Modélisation long terme de signaux sonores

Mathieu Lagrange

MATHIEU LAGRANGE

MODÉLISATION LONG TERME DE SIGNAUX SONORES

HABILITATION À DIRIGER LES RECHERCHES SOUTENUE LE 14 NOVEMBRE 2019,
À L'UNIVERSITÉ DE NANTES DEVANT UN JURY COMPOSÉ DE :

- **FRÉDÉRIC BIMBOT**, DIRECTEUR DE RECHERCHE CNRS, IRISA, RENNES
- **ALAIN DE CHEVEIGNÉ**, DIRECTEUR DE RECHERCHE CNRS, ENS PARIS
- **BÉATRICE DAILLE**, PROFESSEUR, LS2N, NANTES (PRÉSIDENTE DU JURY)
- **PATRICK FLANDRIN**, DIRECTEUR DE RECHERCHE CNRS, ENS LYON
- **STÉPHANE MALLAT**, PROFESSEUR, COLLÈGE DE FRANCE

« L'ART EST NOTRE FAÇON DE DÉCORER L'ESPACE ; LA MUSIQUE, NOTRE FAÇON DE DÉCORER LE TEMPS. »

ALEX CLAY HUTCHINGS

« BEAUCOUP VOYAGERONT ET LA CONNAISSANCE SERA AUGMENTÉE. »

LIVRE DE DANIEL (12, 4)

https://github.com/mathieulagrange/hdrthesis

February 2020

Table des matières

Notice de lecture

Ce document présente de manière synthétique mes contributions, résultats de plus de dix huit années de recherche en modélisation du signal sonore. Il est composé de trois chapitres qui peuvent être lus de manière relativement indépendante. Les lecteurs non familiers avec le traitement du signal audio numérique pourront bénéficier de quelques notions fondamentales exposées dans le chapitre dédié à la modélisation long terme de signaux sonores pour mieux s'approprier le propos du chapitre dédié à la présentation de mon parcours thématique.

Le premier chapitre s'attache à présenter de manière synthétique mon parcours thématique de l'analyse computationnelle de scènes sonores à leur synthèse pour l'expérimentation en écoute artificielle et l'étude de l'impact sur la perception humaine de l'exposition à ce type de stimuli.

Le second chapitre est dédié à une présentation détaillée de la modélisation long terme de signaux sonores. J'y dresse un panorama de l'évolution de l'effort de recherche déployé par la communauté dans ce domaine durant ces dernières décennies, en détaillant certaines contributions que j'ai pu y apporter et en mettant l'accent sur l'identification des verrous majeurs encore existants.

Le troisième et dernier chapitre évoque plus généralement une critique de la méthode scientifique en sciences des données. J'y présente le fruit de mes investigations dans la formalisation et la conception d'outils de conception de protocoles expérimentaux destinés à faciliter la recherche reproductible dans le domaine des sciences des données, ainsi que quelques réflexions concernant la dernière étape de la méthode scientifique : la revue par les pairs.

Je vous souhaite une bonne lecture.

Pour des raisons de lisibilité, je ferai référence à la littérature dans le corps du texte en évoquant le nom du premier auteur de la publication évoquée. La liste des auteurs ainsi que les informations bibliographiques seront systématiquement disponibles à droite, dans cette colonne.

Parcours thématique

Je m'intéresse à l'être humain et en particulier à ses modes de perceptions de l'environnement. Je privilégie la modalité sonore en partie parce qu'elle comporte intrinsèquement un questionnement sur la dimension temporelle et son influence sur notre perception. Le temps est une notion fascinante mais complexe, je me limite donc plus précisément à la notion de causalité : « Je dispose d'un passé, sous forme de mémoires, qui me permet de donner un sens à ce que je perçois ».

Ce sujet d'étude est intrinsèquement multidisciplinaire et investit notamment les disciplines suivantes :

1. neurosciences : étude du système nerveux ;
2. psychologie cognitive : étude des phénomènes psychiques liés aux sensations et aux perceptions,
3. sciences des données : apprentissage, traitement du signal.

Il est bien entendu que les deux première disciplines sont celles auxquelles on pense en premier. Il pourrait donc être opportun, à l'instar de deux estimés collègues, Alain de Cheveigné (directeur de recherche cnrs au laboratoire d'audition de l'École Normale Supérieure) et Jean-Julien Aucouturier (chargé de recherche cnrs à l'ircam), qui, disposant d'une expertise reconnue dans des thématiques relevant des sciences des données, contribuent directement à l'avancée de ces deux thématiques en apportant leur expertise mais également en construisant un savoir et un savoir faire original à l'interface entre ces disciplines et les sciences des données. Malgré cet intérêt partagé pour ces disciplines, j'ai fait le choix de centrer mon effort de recherche en sciences des données, et ce pour plusieurs raisons.

Aujourd'hui incontournable dans de nombreux domaines, la simulation numérique est à mon sens un fantastique outil de compréhension de notre humanité en ce sens que, utilisée en tant que « réplicateur reproductible » de certaines caractéristiques de l'être humain, il nous permet de nous confronter aux limites de notre capacité de modélisation de nos comportements.

La puissance de cet outil ne modifie en rien le processus de questionnement scientifique rythmé par des allers et retours successifs entre processus inductifs (découvertes, avancées techniques, ...) et déductifs (formalisation, théorisation, ...). Il permet simplement d'en accélérer considérablement la cadence.

Cette accélération comporte des risques, notamment celui de entraîner une certaine perte de méthode. En effet, en mettant trop fortement l'accent sur les avancées technologiques possibles au détriment de leur inclusion nécessaire dans un questionnement scientifique qui résistera au temps, il y a un risque majeur de réduire notre niveau de contrôle, pourtant indispensable à la vertu.

Malgré ces éléments qui nous incitent à la vigilance, je reste néanmoins convaincu que la modélisation numérique est un levier important dans le défi de la connaissance de soi.

Γνῶθι σεαυτόν
« Connais toi toi même ».

Candidement armé de ce projet et de ces bonnes intentions, j'ai poursuivi un effort de recherche au sein de plusieurs institutions de recherche et d'une communauté quasi émergente au début de ma carrière dédiée au traitement du signal audio « hors parole » : son musical d'abord puis son environnemental ensuite.

Ces années m'ont permis d'effectuer une transition de la phase d'exploration à la phase de proposition. Cette transition s'est imposée à moi comme nécessaire à la définition d'orientations qui soient intimement motivées par un questionnement et non le résultat d'une affection plus ou moins assumée avec une série de thématiques séduisantes. Se pose alors la question de la recherche d'un équilibre entre isolement et inclusion thématique. C'est un exercice difficile mais néanmoins indispensable pour être à même de maximiser l'impact de mon travail dans la communauté sans en dénaturer les motivations fondatrices.

Dans une volonté de partage d'expérience, je présenterai dans ce chapitre un état des lieux de mes travaux organisé de manière à mettre en lumière l'évolution de mon orientation thématique en modélisation numérique et en traitement du signal sonore.

Analyse computationnelle de scènes auditives

Le système auditif humain reste en grande partie un mystère, même si son organisation physiologique est connue dans ses grandes lignes. Je négligerai volontairement ici les aspects binauraux en considérant le système auditif humain comme mono capteur, ces aspects étant finalement d'une importance assez faible dans notre formidable capacité à inférer une représentation interne adaptée de notre environnement sonore. Je suppose alors que le système auditif humain s'organise en quatre étapes de traitement successives :

1. transfert mécanique (mono directionnel) : tympan, osselets
2. conversion mécanique / électrique : cochlée
3. transfert électrique (bi-directionnel) : éléments spécifique du cerveau moyen
4. traitement : cortex auditif

En plus de cette conversion d'une énergie mécanique vers une énergie électrique ou encore une information analogique à une information digitale, la cochlée opère une décomposition fréquentielle

qui nous permet d'aisément distinguer un son grave d'un son aigu. Le fait que cette décomposition se fasse aussi tôt dans la chaîne de traitement nous indique son importance. En prenant un parti pris évolutionnaire, on peut supposer que la capacité à distinguer les hauteurs a eu un impact déterminant pour la survie. Une large caisse de résonance aura tendance à produire un son grave si elle est mise en vibration. En être alerté rapidement permet d'avoir un avantage certain pour assurer sa survie.

Un autre élément d'importance est que la décomposition se fait sur une échelle logarithmique en fréquence et en amplitude. Le fait que l'amplitude d'un son soit perçue de manière logarithmique est relativement bien compris, ce qui n'est pas le cas de l'axe fréquentiel. Des arguments d'ordre mathématique motivant ce constat seront donnés dans la section dédiée au scattering d'ondelettes.

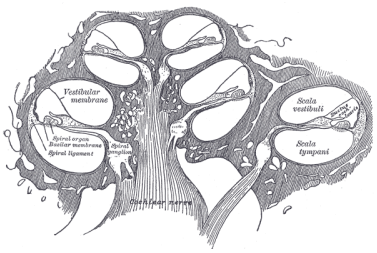


FIGURE 1: Une représentation de la cochlée par Henry Vandyke Carter & Henry Gray (1918) "Anatomy of the Human Body"

Il est important de noter que la communication entre le cortex auditif et la cochlée est bidirectionnelle. De l'information est transmise de la cochlée vers le cortex auditif (direction montante) et du cortex vers la cochlée (direction descendante). Dans le cadre de la psychoperception, la direction montante est relativement aisément étudiable. Un sujet est soumis à un stimulus et doit répondre de manière plus ou moins explicite à des questions sur son expérience. La direction descendante est beaucoup plus difficile à étudier car elle implique de conditionner le sujet et de mesurer l'impact de ce conditionnement sur la performance de l'organe de réception lui même. Cela implique nécessairement une approche « neuroscience » beaucoup plus invasive et complexe à mettre en œuvre comme celle effectuée par Nima Mesgarani avec des êtres humains, montrant l'importance de ces connexions descendantes et leurs capacités à conditionner la sensibilité des couches basses du système auditif. [1]

À la fin du siècle dernier, les outils à disposition étant moins avancés qu'aujourd'hui, on questionnait essentiellement la perception et la cognition humaine par des approches « holistiques » qui étudiaient nos réactions à des stimuli sans pour autant investiguer l'implantation effective dans le cerveau des mécanismes responsables de ces réactions.

Pour le traitement du son, on s'accorde pour considérer que de nombreux éléments structurants sont communément utilisés par le système auditif humain pour inférer une représentation informative de la scène sonore auquel il est soumis. Pour les besoins de l'exemple, on considérera ici que la cochlée peut être fonctionnellement modélisée par une décomposition temps / fréquence et que sa sortie représentée sous forme d'un spectrogramme ou encore d'un scalogramme [2]. Du point de vue physiologique, ces représentations sont une grossière approximation [3] mais elles constituent des références dans le domaine du traitement du signal sonore. C'est donc celles là que nous utiliserons dans ce document.

Les traitements subséquents du système auditif humain peuvent alors être approximés par une opération de segmentation du plan temps / fréquence en des sources données, et ce en fonction de cer-

1. Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233, 2012

2. Les détails de la construction de ces représentations sont disponibles dans la section dédiée au temps / fréquence.

3. Richard F Lyon. *Human and machine hearing*. Cambridge University Press, 2017

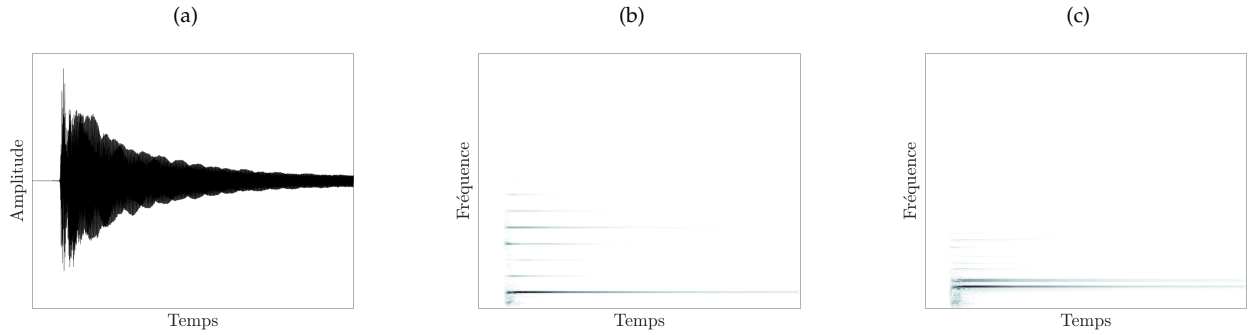
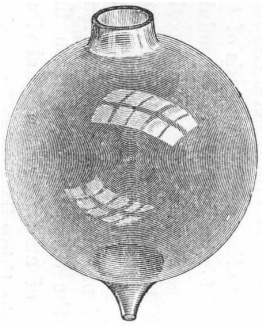


FIGURE 2: De gauche à droite, signal temporel (a), spectrogramme (b) et scalogramme (c) d'une note de piano.

tains critères. On citera, dans l'ordre d'importance communément admis, l'harmonicité, la synchronicité, la proximité en fréquence, en temps, la rythmicité, et bien d'autres. Par exemple, la Figure 2 montre le spectrogramme d'une note de piano. Même si différentes résonances apparaissent comme disjointes sur le plan temps / fréquence nous ne percevons qu'une seule entité sonore. C'est bien qu'il y a eu un processus de formation de source, impliquant majoritairement ici le critère d'harmonicité, *i.e.* des composantes de fortes énergies situées à des fréquences en relation harmonique ont tendance à être perçues comme une seule entité sonore.

De nombreux autres indices ont été étudiés et de nombreux modèles de perception holistiques se basant sur les règles de la Gestalt[4] ont été proposés. En formalisant ces concepts dans une théorie unifiée, appuyée par de nombreuses expériences perceptives, Bregman a fondé l'analyse de scènes auditives (asa).[5] Ces travaux ont suscité un élan d'intérêt dans la communauté traitement du signal et plusieurs modèles computationels ont été proposés comme celui de Dan Ellis,[6] pionnier dans ce domaine communément appelé casa, pour "computational auditory scene analysis". Suivant le principe de formation de sources de l'asa, on peut considérer que le problème casa consiste à attribuer à une des sources d'intérêt chaque point du plan temps / fréquence.



Un « résonateur de Helmholtz ».

4. Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 1935
5. Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994
6. Daniel PW Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Columbia University, 1996

En supposant que l'on a accès aux spectrogrammes des sources composant le mélange, on peut utiliser la technique du masque binaire idéal pour résoudre ce problème. La Figure 3 montre le cas du masque binaire idéal dans le cas d'un mélange de deux sources sonores : deux notes de piano (source harmonique) et une séquence de batterie (source percussive). Dans cet exemple, le masque binaire idéal $M(P,B)$ obtenu à partir des spectrogrammes de piano P et de batterie B est une matrice binaire de même dimension que ces spectrogrammes et se calcule comme suit :

$$M(P,B) = \begin{cases} 1 \text{ si } B(f,t) > P(f,t) \\ 0 \text{ sinon} \end{cases} \quad (1)$$

où f et t sont respectivement les indices de fréquence et de temps dans le spectrogramme. En conséquence, la partie sombre (structure verticale) correspond aux zones du plan temps / fréquence où

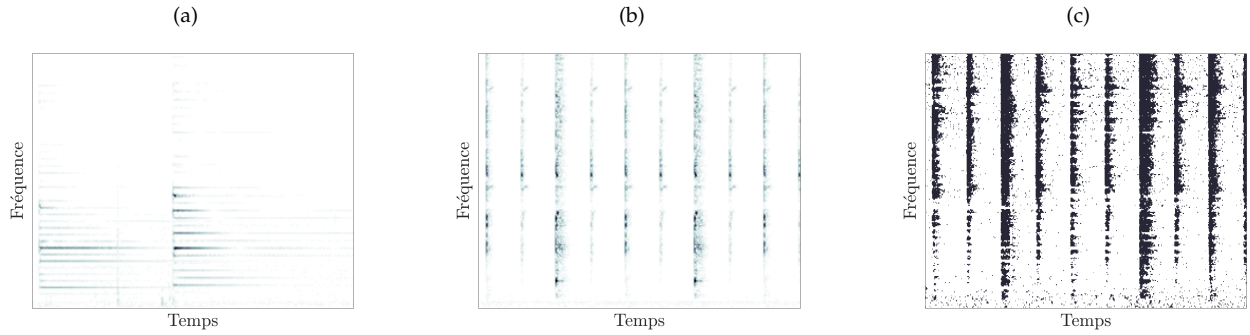


FIGURE 3: Spectrogrammes d'une source harmonique (a), d'une source percussive (b) et masque binaire idéal sélectionnant la source percussive (c).

la source percussive est dominante, la partie blanche (structure horizontale) correspond à la source harmonique.

Ces masques sont dits idéaux parce qu'ils sont estimés à partir de la comparaison avec les spectrogrammes des sources avant mélange. Cette méthode est donc communément utilisé comme référence dite « oracle », car elle nécessite des connaissances qui ne sont pas à disposition du système de séparation que nous cherchons à évaluer. Toute la question est donc d'estimer un masque qui soit le plus proche possible de ce masque idéal.

Séduit par l'aspect neuro-inspiré et la grande diversité algorithmique des approches casa, j'ai investi mon effort de recherche dans ce domaine en proposant plusieurs méthodes de segmentation basées sur des algorithmes de regroupement d'objets, ou "clustering" en anglais.

Soit une scène sonore S composée de plusieurs sources sonores s_m représentée par un ensemble d'atomes temps / fréquence issu d'une décomposition du signal sonore donnée, on cherche alors à regrouper les atomes A en fonction de leur appartenance aux sources s_m. Dans le cas d'une décomposition temps / fréquence comme le spectrogramme, les atomes seront les paniers temps / fréquence $A(f,t)$ du spectrogramme, où t et f dénotent respectivement un indice temporel et fréquentiel. On supposera pour ce faire que la décomposition a été raisonnablement efficace. En effet, dans un objectif de segmentation par regroupement, il est indispensable que les atomes ne soient représentatifs que d'une seule source :

$$\forall A_j, \exists! s_m | A_j \in s_m.$$

Comme on le voit sur la Figure 3, cette condition n'est pas complètement vérifiée dans le cas du spectrogramme car plusieurs composantes temps / fréquence reçoivent de l'énergie provenant des deux sources. On supposera alors que cette contamination est réduite et que cette condition est vérifiée pour un grand nombre d'atomes de forte énergie.

Le problème casa peut donc se formaliser comme une procédure de regroupement d'atomes. Reste à 1) formaliser la procédure du regroupement, et 2) formaliser les critères asa. Ces deux points seront traités dans les deux parties suivantes, avec un niveau de détail plus élevé pour la seconde, cette section étant dédiée à des

problématiques spécifiques au traitement de l'audio.

Algorithmes de partitionnement

Pour formaliser la procédure de regroupement, on supposera maintenant la formalisation des critères asa effectuée, et disponible sous forme d'une notion de dissimilarité entre atomes $d(A_i, A_j)$ telle que :

$$d(A_i, A_j) < d(A_i, A_k) \text{ si } \forall A_i, A_j \in s_m \text{ et } \forall A_k \notin s_m.$$

Le propos des algorithmes de regroupement est de trouver une partition d'un ensemble d'objets (ici des atomes temps / fréquence) en K classes C_k qui minimise la dissimilarité intra classe moyenne :

$$\underset{z_k}{\operatorname{argmin}} \sum_{k=1}^{K} z_k(i,j) d(A_i, A_j) \tag{2}$$

où $z_k(i,j)$ est une fonction indicatrice de l'appartenance des deux atomes i et j à la classe :

$$z_k(i,j) \begin{cases} 1 \text{ si } A_i \text{ et } A_j \in C_k \\ 0 \text{ sinon} \end{cases} \tag{3}$$

En supposant que d procède d'un espace euclidien, l'algorithme des k-moyennes, proposé par Lloyd,[7] constitue la référence pour les problèmes de grande taille en nombre d'objets et d'un nombre de classes petit devant le nombre d'objets. Les relations casa que nous étudierons dans la suite n'étant pas exprimables sous forme d'une distance euclidienne, nous portons notre attention sur des algorithmes alternatifs, adaptés aux espaces non euclidiens.

7. Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982

On peut pour cela formaliser le problème de regroupement comme un problème de partitionnement de graphe, où chaque objet constitue un nœud et chaque dissimilarité pondère un arc entre deux nœuds. L'algorithme des "normalized cuts" permet de résoudre ce problème de partionnement. Cet algorithme a été popularisé dans le domaine de la segmentation d'image,[8] où les pixels de l'image forment les nœuds du graphe et les arcs sont pondérés par proximité sur une échelle de gris avec contrainte de localité spatiale. Comme cela sera détaillé dans la section suivante, nous avons, en adaptant la formulation de ce poids aux critères asa, appliqué avec un certain succès cette approche au problème casa. Néanmoins, pour des problèmes de grande taille (en nombre d'atomes comme en nombre de sources), une alternative à cet algorithme s'est révélée nécessaire. L'algorithme des "normalized cuts" est en effet coûteux, principalement à cause d'une étape de décomposition en valeurs singulières en $\mathcal{O}(n^3)$ d'une version transformée de la matrice de dissimilarité qui est de taille n^2 où n est le nombre d'objets considérés. Dans le cas du problème casa, cela contraint fortement l'horizon temporel exploitable. Par exemple, pour segmenter un spectrogramme composé de 512 bandes fréquentielles calculé avec

8. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000

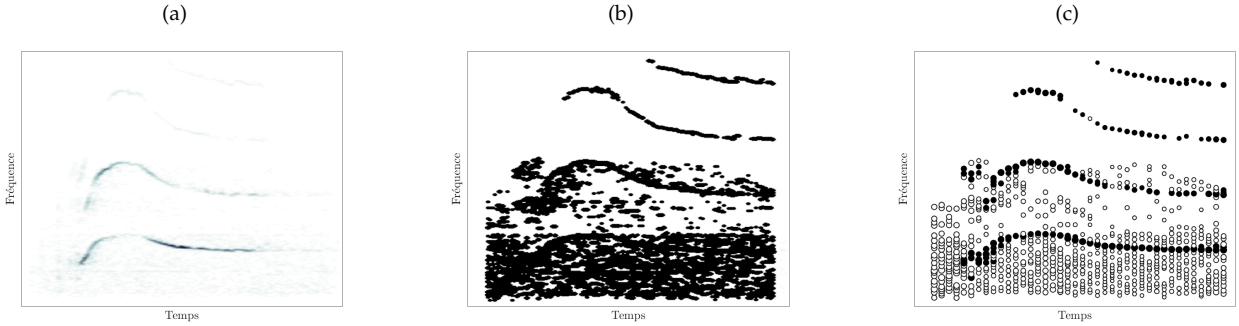


FIGURE 4: Spectrogramme (a), représentation sinusoïdale à court terme (b) et segmentation (c).

un recouvrement de moitié avec un signal échantillonné à 44.1 kHz, la prise en compte d'un horizon temporel d'une seconde nécessitera le traitement de $n = 5.10^5$ atomes. Il nécessite également que la dissimilarité puisse s'exprimer comme un noyau pour garantir sa convergence, *i.e.* que la matrice de dissimilarité soit semi définie positive, ce qui n'est pas le cas en général.

Inderjit Dhillon montre que l'objectif des coupures normalisées peut être formulé comme une fonction de coût optimisable par un algorithme de regroupement à noyau "kernel k-means", ce qui permet d'approcher le problème du regroupement par coupures normalisées avec une complexité considérablement réduite. [9] Il reste que la convergence de ce type d'approche n'est vérifiée que dans le cas où la matrice de dissimilarité est semi définie positive. Il est possible de manipuler une matrice symétrique non semi définie positive pour la rendre semi définie positive [10] mais l'impact de cette manipulation sur le regroupement obtenu est difficile à quantifier. Les relations casa étant complexes à modéliser et ne vérifiant pas en général cette contrainte, il nous a paru important de disposer d'un algorithme de regroupement efficace qui s'affranchisse de cette contrainte.

En collaboration avec Mathias Rossignol, j'ai donc développé un algorithme de regroupement, nommé *k-averages*, dont la convergence est assurée pour tout type de matrice de dissimilarité, pour peu qu'elle soit symétrique, *i.e.* $d(a, b) = d(b, a)$. Cette approche, même si elle considère des relations bien connues, n'avait, à notre connaissance, pas été complètement formalisée en un algorithme effectif. Son évaluation sur un large ensemble de corpus [11] montre des performances équivalentes aux méthodes à noyau, avec un coût de calcul et une empreinte mémoire plus faible, ce qui la rend intéressante pour des problèmes de grande taille. [12]

9. Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11): 1944–1957, 2007

10. Volker Roth, Julian Laub, Motoaki Kawanabe, and Joachim M. Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(12):1540–1551, December 2003

11. Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/

12. Mathias Rossignol, Mathieu Lagrange, and Arshia Cont. Efficient similarity-based data clustering by optimal object to cluster reallocation. *PloS one*, 13(6):e0197450, 2018

Application au problème casa

Dans une première étude menée à l'Université de Victoria en collaboration avec George Tzanetakis, j'ai considéré le problème suivant : comment isoler la partie harmonique dominante d'un enregistrement polyphonique ? Cela peut être le cas de la mélodie dans un enregistrement musical polyphonique ou de vocalisations

de mammifères marins dans un environnement sonore sous-marin bruité. La Figure 4 montre son usage pour une vocalisation d'orque captée grâce à un hydrophone dans les eaux du "Inside passage" au nord de l'île de Vancouver, Colombie Britannique, Canada.

Soit une représentation sinusoïdale à court terme d'un signal sonore où chaque atome A_l^t est paramétrisé par une fréquence f_l^t et une amplitude a_l^t et t et l désignent respectivement l'indice de trame et l'indice de l'atome au sein de la trame [13]. Partant de cette représentation, le problème précité se formule comme suit : comment définir mathématiquement certaines relations asa pour ce type de représentation, les combiner entre elles, et en fonction de ces relations isoler la partie qui se conforme le plus aux critères exprimés ?

La Figure 5 donne le schéma fonctionnel du processus de segmentation proposé [14]. Le module d'analyse sinusoïdal produit pour chaque trame un ensemble d'atomes sinusoïdaux. Ces trames sont ensuite agrégées sur un intervalle de temps correspondant à 150 ms et traitées par l'algorithme de séparation. Les atomes identifiés comme appartenant à la source d'intérêt sont ensuite transmis au module de synthèse.

La contrainte de localité est ici exprimée en temps et de manière implicite par l'utilisation d'une fenêtre de "texture" composée de plusieurs trames. Cette approche permet de considérer un intervalle de temps suffisamment long tout en respectant des contraintes de complexité.

Les indices de proximité locale s'expriment en terme de fréquence et d'amplitude :

$$W_f\left(A_l^t, A_m^p\right) = \mathrm{e}^{-\left(\frac{f_l^t - f_m^p}{\sigma_f}\right)^2} \tag{4}$$

$$W_a\left(A_l^t, A_m^p\right) = \mathrm{e}^{-\left(\frac{a_l^t - a_m^p}{\sigma_a}\right)^2} \tag{5}$$

où σ_f et σ_a permettent de contrôler l'importance relative des deux facteurs dans la mesure de proximité finale.

En revanche la relation d'harmonicité potentiellement entretenue entre deux atomes ne peut s'exprimer de manière robuste d'une manière aussi directe. En effet, le fait que le rapport des fréquences des deux atomes considérés puisse se réduire à un rapport de petits entiers positifs [15] n'est pas une condition suffisante pour exprimer le fait que ces deux atomes appartiennent à une source harmonique.

Nous assignons donc à chaque atome A_l^t un profil harmonique H_l^t composé à partir de l'ensemble des atomes ayant le même indice de trame. Ce profil est ensuite judicieusement manipulé pour maximiser la probabilité d'un bon alignement de ces profils si les deux atomes appartiennent à la même source harmonique [16]. Le critère d'harmonicité s'exprime alors comme une corrélation entre

13. Une représentation sinusoïdale à court terme peut être considérée ici comme une version parcimonieuse et compacte du spectrogramme où l'on a identifié et modélisé les composantes de forte énergie par des sinusoïdes fenêtrées. Le lecteur peut se référer à la section dédiée à la modélisation basée sur des sinusoïdes à court terme pour plus de détails.

