * \psi_{\dot\theta(t)^{-1}\lambda_1}|(t)$ se résume au seul terme $E(\log_2 \lambda_1 - \log_2 \dot\theta(t))$ où l'on a défini $E(\log_2 \lambda_1) = |\widehat{\psi_{\lambda_1}}(k\xi)|$ sur un axe log-fréquentiel. □

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

$$E(\log_2 \lambda_1) = \sum_{k=1}^{K} \delta(\log_2(\lambda_1) - \log_2(k\xi\eta^{-1})). \quad (14)$$

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.

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.

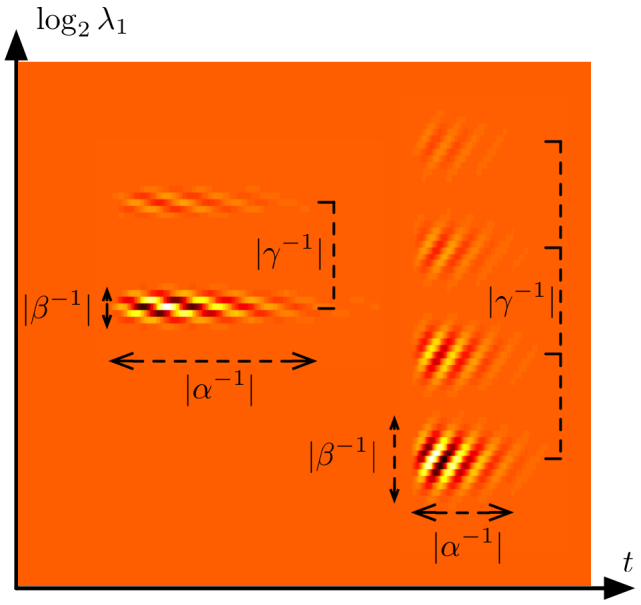

FIGURE 1 – Deux ondelettes en spirale ψ_{λ_2} étalées sur le plan temps-fréquence, présentant des $\lambda_2 = (\alpha, \beta, \gamma)$ différents et une localisation différente sur le scalogramme. À gauche : $\alpha^{-1} = 120\,\text{ms}$, $\beta^{-1} = -0.25\,\text{octave}$, $\gamma^{-1} = +2\,\text{octaves}$.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

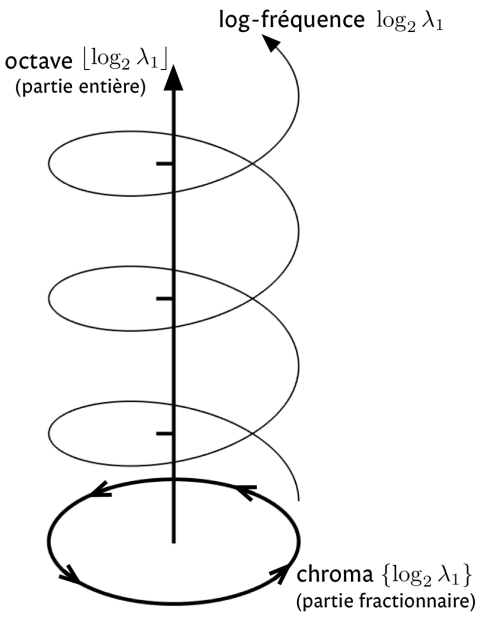

FIGURE 2 – Spirale des log-fréquences

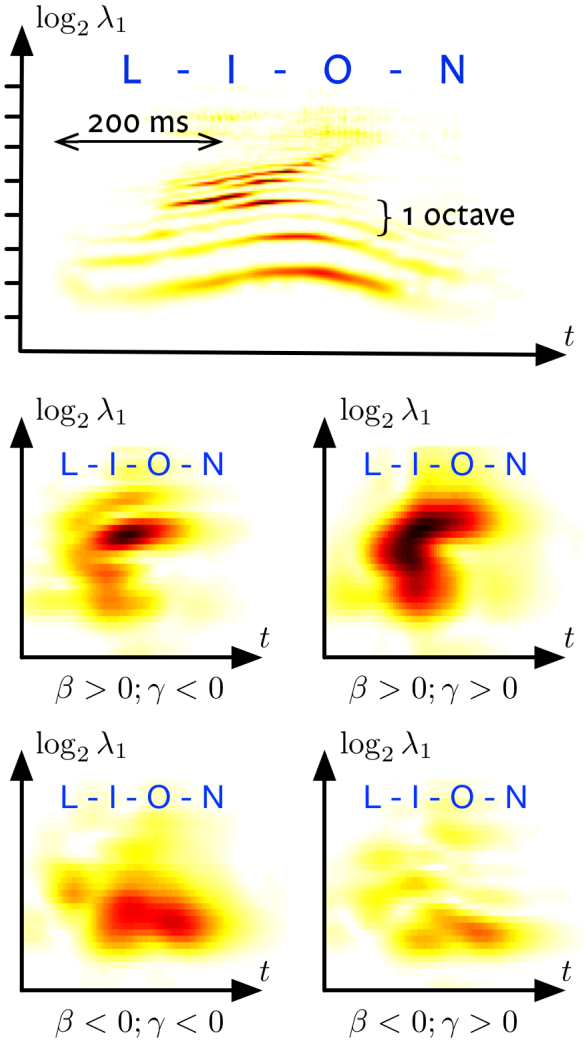

FIGURE 3 – En haut, un scalogramme $x_1(t, \log_2 \lambda_1)$ du mot anglais *lion* (prononcé /'laɪən/). En bas, coefficients de scattering de $x_2(t, \log_2 \lambda_1, \log_2 \lambda_2)$ en fonction du temps t et de la log-fréquence $\log_2 \lambda_1$, pour $\lambda_2 = (\alpha, \beta, \gamma)$ fixé avec $\alpha^{-1} = 120\,\text{ms}$, $\beta^{-1} = \pm 1\,\text{octave}$, $\gamma^{-1} = \pm 4\,\text{octaves}$. On constate que la syllabe /'laɪ/ 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 /'laɪ/, $\ddot{\theta}(t) > 0$ et $\ddot{\eta}(t) > 0$ pour la syllabe /ɪən/. La clarté est inversement proportionelle à l'amplitude des coefficients.