14. Son implantation a été diffusée dans l'environnement de traitement du signal audio Marsyas `http://marsyas.info`

15. Tuomas Virtanen and Anssi Klapuri. Separation of harmonic sound sources using sinusoidal modeling. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 2, pages II765–II768. IEEE, 2000

16. M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized Cuts for Predominant Melodic Source Separation. *IEEE Transactions on Acoustics, Speech and Language Processing*, 2008

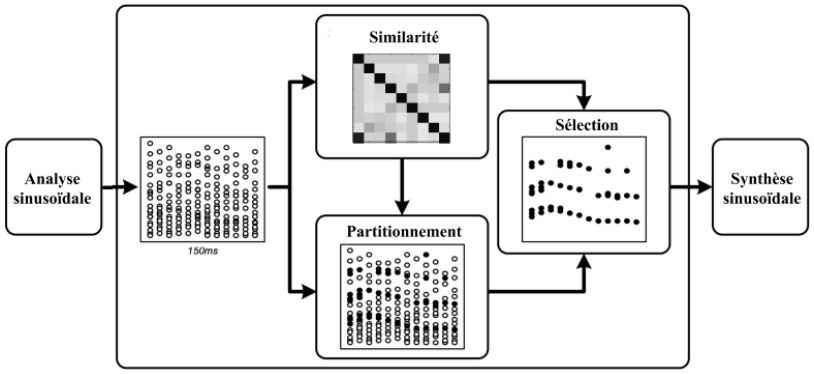


FIGURE 5: Schéma fonctionnel du processus de segmentation basé sur l'algorithme des coupures normalisées. Le module d'analyse sinusoïdal produit pour chaque trame un ensemble d'atomes sinusoïdaux. Ces trames sont ensuite agrégées sur un intervalle de temps correspondant à 150 ms et traitées par l'algorithme de séparation. Les atomes identifiés comme appartenant à la source d'intérêt sont ensuite transmis au module de synthèse.

ces deux profils :

$$W_h\left(A_l^t, A_m^p\right) = \mathrm{e}^{\left(\frac{c\left(H_l^t, H_m^p\right)}{\sqrt{c\left(H_l^t, H_l^t\right)\cdot\left(H_m^p, H_m^p\right)}}\right)^2} \tag{6}$$

où

$$c\left(H_a^b, H_c^d\right) = \sum_i H_a^b(i) \odot H_c^d(i). \tag{7}$$

où $\odot$ désigne le produit de Hadamard. La segmentation effectuée grâce à la combinaison de ces trois critères $W_{afh} = W_a \odot W_f \odot W_h$ permet d'isoler un certain nombre de parties dans chaque fenêtre de texture. La partie la plus "dense" dans l'espace de représentation est ensuite sélectionnée. Soit un ensemble d'atomes d'une même classe

$$C_b = \left\{A_m^t | \,\text{label}\left(A_m^t\right) = b\right\}$$

ce critère de densité s'exprime comme suit :

$$d\left(C_b\right) = \frac{1}{\#C_b^2} \sum_{A_l^t \in C_b} \sum_{a_m^j \in C_b} W_{afh}\left(A_l^t, A_m^j\right). \tag{8}$$

Ce premier système m'a permis d'explorer des critères de type instantanés, c'est-à-dire locaux en temps.

Comme le montrent les études effectuées dans le cadre de l'asa, il est important de considérer également la dimension temporelle dans le processus de structuration. Au sein de l'ircam, et en collaboration avec Mathias Rossignol, je me suis focalisé sur la notion de structure séquentielle, en faisant le postulat suivant : toute segmentation est effectuée sur la base de contraintes de localité dans deux espaces conceptuellement disjoints mais complémentaires. Le premier est un espace analogue à l'espace / temps, et le second est relatif à une notion de forme. Ces deux types d'espaces de représentation induisent respectivement des relations 1) de proximité et de continuité, et 2) de similarité ; leur composition amenant une autre notion, la notion de structure.

Précisons le propos :

1. la continuité signifie que nous identifions de préférence comme des objets des fragments sonores sans rupture de continuité,
2. la similitude signifie qu'en cas de rupture de continuité dans l'observation de l'objet, tous ses composants doivent néanmoins partager des propriétés spectrales communes, puisqu'elles sont produites par une même source,
3. la notion de structure, enfin, reflète la manière dont les auditeurs peuvent reconnaître des objets complexes en identifiant des motifs répétés : par exemple, des sons mécaniques complexes ou des appels d'animaux peuvent être constitués de plusieurs fragments non continus et dissimilaires, mais la répétition de ces formes permet d'en inférer la structure.

L'utilisation conjointe de ces contraintes est particulièrement difficile à formaliser, mais néanmoins indispensable à l'obtention d'une segmentation de qualité. Le schéma fonctionnel de la Figure 6 montre comment ces axes de structuration ont été considérés. Au début de l'algorithme, les éléments de structuration O_{0k} correspond aux trames du spectrogramme. En partant de ce type d'objets élémentaires, il est souhaitable que la notion de proximité temporelle prédomine, car les notions de similarité et *a fortiori* le critère de séquentialité ne peuvent se baser sur une information aussi peu discriminative. Inversement, plus on considère des objets complexes $O_{pk}, p >> 0$, moins la notion de proximité est pertinente et plus la notion de similarité le devient. Forts de ce constat, nous avons donc chercher à concevoir un algorithme de structuration hiérarchique basé sur l'expression de ces critères qui soit adapté à la segmentation temporelle de scènes sonores [17].

L'algorithme proposé effectue pour cela un regroupement hiérarchique ascendant des trames d'un spectrogramme. La première étape ne considère que le critère de continuité, *i.e.* une similarité sous contrainte de localité. Cette étape permet d'avoir des fragments au contenu spectral plus spécifique, un élément important pour obtenir une mesure de séquentialité qui soit pertinente dès le début du processus de structuration. Pour les niveaux suivants, on considère la combinaison d'une mesure de similarité de forme, exprimée comme le coût minimal d'un alignement élastique entre les deux fragments comparés, et d'une mesure de similarité de séquence, calculée à partir d'une quantification des trames associées aux fragments, comme représenté sur la Figure 6. L'algorithme procède comme suit :

1. **Initialisation** par une première classification : regrouper les trames successives qui ont une grande similarité spectrale pour déterminer les fragments de niveau 0
2. **Répéter** pour le nombre spécifié de niveaux :
   - Calcul des similarités entre fragments O_{ik} de niveau i
   - Classifier ces fragments en classes C_i en fonction de ces similarités

17. Mathias Rossignol, Mathieu Lagrange, Grégoire Lafay, and Emmanouil Benetos. Alternate Level Clustering for Drum Transcription. In *European Conference on Signal Processing (EUSIPCO)*, Nice, France, September 2015

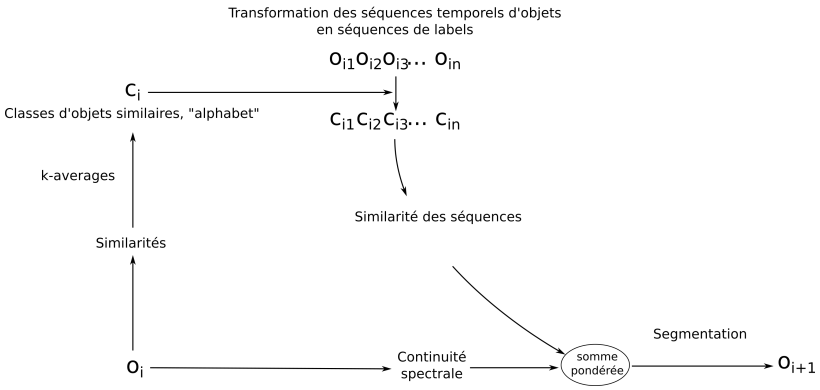


FIGURE 6: Schéma décrivant les principaux traitements utilisés pour passer du niveau i au suivant. Une étape de classification non supervisée des fragments O_{ik} produit des labels C_{ik} qui sont utilisés comme information structurelle. Combinée avec un critère de continuité spectrale, cette information guide la détermination des fragments O_{i+1k} du niveau suivant.

— Calcul de l'information mutuelle entre les O_{ik}, en se basant sur le degré de séquentialité entre C_{ik} et C_{ip} :

$$\mathrm{MI}\left(C_{ik}, C_{ip}\right) = \log\left(\frac{p\left(C_{ik}C_{ip}\right)}{p\left(C_{ik}\right)p\left(C_{ip}\right)}\right)$$

où $p(C)$ désigne la probabilité qu'un fragment appartienne à une classe donnée, et $p(C_{ik}C_{ip})$ est la probabilité que les labels C_{ik} et C_{ip} soient consécutifs.

— Génération d'une courbe de décision le long de l'axe temporel. Cette courbe de décision est fonction de :

  — l'information mutuelle entre les classes des deux fragments consécutifs,

  — la continuité spectrale, définie comme la similarité spectrale entre la fin d'un fragment et le début du fragment suivant.

— Segmenter la séquence de fragments en fonction de la courbe de décision, en combinant les fragments situés entre les minima locaux de cette courbe de décision.

La Figure 7 présente la segmentation de l'introduction du morceau de Deep Purple *Smoke on the Water*. La même phrase musicale est répétée deux fois avec quelques variations d'interprétation. Des motifs émergent à partir du niveau 4 où les motifs élémentaires du riff sont révélés.

Constat

Ces années d'exploration des approches casa m'ont permis de parcourir de larges champs d'expertise scientifiques, de l'informatique au traitement du signal, de la psycho-perception aux neurosciences, mais cela a surtout été pour moi l'occasion de mieux saisir l'importance de certaines problématiques fondamentales en sciences des données :

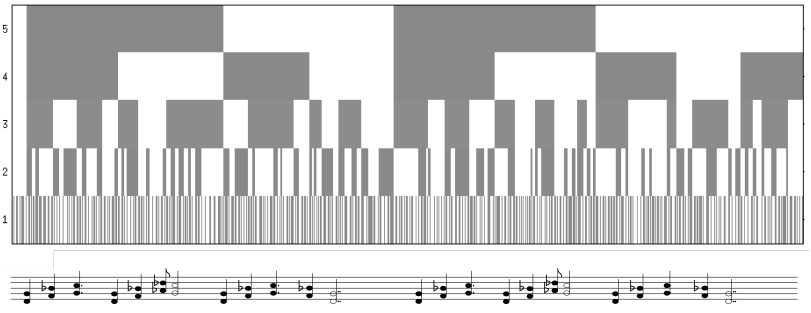


Figure 7: Segmentation du morceau de Deep Purple *Smoke on the Water* (*Machine Head*, Purple Records / EMI, 1972).

1. Comment dévoiler la structure intrinsèque aux données à partir d'observations bruitées ?
2. Comment exploiter le caractère compositionnel de notre compréhension du monde ?
3. Quel rôle peut jouer la notion de mémoire dans les processus pré-cités ?

Ces questions continuent de nourrir mon questionnement, car ces problématiques sont encore largement ouvertes.

La traduction algorithmique de contraintes de structuration est un art difficile, car elle nécessite un effort important pour fiabiliser les algorithmes. La gestion de l'interaction entre les différentes contraintes de structuration l'est aussi.

Enfin, la notion de mémoire est ici exploitée de manière implicite en considérant la redondance d'information dans la scène observée. Je dois faire le constat qu'avec les schémas computationels étudiés, la taille des horizons temporels considérés à l'époque pour rendre l'approche réalisable s'est révélée insuffisante.

Au delà de ces problématiques propres aux approches casa, les approches de séparation de sources se basent généralement sur un traitement en deux temps : 1) représentation du signal audio sous forme de spectrogramme, 2) segmentation. On verra dans le chapitre dédié à la modélisation long terme de signaux sonores, les limites ce que cela impose sur la performance de ces approches.

La synthèse pour l'écoute artificielle

Ces travaux ont également été pour moi l'occasion de réfléchir au protocole d'évaluation de ces systèmes d'écoute artificielle, et en particulier aux données utilisées pour les évaluer. Il est évident que les données, qu'elles soient celles que l'on observe où celles que l'on cherche à prédire, constituent le socle des sciences des données. Cette question est donc d'importance.

Dans le domaine du traitement du signal, on fait souvent la différence entre des signaux dits « réels » et des signaux dits « synthétiques ». Les signaux réels, on parle souvent aussi de données brutes, sont issus d'un processus d'acquisition plus ou moins maîtrisé. Une fois acquises, la seule manipulation que l'on puisse alors

faire pour mieux contrôler ces données, c'est de réduire la taille du corpus. Cette réduction peut se faire selon les différents axes de structuration disponibles. Dans le cas d'un corpus structuré sous forme de classes par exemple, on peut choisir de supprimer une ou plusieurs classes, ou de supprimer des éléments dans chaque classe de manière à obtenir un corpus équilibré, *i.e.* avec un même nombre d'éléments par classe. Elle peut se faire également de manière globale, pour enlever des données aberrantes par exemple. Dans tous les cas, il est important de considérer que ces procédures sont délicates car difficile à motiver. Le risque du "cherry picking", qui consiste à trouver les données qui fonctionnent bien avec l'approche algorithmique défendue est toujours présent.

Je n'évoquerai pas ici la question importante de la représentativité du corpus et de la division du corpus complet en corpus d'entrainement, de test, et de validation. Le lecteur peut se référer à l'excellent cours de Stéphane Mallat sur la notion de risque en apprentissage pour une approche formalisée de ce sujet.
Vidéos : `https://www.college-de-france.fr/site/stephane-mallat`
Notes : `https://www.di.ens.fr/~mallat/CoursCollege.html`.

Les signaux synthétiques sont, eux, totalement contrôlés. Une étude en traitement du signal est canoniquement organisée comme suit : un modèle de signal et une méthode d'estimation des paramètres de ce modèle sont proposés. En utilisant ce modèle de signal pour générer des exemples, on montre que la méthode d'estimation est effective sur ces exemples. Cela permet de faire une démonstration de faisabilité du modèle, mais n'est pas suffisant pour démontrer l'applicabilité de ce modèle à des problématiques concrètes.

La méthode couramment prise est donc d'estimer ce niveau d'applicabilité en considérant les performances obtenues par le système en considérant des données brutes. Le principal risque de donner une foi totale dans les résultats obtenus en utilisant les données brutes est précisément que nous disposons souvent d'un niveau de contrôle très faible sur ces données, et ce qui à permis de les obtenir. Par exemple, le processus d'acquisition pour les données observées, le processus d'annotation pour les données que l'on cherche éventuellement à prédire sont des étapes non triviales où de nombreux choix vont fortement conditionner le potentiel des corpus construits.

Des problèmes se posent également du côté de l'évaluation de l'estimateur. Prenons l'exemple du calcul d'une statistique telle que la moyenne arithmétique : un ensemble d'outils nous permet de valider la pertinence de cette statistique, comme l'étude de la distribution, l'écart à la médiane, les quartiles, etc. Pour un modèle avec plusieurs millions de paramètres libres comme c'est le cas pour les architectures de traitement de l'information telles que les réseaux de neurones profonds, les outils d'inspection sont nettement plus ardus à mettre en place. La valeur de la métrique de qualité obtenue par le système doit donc être appréciée avec beaucoup plus de retenue.

Le modèle proposé est-il pertinent ? Est-ce que l'information captée par la méthode d'estimation se rapporte à des phénomènes d'intérêts ? Pour tout problème qui sort de la trivialité, nous ne disposons pas à ce jour d'outils qui permettent de répondre de manière solide à ces questions. Dans une forme moderne incluant la notion de recherche reproductible comme celle développée dans

la section dédiée à la revue par les pairs, l'inspection par les pairs constitue probablement la seule réponse efficace dont nous disposons à l'heure actuelle. Précisons ce propos avec deux exemples.

La communauté d'acoustique environnementale se base principalement sur l'étude de mesures de niveaux acoustiques. De nombreuses études issues de cette communauté montrent que ces mesures ont leurs limites et qu'il serait utile de pouvoir décrire plus finement l'environnement sonore. [18] Une étude utilisant des techniques d'apprentissage semble montrer qu'il est possible de prédire le type d'environnement sonore urbain (rue, parc, boulevard, ...) avec une excellente précision. [19] Cette étude a été très bien reçue par la communauté et pris pour fait acquis, en partie parce que cette étude appliquait des techniques mal connues par les membres de cette communauté.

Intrigué par ces résultats qui contrastaient avec des résultats obtenus dans d'autres études, j'ai entrepris de répliquer ces travaux. Cette tentative de réplication, effectuée en collaboration avec les auteurs, à été fructueuse et m'a permis de comprendre que les bons résultats obtenus étaient dus à une faille dans le protocole expérimental [20]. Les scènes présentes dans le corpus ont été enregistrées durant une période longue (environ une heure) et découpées en tronçons d'une minute. Etant donné la stationnarité de ce type de scènes sonores en terme de composition, deux tronçons successifs sont la plupart du temps quasiment identiques en terme de contenu. Le protocole expérimental n'empêchant pas que l'algorithme d'apprentissage exploite des tronçons d'une scène donnée pour prédire le label d'un tronçon de cette même scène, le problème s'en trouvait, dans son incarnation numérique, considérablement simplifié.

Prenons un second exemple. Le scattering d'ondelettes, un modèle de signal d'intérêt présenté dans la section dédiée aux représentations temps / fréquence / modulations, a notamment été utilisé dans le cadre de la classification de genre musical [21]. Le scattering est en charge de produire une représentation du signal sonore qui soit informative mais de faible dimensionalité. Cette représentation est ensuite utilisée par un algorithme de classification supervisée chargé de prédire à quel genre musical le morceau analysé appartient parmi une typologie déterminé.

Dans une étude de réplication et d'analyse de la robustesse de ce système de prédiction du genre musical, Rodriguez et Sturm montrent que les bonnes performances de l'approche basée sur le scattering se basent (au moins en partie) sur 1) une base de données mal structurée, et 2) l'utilisation d'information coïncidentes inaudibles pour l'être humain. [22]

La base de données gtzan proposée par George Tzanetakis a été pendant une dizaine d'années la base de donnée de référence sur le sujet, car elle a été conçue pour évaluer le premier système de prédiction du genre musical. [23] Plusieurs études ont montrées les limites de cette base de données notamment en terme de structure

18. Catherine Lavandier and Boris Defréville. The contribution of sound source characteristics in the assessment of urban soundscapes. *Acta acustica united with Acustica*, 92(6):912–921, 2006

19. Jean-Julien Aucouturier, Boris Defreville, and Francois Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2): 881–891, 2007

20. Mathieu Lagrange, Grégoire Lafay, Boris Défréville, and Jean-Julien Aucouturier. The bag-of-frames approach: A not so sufficient model for urban soundscapes. *The Journal of the Acoustical Society of America*, 138(5):487–492, 2015
Ce phénomène est similaire à l'effet « album » en musique. Les conditions d'enregistrements (microphones, mixage, mastering) sont souvent spécifique à l'album enregistré, spécificité que les algorithmes d'apprentissage exploitent pour minimiser leur erreur de prédiction. Pour mitiger cet effet, il convient alors de ne pas permettre d'avoir des chansons d'un même album dans les corpus d'apprentissage et de test.

21. Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16): 4114–4128, 2014

22. Francisco Rodriguez Algarra, Bob L Sturm, Hugo Maruri-Aguilar, et al. Analysing scattering-based music content analysis systems: Where's the music? In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2016

23. George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002

d'annotation et de constitution de partitions d'apprentissage et de test. L'utilisation d'un partitionnement plus rigoureux de la base fait décroitre les performances des méthodes testées de plus de 20 %. Par ailleurs, l'étude montre que des informations présentes dans les bandes de fréquence inférieures à 20 Hz sont déterminantes pour la qualité de la prédiction. En effet, en considérant une version où les bandes de fréquence inférieures à 20 Hz sont fortement atténuées, les performances se réduisent de 15 % environ. L'oreille ne percevant pas les signaux en dessous de 50 Hz, on peut dire que le système d'apprentissage se focalise sur des composantes inaudibles pour l'être humain.

Ces deux exemples de réplication montrent combien il est difficile d'apprécier la performance d'un algorithme de traitement de données, et que, comme cela sera discuté dans le chapitre dédié à la méthode scientifique en sciences des données, la réplication par les pairs constitue un outil privilégié pour construire une connaissance plus solide.

Dans une posture particulièrement sceptique, Bob Sturm désigne tout ces algorithmes comme des potentiels "horses", en référence à "Clever Hans", le cheval intelligent.

Néanmoins, le fait d'utiliser une corrélation fortuite entre différentes informations présentes dans les données, n'est pas nécessairement dommageable. Il est par exemple courant, pour des raisons axiomatiques concernant le type des données, ou encore empiriques, d'optimiser une fonction de coût donnée que l'on suppose corrélée avec l'objectif principal (prédire sur de nouvelles données, séparer des sources, ...). Dans la mesure où ces choix sont assumés, on en connaît mieux les limites. Il est par contre fondamental d'avoir conscience de ces problématiques pour être à même d'apprécier avec un regard suffisamment critique les mesures de performance dérivées de tel ou tel algorithme ou protocole expérimental. [24] Par exemple, une technique classique lors de l'optimisation consiste à procéder à l'"early stopping", c'est-à-dire ne pas optimiser jusqu'à convergence de la fonction de coût optimisée, mais jusqu'à un certain nombre d'itérations amenant au maximum, empiriquement observé, de la fonction objectif.

On a donc vu que l'usage de données synthétiques ne permettait qu'une validation du concept, et que l'usage de données brutes doit être encadré avec soin. En effet, même si le but d'un algorithme de traitement de données est d'être capable *in fine* de traiter des données brutes, le niveau de contrôle sur ce type de données est réduit, ce qui affaiblit considérablement la confiance que l'on peut avoir en ces métriques de performance obtenues. Je n'évoquerai pas ici le problème de l'annotation des données brutes qui se pose également, et qui, dans le cas de certaines tâches peut s'avérer particulièrement problématique. Deux exemples permettront de saisir l'ampleur de ces difficultés :

1. Comment annoter la fin d'un évènement sonore dans une scène sonore polyphonique ?

"Clever Hans" était un cheval supposément capable de compter. Une personne lui donnait une opération mathématique à résoudre, et Hans tapait du sabot au sol et s'arrêtait de taper quand le compte de coups de sabot correspondait à la solution de l'opération. Confronté par un protocole rigoureux, il s'est avéré que Hans n'était capable de donner la bonne réponse que si la personne qui lui posait la question connaissait la réponse. Hans était capable de détecter les subtiles modifications de posture de la personne qui lui avait posé la question à l'approche du bon nombre de coups de sabots. Dire si Hans était intelligent nécessiterait de s'entendre sur ce terme mais la manière dont Hans répondait à la sollicitation n'était pas d'ordre calculatoire.

24. Mathieu Lagrange and Mathias Rossignol. Computational experiments in Science: Horse wrangling in the digital age. In *Research workshop on "Horses" in Applied Machine Learning*, London, United Kingdom, October 2016

2. Comment déterminer le genre musical d'un morceau de musique comme « Psyché Rock » de Pierre Henry ?

Dans le but de mitiger ces problèmes, il semble convenable de complémenter ces deux approches par une approche intermédiaire en considérant ce que j'appellerai des données « simulées ». On verra dans la suite que, même si ces données simulées ne constituent pas une validation en soi, elles sont un outil particulièrement efficace pour une approche méthodologique de conception algorithmique qui soit mieux contrôlée et plus propice à être fondée sur un questionnement scientifique.

Prenons le cas de la tâche de détection d'évènements sonores dans des scènes sonores complexes. Ces scènes sonores sont composées d'évènements sonores qui peuvent être présent de manière simultanés ou non, et où un bruit de fond plus ou moins élevé est présent. On souhaite que le système soit capable :

1. d'être robuste à la présence du bruit de fond, *i.e.* ne pas détecter d'évènements quand il n'y a que du bruit de fond dans l'intervalle d'observation et ne pas se tromper de classes d'évènements lorsque le bruit de fond se superpose avec l'évènement dans l'intervalle d'observation ;
2. d'être capable de gérer la diversité intra classe de chaque classe d'évènements sonores et de type de bruit de fond, *i.e.* dans le cas de la classe « frappe sur clavier d'ordinateur », le système doit être capable de gérer une grande diversité de marques de claviers et de style de frappes ;
3. être robuste aux changements de conditions d'enregistrements, *i.e.* le type de microphone, la qualité de l'enregistrement et les propriétés de réverbération de la pièce doivent pouvoir changer sans influencer les performances du système.

Evaluer ces éléments nécessite d'avoir à disposition des corpus séparés avec, dans le premier cas, un niveau de bruit de fond qui augmente, dans le deuxième cas, une diversité qui varie, et dans le troisième cas des conditions d'enregistrements qui changent. Bien entendu, la prise de son et l'annotation de scènes sonores brutes pour chacune de ces conditions expérimentales sont possible mais sensiblement plus coûteuses à mettre en place pour obtenir le niveau de précision souhaité.

Nous avons donc considéré un modèle de scènes sonores [25] qui permet d'aisément générer des scènes sonores tout en contrôlant ces paramètres compositionnels, et ce à partir d'un corpus de sons isolés où chaque enregistrement audio fait partie d'une *collection*. Deux enregistrements font partie d'une même collection si ils sont appréciés de manière équivalente par des auditeurs dans un contexte d'écoute donné. Pour générer des scènes avec une certaine diversité en terme de contenu, il suffit alors de choisir différent éléments de la collection. Pour générer des scènes avec une certaine diversité en terme de composition, le modèle se base sur un processus de séquencement des sons où l'amplitude et l'écart temporel

Une implantation de ce modèle dédiée à l'évaluation a été réalisée en Matlab® (`https://bitbucket.org/mlagrange/simscene`). Un outil alternatif a été développé en Python par Justin Salamon (`https://github.com/justinsalamon/scaper`).

25. Grégoire Lafay. *Simulation de scènes sonores environnementales : application à l'analyse sensorielle et à l'analyse automatique*. Thèse de doctorat, École Centrale, Nantes, 2016

entre deux sons d'une même source sont contrôlés par des processus stochastiques dont les paramètres sont estimés à partir de scènes sonores annotés en terme de présence de source.

Nous avons mis en application ce modèle dans la première édition du challenge dcase, pour "detection and classification of acoustic scenes and events". Ce challenge, soutenu par le comité technique ieee "audio and acoustic signal processing technical committee" (aasp), se propose de promouvoir l'activité scientifique et technique autour de la détection d'évènements et la classification de scènes dans un contexte sonore environnemental. Nous avons proposé dans ce cadre une tâche de détection d'évènements sonores se basant sur l'approche par simulation, à partir d'un corpus de scènes simulées grâce à l'outil simScene. Ce corpus utilisait des échantillons sonores et des bruits de fonds enregistrés à l'université de Queen Mary (qmul). Ce corpus sera nommé *testQ*. Plusieurs laboratoires ont soumis leurs algorithmes, qui ont été transmis aux organisateurs qui se sont chargés de calculer les prédictions de ces algorithmes sur les données de test gardées secrètes lors du déroulement du challenge.

Ces résultats publiés,[26] nous avons ensuite voulu tester la capacité de généralisation de ces algorithmes, *i.e.* leur aptitude à maintenir des performances de détection similaires sur plusieurs corpus de scènes présentant des conditions expérimentales différentes. Pour ce faire, nous avons considéré un autre corpus d'échantillons sonores et de bruits de fonds cette fois ci enregistrés à l'irccyn, Nantes pour générer de nouvelles scènes de test. Avec la permission des auteurs des différents systèmes soumis au challenge, ces derniers ont été évalués sur les corpus de scènes simulées, en utilisant les mêmes serveurs de calcul que ceux utilisés pour le challenge.

26. Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, 17(10), 2015

De nombreuses métriques sont disponibles pour évaluer ce type de système. On se concentrera ici sur la métrique Fcw_{eb}, une F-mesure calculée en prenant en compte les débuts des événements bien identifiés, et en normalisant les résultats par classe. Un évènement est considéré correctement identifié si le système identifie le début d'un évènement de la classe correspondant à la classe annotée. Une imprécision temporelle de 200 ms est tolérée.

La capacité de généralisation est considérée suivant trois angles :

1. robustesse à la diversité des échantillons : évaluer la capacité de généralisation sur des corpus de scènes possédant les mêmes caractéristiques structurelles (intensité sonore des échantillons, positionnement des échantillons), mais composés d'une sélection de échantillons différents.

2. robustesse à la diversité structurelle : évaluer la capacité de généralisation sur des corpus de scènes composés des mêmes échantillons, mais dont les caractéristiques structurelles (intensité sonore des échantillons, positionnement/espacement moyen des échantillons) diffèrent ;

3. robustesse au niveau du bruit de fond. Ce niveau est mesuré en termes relatifs, avec la mesure ebr (pour "event-to-background ratio"), exprimée en décibels.

Dans le premier cas, on considère une structure de scène de référence (correspondant aux annotations effectuées sur *testQ*) pour générer deux nouveaux corpus *insQ*, et *insI*. Le premier est généré à partir d'échantillons enregistrés à qmul, le second à partir d'échantillons enregistrés à l'irccyn. Le premier ayant une structure et un contenu sonore équivalent à *testQ*, une certaine équivalence des performances est attendue. Comme on peut le voir sur la Figure 8 c'est globalement le cas, sauf pour deux méthodes qui se retrouvent avec des performances drastiquement réduites. Une analyse d'erreur a montré que ces changements de comportement étaient dûs à des phénomènes d'instabilité de certains prédicteurs. Le second corpus (*insI*) permet d'évaluer la capacité de généralisation aux types d'échantillons. A l'exception du système scs, la chute de performance de tous les autres algorithmes à des niveaux de performance équivalent à l'algorithme de référence (« Baseline ») lors du changement de corpus d'échantillons (de *Q à *I) montrent que l'ensemble de ces algorithmes ne sont pas capable de généraliser. On verra dans l'étude suivante que la bonne robustesse du système scs est probablement due à une bonne gestion du bruit de fond.

Dans le deuxième cas, on considère le modèle de génération de scènes précédemment décrit pour générer deux corpus avec une certaine diversité structurelle *absQ* et *absI*. Le premier considère des échantillons enregistrés à qmul et le second à l'irccyn. Les paramètres du modèle de haut niveau sont estimés à partir des annotations du corpus *testQ*. L'équivalence des résultats obtenus entre les corpus *insQ* et *absQ* ainsi que *insQ* et *absQ* montrent une faible dépendance des algorithmes à la structure temporelle des scènes analysées. Ce résultat est attendu car la majorité des algorithmes ont peu de modélisation des dépendances temporelles.

Dans le troisième cas, on considère quatre corpus, avec la même structure temporelle, les mêmes échantillons sonores, mais avec un niveau relatif de bruit de fond différent. L'ebr varie de -12 dB à 6 dB, le niveau de référence (équivalent à celui mesuré sur le corpus *testQ*) étant de 0 dB. Les résultats, affichés sur la Figure 9, montrent une forte sensibilité au niveau d'ebr pour toute les systèmes, sauf pour le système scs, doté d'un algorithme de gestion du bruit de fond particulièrement efficace. [27]

Notre implication dans le challenge dcase s'est confirmée en 2016, par la proposition d'une tâche similaire à celle proposée en 2013, avec quelques raffinements. [28] Pour cette édition, nous souhaitions étudier l'influence de la répartition temporelle des évènements dans la scène, qui, dans le type de structures utilisées dans le l'édition de 2013 avait peu d'impact, probablement à cause d'une répartition assez parcimonieuse, sans recouvrement temporel entre les évènements. On distingue donc maintenant deux types de scènes, les scènes polyphoniques, scènes où les événements de différentes

27. Grégoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, and Axel Roebel. A morphological model for simulating acoustic scenes and its application to sound event detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(10):1854–1864, 2016

28. Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D. Plumbley. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(2):379–393, February 2018

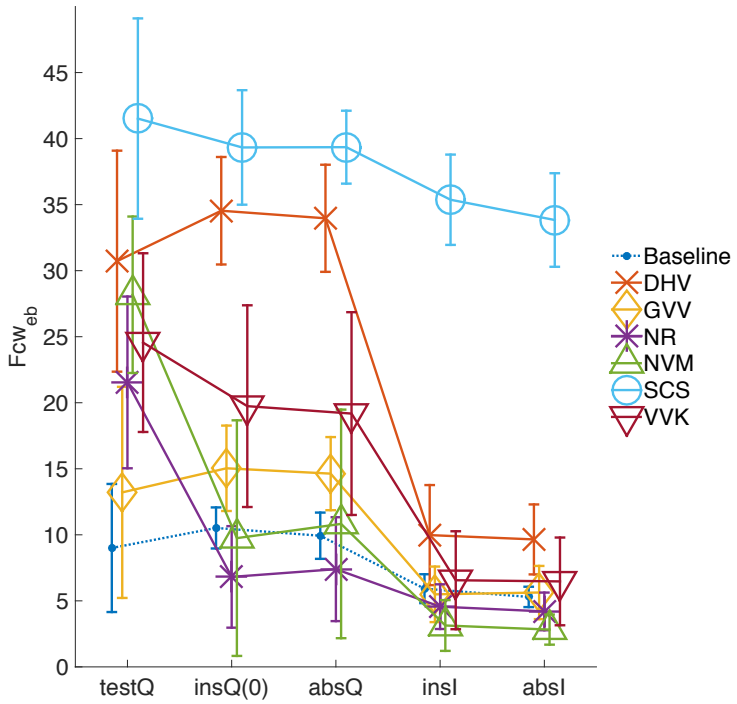


FIGURE 8: Performances des systèmes évalués dans le cadre du challenge dcase 2013 sur les corpus qmul et irccyn en considérant la métrique Fcw_{eb}.

classes peuvent se recouvrir temporellement, et les scènes non-polyphoniques, scènes où un seul événement peut être actif à un moment donné.

Au sein de ces deux types de scènes, deux paramètres sont considérés pour contrôler la simulation des scènes sonores : 1) le rapport moyen entre les niveaux des événements et du bruit de fond, et 2) le nombre d'événements présents pour chaque classe.

Une analyse statistique des résultats [29] montre un impact non significatif de la polyphonie et de la densité d'évènements, les systèmes réagissant différemment à la modification de ces facteurs expérimentaux, chaque système ayant opté pour un compromis précision / rappel différent. L'influence de l'intensité des évènements sonores par rapport au bruit de fond est, par contre, elle, significative. Ce dernier résultat indique qu'il est capital, afin d'améliorer les performances des algorithmes, de proposer des systèmes dotés d'une gestion du bruit de fond efficace.

29. Grégoire Lafay, Emmanouil Benetos, and Mathieu Lagrange. Sound Event Detection in Synthetic Audio: Analysis of the DCASE 2016 Task Results. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, United States, September 2017

Je terminerai cette section par des éléments de réflexion que j'ai amené lors du workshop dcase 2018 où le constat était fait par une large communauté de chercheurs que les approches d'apprentissage profond et les méthodes ensemblistes appliquées à des classifieurs forts, apportent certes des performances élevées, mais finalement peu d'éléments de réflexion et compréhension.

A mon avis, si la question posée est d'ordre pratique, ce n'est pas un problème que les réponses données ne soient que techniques. Comme nous en discuterons plus en détail dans le chapitre dédié à la méthode scientifique en sciences des données, encore faut-t-il que ces réponses aient un intérêt pour la communauté. Pour cela,

Sans contraintes explicites, tous les moyens sont bons pour résoudre le problème.

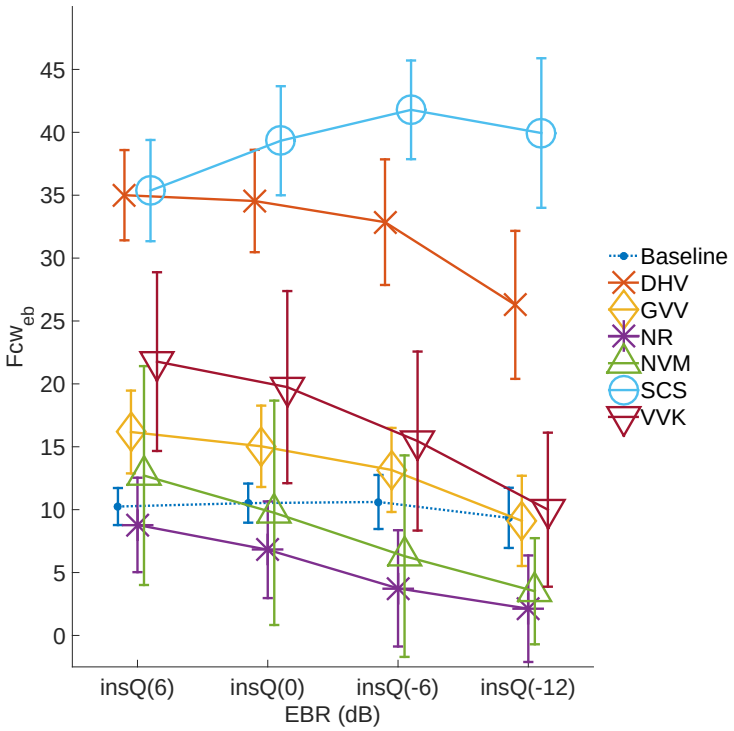


FIGURE 9: Performances des systèmes évalués dans le cadre du challenge dcase 2013 sur les corpus *insQ* simulés avec différents ebr (6, 0, -6 et $-12dB$).

la poursuite d'une méthode rigoureuse d'investigation reste d'importance (protocole expérimental approprié, précision du rapport, publication du code, réplicabilité, ...).

Ceci ne doit pas en revanche être confondu avec une approche d'investigation scientifique qui doit, elle, avant tout se baser sur une question. De la question naît un corpus d'observations qui peut servir ensuite de base à l'investigation qui peut comporter une partie technique, mais qui reste au service d'un questionnement qui lui est, par essence, scientifique. Il est donc du devoir de celui qui propose une tâche, s'il y a lieu, de réfléchir à la pertinence de cette tâche au vu d'un certain questionnement scientifique pertinent pour la communauté.

La synthèse pour la psychologie expérimentale

Ces travaux nous ont bien entendu amenés à écouter un grand nombre de scènes simulées, et à notre grande surprise, ce matériau sonore destiné originellement à n'être écouté que par des machines, s'est révélé particulièrement crédible en terme de qualité d'écoute. Ceci nous a amené à nous questionner sur la pertinence de ce type de stimulus pour nourrir des études de psychologie cognitive.

J'ai eu le plaisir de confirmer cette intuition en contribuant au domaine de la psychologie expérimentale, en collaboration avec Grégoire Lafay et Nicolas Misdariis de l'ircam. Nous avons souhaité questionner la notion d'agrément sonore en zone urbaine, sujet d'une grande importance sanitaire.

Pour cela, le modèle de scènes décrit précédemment, de part son réalisme et sa simplicité de manipulation s'est révélé être un outil pertinent pour questionner d'une nouvelle manière des notions de qualité perçue comme l'agrément sonore, *i.e.* la qualité hédonique de l'environnement : « Est-ce que l'environnement est agréable ou désagréable ? ». L'étude de référence sur le sujet de l'agrément sonore en zone urbaine, effectuée par Catherine Guastavino, [30] considère un paradigme expérimental usuel, basé sur la description faite par le sujet de l'objet d'étude. Chaque sujet est invité à verbaliser les propriétés d'un environnement sonore idéal ou non idéal. Ces descriptions textuelles sont ensuite traitées par une analyse psycho linguistique, des invariants dans les descriptions données par les sujets sont alors dégagés qui permettent de conclure quant aux propriétés de ces environnements.

30. Catherine Guastavino. The ideal urban soundscape: Investigating the sound quality of french cities. *Acta Acustica united with Acustica*, 92(6): 945–951, 2006

On note ici que ces environnements sont pensés par le sujet, puis décrits, ce qui place l'étude dans de la cadre de la théorie classique de la cognition. Les percepts subissent une étape de transduction en informations amodales que l'on peut alors questionner par le langage. Des théories alternatives existent, comme la théorie ancrée [31] dont le propos est de ne pas distinguer perception et cognition en terme d'objets manipulés, mais en terme d'objectifs. Elle met à ce titre l'accent sur le contexte et la notion d'expérience, d'"embodiement". Dans cette théorie, les percepts sont le fondement, la matière brute des constructions cognitives qui en découlent. La cognition n'est plus en charge d'une transduction mais d'une extraction d'invariants qui serviront à construire des représentations de plus en plus abstraites. [32]

31. Lawrence W Barsalou. Grounded cognition: Past, present, and future. *Topics in cognitive science*, 2(4):716–724, 2010

Notre proposition s'inscrit dans ce schéma de pensée, en demandant non plus au sujet de verbaliser son image mentale mais de « construire » une représentation de l'image mentale que le sujet a d'un environnement sonore idéal. Comme on peut s'en rendre compte à l'écoute, de telles scènes sont plausibles mais ne sont pas pour autant réalistes au sens strict du terme. On notera que l'on cherche ici, et ce avec un certain nombre de contraintes (durée de la scène, choix des éléments de base, paradigme de séquencement), en quelque sorte à exemplifier ces images mentales de haut niveau. Sans parler de caricatures, le côté pittoresque des scènes produites, notamment pour les scènes idéales, inscrit finalement bien le protocole proposé dans la théorie ancrée de la perception.

32. En prenant un point de vue science des données, on peut s'amuser ici à rapprocher cet argumentaire de celui couramment utilisé pour motiver l'effectivité des architectures des réseaux neuronaux profonds.

Les scènes produites par les sujets sont disponibles pour écoute : `http://soundthings.org/research/urbanSoundScape/XP2014`.

Cette étude a permis de confirmer certains résultats obtenus dans l'étude de Catherine Guastavino et d'en questionner d'autres. On citera par exemple, la présence systématique des bus dans les scènes idéales imaginées « par écrit », alors que notre étude les placent plutôt dans les scènes non idéales. On voit ici, comment d'un point de vue conceptuel, les transports en commun sont positivement connotés dans un environnement urbain, alors qu'ils restent des objets mobiles lourds avec une cylindrée conséquente, générant des niveaux de bruits mécaniques élevés. [33]

33. Grégoire Lafay, Mathias Rossignol, Nicolas Misdariis, Mathieu Lagrange, and Jean-François Petiot. Investigating soundscapes perception through acoustic scenes simulation. *Behavior Research Methods*, 2(51), October 2018

Au travers de ces deux exemples, on voit comment la capacité

à manipuler la matière sonore nous permet de mieux questionner notre connaissance de la manière dont nous, humains, percevons et faisons sens de notre environnement. Le modèle présenté ici est à proprement parler un modèle de séquencement de sons, donc d'assez haut niveau. Nous étudierons dans le chapitre suivant les nombreux modèles de sons disponibles pour manipuler plus finement la matière sonore.

Modélisation long terme de signaux sonores

La plupart des modèles de sons utilisés dans la littérature sont des modèles à court terme, c'est-à-dire qu'ils modélisent le signal sur une courte période de temps. Les raisons de ce choix de conception sont nombreuses et seront détaillées ci-après. Une originalité de mon travail de recherche a été d'étudier les limitations de ce type d'approche et de contribuer à des alternatives qui, sans être totalement satisfaisantes à ce jour, nous permettent de mieux comprendre les défis scientifiques et techniques pour parvenir à une modélisation des signaux sonores qui soit à la fois fidèle, versatile, et expressive.

Préambule

On considérera ici qu'un signal sonore est un signal à valeurs réelles, échantillonné à une fréquence de 40 kHz environ. À titre d'exemple, la Figure 10 montre la forme d'onde d'une note de piano. Ce signal peut provenir d'une version quantifiée, mais on supposera que le pas de quantification est suffisamment réduit pour n'introduire qu'une dégradation perceptivement négligeable. On notera ce signal $x[n]$. Il peut être observé sur un intervalle de temps donné, correspondant à un certain nombre d'échantillons temporels $x[n], n \in [1, ..., N]$ souvent appelé intervalle d'observation, trame, support ou encore champ réceptif. Je n'utiliserai pas dans ce document la version continue du signal, l'expression $f(x)$ devant se comprendre comme l'application d'une fonction f à un signal x.

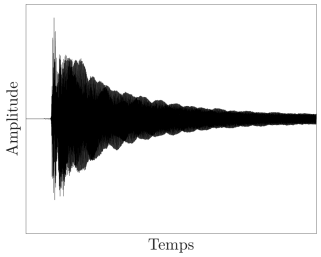


FIGURE 10: Signal temporel ou forme d'onde d'une note de piano.

Le terme modélisation étant fortement polysémique, précisons l'usage qui sera fait du terme « modèle de son ». Un modèle de son est une représentation du signal sonore qui permet de modifier ce signal d'une manière qui soit d'intérêt pour une tâche donnée et ce de manière cohérente avec certaines contraintes motivées par 1) des considérations physiques ou 2) la perception qu'en aurait un être humain. On notera $\mathrm{Pe}(x)$, l'opérateur de perception humaine du signal x. Le modèle est utilisé en considérant deux étapes de transformation, respectivement appelées « analyse » et « synthèse ». La phase d'analyse An, à partir d'un signal x, produit un modèle X et la phase de synthèse Sy permet, à partir de ce modèle X, de produire un signal $\tilde{x} = \mathrm{Sy}(\mathrm{An}(x))$.

Trois critères d'évaluation sont d'importance pour qualifier ce modèle pour des activités de conception, création et manipulation :

1. **fidélité** : un modèle fidèle est capable de ne pas produire d'artefacts. Ce critère peut s'exprimer formellement de la manière suivante : $\mathrm{Pe}(x) \sim \mathrm{Pe}(\tilde{x})$.
2. **expressivité** : un modèle expressif est capable de mettre à disposition des mécanismes de manipulation tels que la modification du modèle produise une modification cohérente de la perception du signal résultant. Ce critère peut s'exprimer formellement de la manière suivante : $\mathrm{Pe}(\mathrm{Sy}(X')) \sim \mathrm{Pe}(x')$, où X' désigne une version du modèle X modifié avec des opérateurs simples, et x' désigne une version modifiée du signal x à partir d'opérateurs complexes : étirement, transposition, etc.
3. **versatilité** : un modèle versatile s'applique à tout type de signal sonore d'intérêt, *i.e.* les conditions de fidélité et d'expressivité sont remplies pour tout signal x d'intérêt pour une tâche donnée.

Je ne considère pas ici la notion d'efficacité, *i.e.* la capacité à modéliser des signaux avec peu de paramètres. Cette notion sera un élément prépondérant d'évaluation par exemple pour un algorithme de compression. De par mon expérience, les notions d'expressivité et d'efficacité sont subtilement liées. Il est néanmoins important de noter qu'un modèle efficace n'est pas forcément expressif, et qu'un modèle expressif n'est pas forcément efficace.

L'expressivité sera évaluée ici au regard des modifications canoniques du signal sonore, à savoir :

— la modification du volume ;

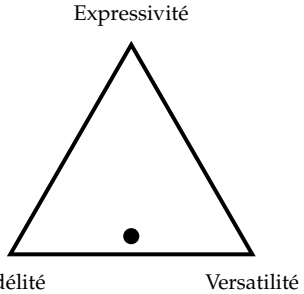
— la modification de la durée ;

— la modification de la hauteur ;

— la modification du timbre.

Bien d'autres peuvent être imaginées ou désirées, en fonction de projets de création spécifiques.

Avec ces trois critères d'évaluation, se dessine ici un triangle de contraintes qui nous permettra d'apprécier le compromis inhérent au choix de conception de chacun des modèles que nous discuterons. Par exemple, la forme d'onde est un modèle versatile et fidèle mais d'une expressivité limitée, car il n'y a guère que le volume que l'on puisse manipuler aisément.



La forme d'onde est un modèle de son versatile et fidèle mais d'une expressivité limitée.

L'intuition ici, largement suivie dans la communauté des sciences des données est que, pour tout processus observable, il existe un modèle de ce processus pour lequel des manipulations basées sur des opérateurs mathématiques simples dans le domaine transformé « font sens » dans le domaine signal. On jugera alors la qualité du modèle par sa capacité à proposer de nouvelles réalisations du processus modélisé qui soient plausibles, dans notre cas perceptivement valides. Par exemple, en considérant le signal temporel, on peut aisément manipuler l'intensité perçue en multipliant tous les échantillons par une constante. Ce modèle ne permet par contre pas dans le cas général de manipuler aisément la hauteur perçue.

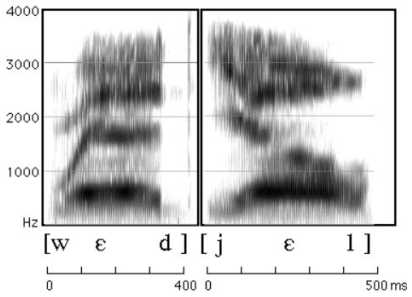
Cette approche se distingue clairement des approches par « modèle physique » qui modélisent explicitement le processus mécanique, la qualité du son étant alors censée être la résultante d'une bonne modélisation du système physique dont la mise en vibration est à l'origine du signal sonore d'intérêt. Le distinguo est ici fait sur l'objet d'intérêt, à savoir que l'on s'intéresse aux propriétés du signal et non à celles de la source. Ceci ne veut pas dire que l'on ne puisse pas, avec une approche sciences des données, se focaliser sur le signal d'un type de source sonore en particulier. Mais, dans ce cas, on s'attachera à exploiter les régularités du signal pour motiver notre modèle.

Dans la communauté traitement du signal, on parle souvent de modèle paramétrique, dans le sens où le modèle dispose d'un ensemble réduit de paramètres qui peuvent être explicitement manipulés pour influer sur le son modélisé. L'usage du terme « paramétrique » est malheureusement utilisé avec peu de cohérence dans la littérature. Je préférerai donc à ce terme le distinguo entre « modèle » et « transformée ». Un modèle expose donc pour moi un ensemble de paramètres manipulables, une transformée étant plutôt un outil de description. Ceci n'empêche pas qu'une transformée puisse avoir un modèle sous-jacent, ni que les étapes d'analyse et / ou de synthèse d'un modèle ne se basent sur une ou plusieurs transformées.

Précisons maintenant l'intervalle temporel dénommé par « long terme ». Cela désignera ici une durée à partir de laquelle le signal sonore ne peut plus raisonnablement être considérée comme stationnaire. De par nos connaissances sur les propriétés physiques de l'organe phonatoire, il est par exemple commun de considérer que le signal de parole est localement stationnaire sur un intervalle de 25 ms. Comme on peut le voir en observant des spectrogrammes de vocalisations humaines,[34] dès que l'on considère des consonnes, ou lorsque la prosodie évolue de manière conséquente, le signal sonore produit par l'organe vocal n'est plus stationnaire. Il me paraît donc plus juste de motiver ce choix par l'assertion suivante : « il est raisonnable de modéliser le signal de parole avec un modèle stationnaire par morceaux à partir du moment où ces morceaux sont d'une durée inférieure à 25 ms ».

La dénomination « long terme » étant dépendante du degré de stationnarité du signal étudié, il n'est donc pas aisé de fixer une durée pour tout type de signaux sonores à partir de laquelle la dénomination « long terme » est appropriée. On considérera néanmoins dans les études discutées dans ce document qu'un intervalle de temps « long terme » désigne un intervalle de temps supérieur ou égal à la seconde.

34. Peter Ladefoged and Keith Johnson. *A course in phonetics*. Nelson Education, 2014



Spectrogrammes de séries de phonèmes (figure issue de la référence n° 34).

Typologie sonore

Il est bien entendu que même si l'on souhaite que notre modèle soit versatile, on ne souhaite pas être capable de modéliser tous

les signaux possibles. On supposera donc que certaines propriétés d'invariance existent, c'est-à-dire que même si une seconde de son échantillonné à 44 kHz est représentée dans R^{44000}, les paramètres du modèle évoluent dans un espace de dimension beaucoup plus réduit. Je considère ce postulat raisonnable essentiellement pour des raisons d'observations physiques. En prenant l'exemple du signal de parole, il est bien entendu que le son produit est le résultat d'une interaction complexe entre un ensemble d'éléments qui ont une certaine plasticité. Il reste néanmoins évident que les déformations possibles de l'ensemble de ces éléments sont bornées.

Dans le but de préciser le domaine d'investigation, je présente ici cinq grandes familles de signaux sonores :

— **parole** : sons voisés <a>, <o> / sons plosifs <pe>, <qe> ;

— **communication animale** : hululement de chouettes / clics de localisation de chauve souris ;

— **musique** : chant lyrique / percussions ;

— **mécanique** : ventilation / marteau piqueur ;

— **environnementaux** : vent faisant siffler des câbles / gouttes de pluie tombant sporadiquement.

Cette typologie est ordonnée par période approximative d'intérêt de la communauté de traitement du signal. Le traitement de la parole a connu un essor dans les années 1970, la bio acoustique un peu plus tard, la musique dans les années 2000 et les signaux environnementaux dans les années 2010.

Les exemples donnés pour chaque classe sont choisis à dessein pour illustrer une dichotomie présente dans toutes ces classes. Elle nous servira de base pour notre critique des outils de modélisation à notre disposition et pour identifier les défis majeurs à relever.

Suivant cette dichotomie, certains sons portent une structuration fréquentielle forte et d'autres une structuration temporelle forte. La Figure 3 montre deux exemples dans le domaine instrumental. Ces exemples sont volontairement typés pour expliciter le propos, mais la plupart des sons naturels sont une composition de ces deux types de structures. Il est donc nécessaire pour atteindre nos objectifs de disposer d'outils performants pour modéliser conjointement ces deux types de structure, temporelle et fréquentielle.

Temps / fréquence

Lorsque l'on cherche à décrire un son, deux dimensions émergent : l'intensité et la hauteur. Ces deux dimensions étant respectivement intimement liés à l'amplitude et à la fréquence, il est normal de constater que la plupart des modèles de son permettent de manipuler ces deux dernières quantités de manière dynamique, *i.e.* au cours du temps.

En supposant le signal sonore stationnaire sur l'intervalle d'observation, l'outil privilégié pour ce faire est la transformée de Fourier discrète (tfd). Cette transformée projette le signal sur une base

d'exponentielles complexes :

$$X[m] = \sum_{n=0}^{N-1} x[n] \cdot e^{\frac{-2j\pi nm}{N}} \quad (9)$$

$$x[n] = \frac{1}{N} \sum_{m=0}^{N-1} X[m] \cdot e^{\frac{2j\pi mn}{N}}. \quad (10)$$

Cette transformée est donc inversible.

On calcule la transformée de Fourier de manière efficace grâce à des algorithmes de transformée de Fourier rapides qui exploitent, pour le plus connu d'entre eux, des longueurs de supports en puissance de deux. [35]

Comme on peut le voir sur la Figure 11, en considérant le module de cette transformée, on manipule directement l'intensité du son à une fréquence donnée. Le spectre de phase est lui souvent négligé dans les modèles, car il est beaucoup plus difficile à interpréter. Ceci constitue un handicap majeur pour tous les modèles de sons basés sur la transformée de Fourier car cela pose bien souvent une limite sur la propriété de fidélité, le module et l'exposant étant mathématiquement liés.

En considérant qu'une partie des sons que l'on cherche à modéliser possèdent une structure temporelle forte comme une note de piano ou un sifflement d'oiseau, on s'intéresse à l'estimation des paramètres de composantes périodiques simples, les sinusoïdes. Considérons pour cela un premier signal « éprouvette » constitué d'une sinusoïde de fréquence et d'amplitude constante :

$$x_s[n] = a_s \sin(2\pi f_s n / f_e + \phi_s) \quad (11)$$

Les paramètres d'amplitude et de fréquence de cette composante s'estiment aisément grâce au spectre de Fourier de ce signal :

$$\hat{f}_s = \frac{M}{N} . f_e \quad (12)$$

$$\hat{a}_s = X_s[M] \quad (13)$$

avec

$$M = \underset{m}{\operatorname{argmax}} |X_s[m]|. \quad (14)$$

Dans le cas de plusieurs sinusoïdes, on remplacera l'équation 14 par une sélection de maxima locaux dans le spectre de puissance et on considèrera l'usage d'une fenêtre de type co-sinusoïdale (hamming, hann, ...) pour améliorer la localité fréquentielle de la transformée. [36] En effet, considérer le signal sur un intervalle de temps fixé par la largeur de la trame revient implicitement à fenêtrer le signal par une fenêtre rectangulaire qui a des propriétés spectrales non désirables dans le cas multi composantes. Avec des fenêtres co-sinusoïdales, le lobe principal s'élargit, d'où une perte locale de résolution, mais permet de réduire considérablement l'importance des lobes secondaires, ce qui s'avère indispensable dans un contexte multi composantes.

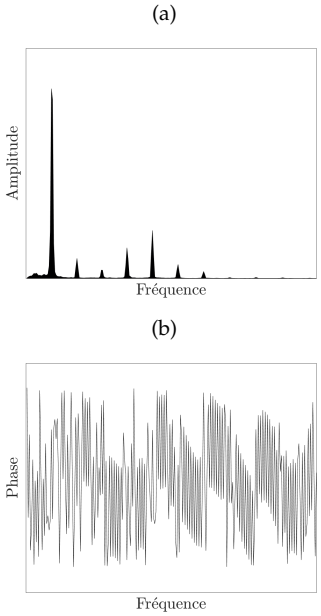
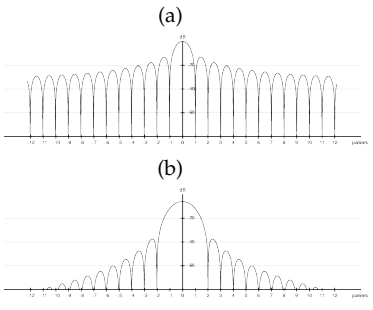


FIGURE 11: Spectre d'amplitude (a) et de phase (b) d'une note de piano.

35. James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90): 297–301, 1965



Spectre d'amplitude de la fenêtre rectangulaire (a), et de la fenêtre de Hann (b).

36. Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978

On peut améliorer la précision de ces estimations en réduisant le pas de quantification de l'axe fréquentiel. On utilise pour cela la technique du "zero padding", qui consiste à concaténer une série de valeurs nulles au signal observé avant d'opérer la transformation. La transformée s'opérant alors sur plus de points, son pas de discrétisation fréquentielle s'en trouve augmenté. Par conséquent, la précision des estimateurs de l'équation 13 en est augmentée.

Un large échantillon de méthodes sont disponibles pour améliorer la précision de ces estimations par d'autres moyens moins directs, mais souvent plus efficaces. Celles basées sur le réassignement [37] étudient les relations entre le spectre du signal fenêtré et le spectre du signal fenêtré par la dérivée de la fenêtre. Les méthodes basées sur la phase étudient elles les relations entre le spectre du signal et celui de sa dérivée. Ces deux méthodes, conceptuellement proches, sont en fait équivalentes théoriquement et numériquement. [38]

Il est important de bien garder en mémoire que seule la précision des estimateurs est augmentée, la résolution de la transformée restant dépendante de la largeur du support temporel. Augmenter l'intervalle d'observation va donc non seulement influer sur la précision, mais également sur la résolution de la transformée. Ainsi, deux sinusoïdes de fréquences proches ne pourront être « résolues », c'est-à-dire que leurs paramètres ne pourront être estimés en observant directement le spectre d'amplitude que par l'extension suffisante du support temporel de la transformée. Comme nous allons le voir dans l'exemple suivant, cette extension a un coût.

Considérons un autre signal « éprouvette » constitué cette fois ci d'une impulsion de Dirac :

$$s_d[n] = \begin{cases} a_d \text{ si } n = n_d \\ 0 \text{ sinon} \end{cases} \tag{15}$$

L'estimation du paramètre d'amplitude a_d et de sa position temporelle n_d est triviale dans le domaine temporel :

$$\hat{n_d} = \underset{n}{\operatorname{argmax}} |s_d[n]| \tag{16}$$

$$\hat{a_s} = s_d[\hat{n_d}]. \tag{17}$$

Cette dernière est par contre beaucoup plus difficile à estimer dans le domaine spectral, du fait du caractère non stationnaire du signal analysé.

Pour palier au caractère non stationnaire à long terme de la plupart des signaux naturels, il est courant d'utiliser la transformée de Fourier à court terme (tfct). Son principe est d'effectuer une série de transformée de Fourier, chacune sur des fenêtres espacées dans le temps :

$$X[m,t] = \sum_{n=-\infty}^{\infty} x[n]w[n-t]\mathrm{e}^{\frac{-2\mathrm{j}\pi mn}{N}} \tag{18}$$

où w est la fenêtre d'observation. En prenant le module de la tfct, on obtient le bien connu spectrogramme, un outil de manipula-

37. François Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on signal processing*, 43(5):1068–1089, 1995

38. M. Lagrange and S. Marchand. Estimating the instantaneous frequency of sinusoidal components using phase-based methods. *Journal of the Audio Engineering Society*, 2007

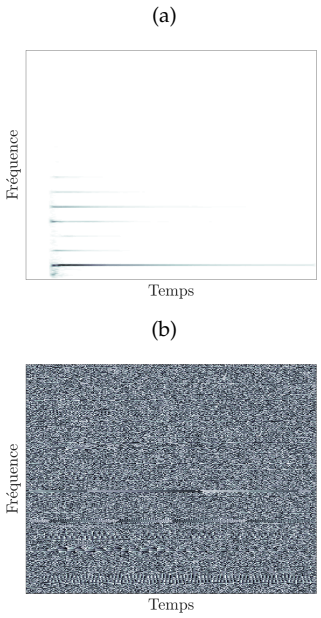
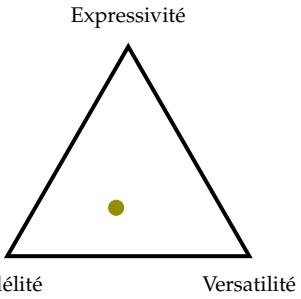


FIGURE 12: Spectrogramme d'amplitude (a) et de phase (b) d'une note de piano où la teinte représente la phase. TODO HSV

tion et de visualisation qui trouve son utilité dans de nombreux domaines d'application. Pour exemple, la Figure 12 montre le spectrogramme d'amplitude et de phase d'une note de piano. Pour ce cas simple avec une bonne parcimonie et une structure temporelle élevée, des régularités sont observables dans le spectrogramme de phase aux fréquences des harmoniques de la note de piano. Pour des sons plus complexes, l'interprétation du spectre de phase devient ardue.



Le spectrogramme est à ce jour le compromis le plus apprécié.

L'espoir ici est que les erreurs de modélisation des non stationnarités du signal seront négligeables du fait du faible support temporel des fenêtres sur lesquels on applique successivement la tfd. Dans le cas où les fenêtres se recouvrent de moitié, on peut utiliser la transformée en cosinus discret modifiée (tcdm) pour conserver un échantillonnage critique. [39]

39. John Princen and Alan Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5): 1153–1161, 1986

En considérant la tfct, il est alors possible de construire un estimateur « spectral » de la position temporelle de notre dirac :

$$\hat{n}_d = \underset{t}{\operatorname{argmax}} \sum_m \|X[m,t]\| \tag{19}$$

À condition d'utiliser une fenêtre » à rampe « , on peut améliorer la précision temporelle de cet estimateur en réduisant le recouvrement entre fenêtres. La encore, il s'agit d'une amélioration de la précision et non de la résolution. Seule la réduction de l'intervalle d'observation permettra de résoudre la position temporelle de deux diracs proches.

Au travers de ces deux exemples, on voit qu'il est possible d'améliorer dans une certaine mesure la précision des estimations effectuées à partir d'un plan temps / fréquence obtenu grâce à la tfct. Néanmoins, les résolutions temporelles et fréquentielles restent conflictuellement contraintes par la taille de la fenêtre choisie, [40] comme énoncé par le principe d'incertitude d'Heisenberg. On parle ici de variables complémentaires pour le temps et la fréquence, au même titre que la position et la quantité de mouvement en mécanique.

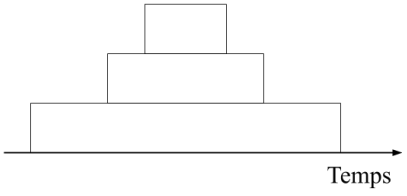
40. David Slepian. Some comments on fourier analysis, uncertainty and modeling. *SIAM review*, 25(3):379–393, 1983

Lorsque l'on considère des signaux qui possèdent à la fois une localité temporelle et fréquentielle, ce qui est le cas de la plupart des signaux sonores naturels, il est alors impossible de déterminer une résolution convenable.

Un exemple iconique étant la note de piano dont la qualité s'exprime autant du côté de la localité temporelle (netteté de l'attaque) que de la localité fréquentielle (pureté des harmoniques).

Il paraît alors naturel de considérer des approches à résolution multiples. Tout le problème consiste alors à trouver une représentation qui maximise le potentiel apporté par la multirésolution en terme de fidélité tout en minimisant la perte en expressivité. En effet, la manipulation de données non régulièrement échantillonnées ou recopiées sous divers formats, sont un frein à l'aisance de manipulation.

Pour parvenir à une représentation à résolution multiple, l'approche la plus simple conceptuellement consiste à « empiler » plusieurs tfd de largeurs de fenêtre différentes. Cette technique peut se révéler utile dans le cas d'inspection de données ou de prises de décision. Il est par exemple courant de procéder à des fusions



Analyses fréquentielles effectuées sur des supports temporels différents.

ensemblistes de classifieurs d'architectures équivalentes effectuant leurs prédiction à partir de représentations spectrales de résolution différentes. Dans le cas d'un modèle de son, cela induit malheureusement une perte considérable en manipulabilité du fait de la perte d'unicité de la représentation.

Pour éviter cette perte, il est nécessaire de baser notre modèle sur une représentation qui soit à même de partager judicieusement le plan temps / fréquence de manière à mitiger au mieux la contrainte posée par le principe d'incertitude. Un immense travail de la communauté de traitement du signal a été de concevoir ce type de transformée, connue à présent sous le nom de transformée en « ondelettes ».[41]

Soit $\psi(t)$ un filtre de fréquence centrale ξ et de largeur ξ/Q, où Q est le facteur de qualité du filtre, un banc de filtre en ondelettes est construit en dilatant $\psi(t)$ par une séquence géométrique d'échelles $2^{\lambda/Q}$ pour obtenir

$$\psi_\lambda(t) = 2^{-\lambda/Q}\psi(2^{-\lambda/Q}t). \tag{20}$$

La variable λ est un facteur d'échelle prenant des valeurs entières entre 0 et $(JQ-1)$, où J est le nombre d'octaves couvertes par le banc de filtre.

La transformée en ondelettes d'un signal $x(t)$ est obtenue par convolution avec tous les filtres. Un scalogramme, comme celui de la Figure 13, s'obtient par application du module au résultat :

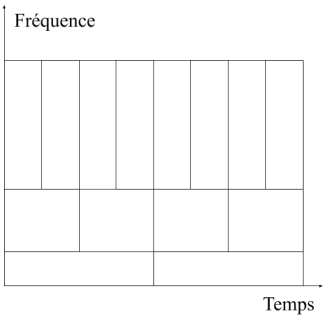
$$x(t,\lambda) = |x * \psi_\lambda|(t). \tag{21}$$

En considérant ce type de décomposition, on obtient alors un pavage du plan temps / fréquence alternatif à celui du spectrogramme. Du fait du principe d'incertitude, les pavés ont la même aire, mais le rapport entre résolution fréquentielle et temporelle est optimalement adapté, contrairement à celui du spectrogramme qui lui reste constant. La Figure 13 montre le scalogramme d'une impulsion de Dirac. On y observe clairement l'augmentation de la résolution temporelle en fonction de la fréquence.

En considérant que le banc de filtres est construit de manière itérative en partant des hautes fréquences, et en notant que le signal observé est d'une durée limitée, une partie du spectre en basse fréquence ne peut être capturé. On ajoute alors un filtre passe bas dit « de terminaison » qui permet de préserver l'unicité de la représentation. Ce dernier filtre, souvent appelé « fonction d'échelle » est peu utilisé dans le cas du traitement de l'audio car il est aisé de capturer le spectre audible avec le banc de filtre en ondelettes et ainsi ne pas considérer les infrasons. Les sorties de ce banc de filtres peuvent être ensuite re-échantillonnées à une fréquence de Nyquist correspondante à la largeur de la bande d'octave du filtre considéré.

Comme le montrent les travaux pionniers effectué par Richard Kronland-Martinet,[42] l'application de la transformée en ondelettes paraît avoir d'excellentes propriétés pour le traitement de l'audio,

41. Stéphane Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989



Pavage du plan temps / fréquence de la transformée en ondelettes.

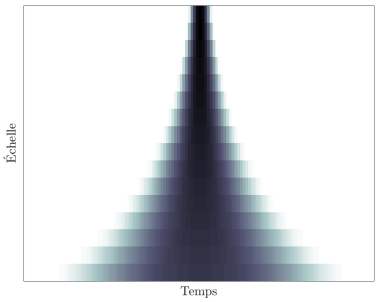


Figure 13: Scalogramme d'une impulsion de Dirac.

42. Richard Kronland-Martinet, Jean Morlet, and Alexander Grossmann. Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(02):273–302, 1987

en particulier cette capacité à proposer une analyse multirésolution en préservant la propriété d'unicité.

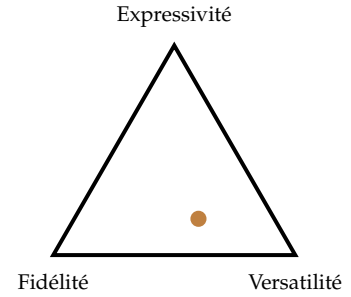
Quelques décennies plus tard, on doit malheureusement faire le constat que l'utilisation de la transformée en ondelettes reste marginale en traitement de l'audio. Dépasser ce constat arbitraire pour amener des éléments concrets nécessiterait une étude approfondie qui sort du cadre de cet exposé. On peut néanmoins avancer quelques éléments de réflexion.

Une première explication peut être que les fonctions de bases de la transformée en ondelettes, *i.e.* les réponses impulsionnelles des filtres passe bande ont des formes particulièrement irrégulières, non adaptées à l'audio. On peut en effet aisément retenir cet argument pour des ondelettes de Daubechies de faible longueur, mais devient moins recevable pour des longueurs plus élevées, ou pour les ondelettes de Morlet. Cet argument pourrait par contre expliquer pourquoi cette transformée à plutôt pris son essor dans le domaine du traitement d'image, où les discontinuités sont beaucoup plus fortes que pour les signaux sonores. En effet chaque contour d'un objet amène une discontinuité importante dans l'espace couleur, qui, dans le cas de la reconnaissance apporte des éléments d'interprétation très importants, et, dans le cas du codage, se doit d'être représenté en détail, sous peine d'engendrer des artéfacts de type « lissage » très pénalisants.

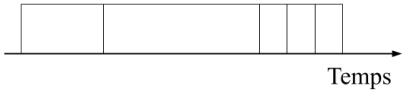
Une seconde explication, finalement peut être plus déterminante, peut avoir trait à la faible expressivité de la transformée. L'interprétation du spectrogramme est intuitive et permet une manipulation fine du contenu spectral d'un signal audio par exemple pour des opérations de masquage présentées dans la section dédiée à l'analyse computationnelle de scènes auditives. L'échantillonnage non régulier en temps et en fréquence de la transformée en ondelettes ne simplifie pas l'expression des transformations que l'on souhaite généralement opérer (transposition, étirement, ...). La version filtrée pour obtenir un échantillonnage régulier en temps est plus appropriée, mais au prix de la perte de toute notion de multirésolution.

En tout état de cause, dans le domaine de l'audio, la grande majorité des systèmes de traitement considère une représentation de type tfct, avec des paramètres de résolution et de précision constants, adaptés à la tâche. Dans les applications de codage, on trouve des mécanismes à résolution multiple de type alterné, implantés grâce des tcdm à plusieurs largeurs de support.[43] Au sein du codeur, une heuristique détermine parmi les résolutions possibles, la résolution la plus adaptée au vu du signal contenu dans l'intervalle d'observation considéré. Cet échantillonnage non régulier de l'axe temporel permet une meilleure fidélité et flexibilité si l'heuristique est adaptée, ceci au détriment de l'expressivité, le plan temps / fréquence induit étant non régulier, ce qui n'est généralement pas un facteur limitant dans des chaînes de codage.

On considère donc ici que dans l'intervalle d'observation, une



La transformée en ondelettes est en théorie assez versatile grâce à la multirésolution, mais peu de travaux démontrent son expressivité.



Analyses fréquentielles effectuées sur des support temporels différents séquencées dans le temps.

43. Karlheinz Brandenburg. Mp3 and aac explained. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society, 1999

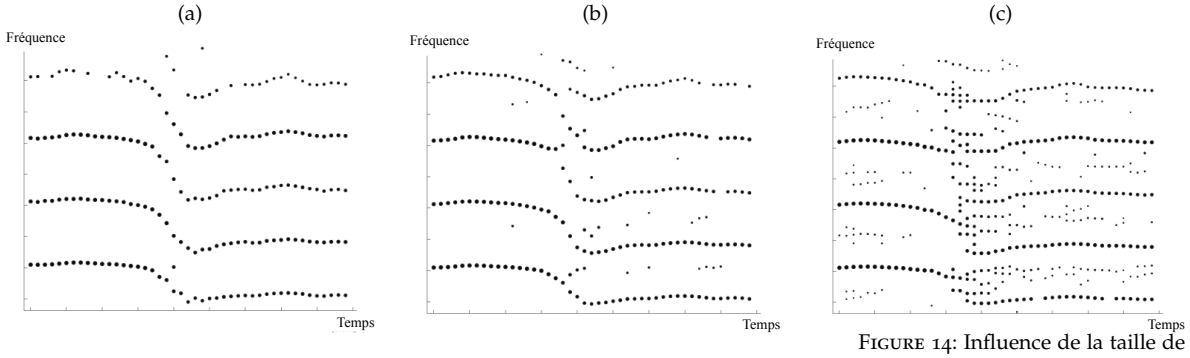
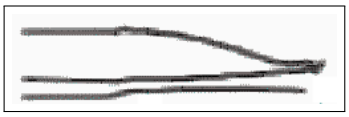


FIGURE 14: Influence de la taille de fenêtre de la tfct utilisée pour estimer un modèle sinusoïdal à court terme d'un glissando de trombone. De gauche à droite, la taille est de 25 (a), 50 (b), et 100 ms (c), pour un pas d'avancement de 10 ms. Chaque point correspond à une composante à court terme $p_{t,c}$ et sa taille est fonction de l'amplitude de la composante $a_{t,c}$.

source sonore est dominante et que la source est structurée soit de manière temporelle soit de manière fréquentielle. Dans le cas d'un contenu polyphonique avec une structure à la fois en temps et en fréquence, cette approche se révèle bien entendu inopérante. La définition d'une méthode d'analyse/synthèse multirésolution reste donc, avec les éléments présentés ici, un champ de recherche encore largement ouvert.

Sinusoïdes à court terme

Lorsque l'on observe le spectrogramme d'un instrument de musique tonal, comme celui d'une note de piano, il apparaît clairement que l'énergie est concentrée en un ensemble réduit de zones fréquentielles. On dit que ces signaux sont parcimonieux en fréquence. C'est, dans une certaine mesure, le cas de la parole. Comme le montrent les exemples de "sine-wave speech", un nombre réduit de sinusoïdes judicieusement placées aux zones de résonances formantiques suffisent à rendre le signal intelligible.



Un exemple de ce type de synthèse de parole est disponible pour écoute : http://webpages.mcgill.ca/staff/Group2/abregm1/web/snd/Track23.mp3.

En suivant cette observation, on peut modéliser chaque zone de forte énergie du plan temps / fréquence par une composante sinusoïdale à court terme. Pour cela, il est commun d'identifier des maxima locaux dans chaque spectre du spectrogramme pour identifier un ensemble de « pics spectraux », ou « atomes » temps / fréquence. On obtient alors le modèle de son suivant :

$$x[n] = \sum_{c=1}^{C_t} a_{t,c} \sin(2\pi f_{t,c} n + \phi_{t,c}) \tag{22}$$

où t et c désignent respectivement l'indice temporel de trame et l'indice de composante dans cette trame et C_t désigne le nombre de composantes par trame t.

Les paramètres de fréquence $f_{t,c}$, d'amplitude $a_{t,c}$, et de phase $\phi_{t,c}$ de chaque atome $p_{t,c}$ peuvent être estimés grâce aux techniques évoquées dans la section précédente, qui se basent sur le spectrogramme. On voit par exemple sur la Figure 14 l'impact de l'augmentation de la taille de fenêtre de la tfct utilisée pour identifier les composantes à court terme. Dans le cas de signaux polyphoniques, on voit qu'il peut être ardu de préserver un bon compromis résolution temporelle / résolution fréquentielle.

L'effectivité du modèle sinusoïdal à court terme est motivée par la parcimonie des signaux sonores. Elle se révèle néanmoins très robuste à la réduction de cette dernière pour peu que le nombre de composantes soit suffisant. En effet, l'expansion de Karhunen-Loève de signaux bruités montre qu'une représentation sinusoïdale de ce type de signaux est valide, à condition que les atomes soient suffisamment nombreux et de fréquences suffisamment proches pour que la densité spectrale de puissance dans ces zones varie lentement dans le temps. Le nombre de composantes par intervalle de temps C_t est donc un facteur déterminant. En pratique, dans un compromis favorisant la fidélité et la versatilité au détriment de l'expressivité, on choisira C_t en fonction de contraintes de complexité, *i.e.* le plus grand nombre possible.

Dans un objectif d'expressivité, il est à contrario important que l'observation soit en adéquation avec le modèle utilisé, c'est-à-dire que les composantes sinusoïdales modélisent effectivement des composantes quasi-périodiques présentes dans le signal analysé. En effet, même si une partie bruitée du signal peut être raisonnablement modélisée avec des composantes sinusoïdales de fréquence proches et de phase arbitraire, il sera difficile avec cette modélisation de préserver une bonne qualité d'écoute lors de la manipulation de ce modèle, par exemple pour effectuer un étirement ou une transposition.

Il convient donc d'être à même de sélectionner parmi les composantes candidates, celles qui sont les plus appropriées en fonction de critères dits de « sinusoïdalité ». Une manière de faire consiste à corréler, pour chaque pic candidat, sa réponse fréquentielle à celle du spectre d'une sinusoïde dont les paramètres sont résultant du processus d'estimation.[44] Le degré de corrélation permet alors de quantifier la pertinence du modèle vis-à-vis de l'observation. Une étude comparative semble montrer l'intérêt de cette approche.[45] Dans cet algorithme, les références au modèle sont construites dynamiquement mais peuvent aussi être échantillonnées. On se rapproche alors du principe des algorithmes de poursuite de sous espaces.

Le principe de ces algorithmes est de s'affranchir des contraintes de la base de Fourier en disposant d'un (très) large dictionnaire d'atomes redondants et de déterminer la combinaison optimale de ces atomes permettant de réduire l'erreur d'approximation. Dans de la modélisation de données audio-numériques, ces atomes sont typiquement des sinusoïdes tronquées ou fenêtrées. L'algorithme "matching pursuit" (mp) proposé par Stéphane Mallat[46] permet de résoudre efficacement ce problème en considérant un algorithme glouton, où l'atome mis à l'échelle ayant la meilleure corrélation avec le signal résiduel (égal au signal d'origine au début de l'algorithme) est itérativement sélectionné puis soustrait. L'application de cet algorithme à des signaux musicaux peut poser des problèmes, dont certains ont été étudiés par Rémi Gribonval.[47]

La Figure 15 illustre deux cas typiques de non stationnarités

44. M. Lagrange, Sylvain Marchand, and J. B. Rault. Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 59–64, September 2002

45. Jeremy J Wells and Damian T Murphy. A comparative evaluation of techniques for single-frame discrimination of nonstationary sinusoids. *IEEE transactions on audio, speech, and language processing*, 18(3):498–508, 2010

46. Stéphane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993

47. Rémi Gribonval, Philippe Depalle, Xavier Rodet, Emmanuel Bacry, and Stéphane Mallat. Sound signals decomposition using a high resolution matching pursuit. In *Proceedings of the International Computer Music Conference*, pages 293–296, 1996

rencontrées dans les signaux musicaux. Le premier consiste en une composante sinusoïdale dont l'amplitude est modulée de manière sinusoïdale, Figure 15(a). C'est une modulation typique d'un tremolo. Le second consiste en une composante sinusoïdale dont l'amplitude est modulée par la multiplication d'une marche d'escalier et d'une exponentielle décroissante, Figure 15(b). C'est une modulation typique d'une corde frappée ou pincée comme le piano ou le clavecin.

Dans le premier cas, l'algorithme mp identifie d'abord une composante avec un support temporel couvrant l'intégralité du signal. Deux autres composantes viennent ensuite annuler certaines parties de cette première. Dans le second cas, le changement abrupt d'énergie étant difficilement modélisable aves les fonctions de base à disposition, de nombreuses composantes d'annulation sont nécessaires, amenant à un effet de pré écho, l'énergie ajoutée n'étant que partiellement enlevée. Pour pallier ces problèmes, des contraintes de non négativité ont été proposées pour améliorer l'algorithme de poursuite, alternative nommée "high resolution matching pursuit" (hrmp). En réduisant les contributions destructives, *i.e.* des atomes sélectionnés pour se soustrairent à des contributions résiduelles d'atomes précédemment sélectionnés, cette approche est plus efficace au sens où l'énergie globale des atomes (somme des enveloppes des atomes utilisés) est plus faible. Dans des approches codage, le fait d'utiliser un nombre de composantes réduit est désirable pour des raisons de contrôle du débit. Cette propriété est, dans une certaine mesure, souhaitable également pour des applications de manipulation car cela permet d'avoir un nombre de paramètres réduit à manipuler. Il est par contre important que le gain en compacité ne se fasse au détriment de l'expressivité.

De manière générale, il est difficile de critiquer les approches basées poursuite dans le sens où leur pertinence vis-à-vis de la tâche va grandement dépendre du dictionnaire choisi. On notera tout de même que ces approches sont construites avec un objectif d'approximation et non d'identification. Par exemple dans le cas de la sinusoïde modulée de manière sinusoïdale, Figure 15(a), les deux algorithmes de poursuites sont tout deux à même de fournir une bonne approximation du signal, mais « n'identifient » pas le phénomène, le dictionnaire n'ayant pas ce type de forme à disposition. Dans une application de type codage, on préfèrera sans doute la seconde approximation pour des raisons d'efficacité de la représentation. Pour ce qui est de la notion d'expressivité, le choix entre les deux modèles seront une affaire de compromis en fonction du type de manipulation souhaité. On se trouve donc avec un compromis de modèle favorisant la fidélité et la versatilité, au détriment de l'expressivité. Il est en effet difficile de déterminer des fonctions simples de manipulation dans le cas d'un dictionnaire très largement redondant.

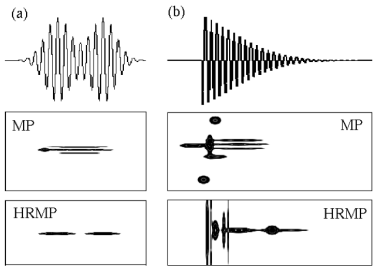
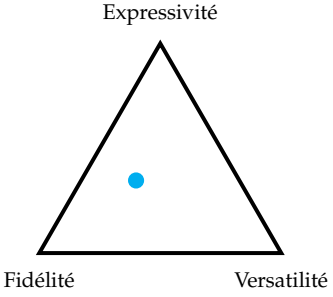


Figure 15: Deux composantes sinusoïdales dont l'amplitude est modulée a) sinusoïdalement, b) exponentiellement et décompositions correspondantes par deux algorithmes de poursuite (mp et hrmp). Figure issue de la référence n° 48.



Les modèles sinusoïdaux à court terme sont des modèles de son avec une expressivité améliorée par rapport au spectrogramme au prix d'une perte relative en fidélité et versatilité.

Sinusoïdes à long terme

L'approche sinusoïdale à court terme effectue des observations discrètes d'un processus de production sonore qui est par essence continu. Dans le cas de nombreux instruments de musique, il est raisonnable de considérer que ce processus continu peut être représenté par une somme de sinusoïdes dont les paramètres varient *lentement* dans le temps. Une variation lente s'entend ici de l'ordre de la dizaine de Hertz. Dans ce cas, le modèle sinusoïdal à long terme est particulièrement à propos. Ce modèle peut s'écrire comme suit :

$$x[n] = \sum_{l=1}^{L} a_l[n] \sin\left(\frac{2\pi}{F_s} f_l[n] \cdot n + \Phi_k\right) \tag{23}$$

où $a_l[n]$ et $f_l[n]$ sont des signaux contrôlant respectivement l'amplitude et la phase des L oscillateurs composant le modèle. Au contraire du modèle court terme de l'équation 22, il est donc possible de modéliser des signaux non stationnaires et s'avère donc pertinent pour modéliser des signaux à long terme.

Chaque composante sinusoïdale est souvent appelée *partiel* pour des raisons historiques. En effet, ce modèle a été appliqué originellement à des sources monophoniques, dont les composantes périodiques étaient appelées partiels, l'ensemble des partiels composant la note. Dans le cas de signaux harmoniques, les partiels sont également souvent appelées harmoniques et la première harmonique, l'harmonique fondamentale.

Ce modèle est particulièrement expressif car il permet des modifications du signal d'intérêt comme la translation, la transposition, ou encore l'étirement, simplement en manipulant les paramètres des partiels. Ces transformations sont de plus de faible coût car les paramètres sont échantillonnés à une fréquence environ 1000 fois plus faible que le signal audio-numérique. Le processus de synthèse est également aisé à mettre en œuvre avec les moyens computationnels actuels. Dans le cas d'un très grand nombre de composantes, des algorithmes d'élagages perceptuels [48] peuvent être mis en place. L'étape d'analyse, *i.e.* l'estimation des paramètres à partir d'un signal sonore est par contre, particulièrement ardue dans le cas général.

Dans notre présentation des algorithmes d'estimation des paramètres long terme, on supposera que le signal analysé est composé d'une ou plusieurs sources dont une grande partie de l'énergie peut être correctement approximée par ce modèle. On fera état ici des approches dites de suivi de partiels, dont le but est de « restaurer » la continuité temporelle entre les composantes à court terme d'un même partiel. Le résultat de cette étape de continuation est une version échantillonnée des paramètres de contrôle nécessaire au modèle de l'équation 23 :

$$\mathcal{P}_p = \left(\mathcal{P}_p[t]\right)_{t \in [b_p, \cdots, b_p + l_p - 1]} \tag{24}$$

où t est l'indice de trame du spectrogramme, b_p est l'indice de

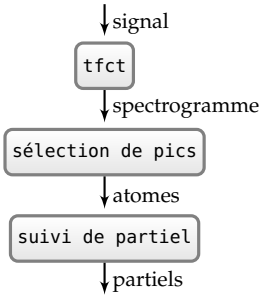


Schéma fonctionnel des approches d'identification de composantes long terme (les partiels) par restauration de continuité. Le signal sonore est projeté sur le plan temps / fréquence où des atomes court terme $p_{t,c}$ correspondant à des maxima locaux sont identifiés. Certains atomes de trames successives et de fréquence proche formeront des partiels $\mathcal{P}_p$.

48. M. Lagrange and Sylvain Marchand. Real-Time Additive Synthesis of Sound by Taking Advantage of Psychoacoustics. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 5–9, December 2001

trame de début du partiel, et l_p le nombre de trame où le partiel est actif. Chaque composante court terme se décrit comme suit :

$$\mathcal{P}_p[t] = \left(n, \phi_p[t], a_p[t], f_p[t]\right) \tag{25}$$

Le passage de cette version échantillonnée à court terme des paramètres à l'échantillonnage signal peut se faire par différentes techniques d'interpolation.

Dans l'article fondateur de George Mc Aulay, [49] il est proposé « d'identifier » ces partiels en reliant entre eux des atomes de trames successives. L'algorithme proposé est itératif. Supposant un ensemble de partiels identifié à la trame t, on va chercher à prolonger ces partiels à la trame $t+1$ en commençant par le partiel de plus basse fréquence, tel que la différence entre la fréquence du partiel $\mathcal{P}_p$ à la trame t et la fréquence d'un atome disponible à la trame $t+1$ est minimale :

$$c_{\min} = \underset{c}{\operatorname{argmin}} \left| f_{t+1,c} - f_p[t] \right| \tag{26}$$

Si cette continuation n'engendre pas un saut en fréquence trop conséquent :

$$\left| f_{t+1,c_{\min}} - f_p[t] \right| < \Delta_f \tag{27}$$

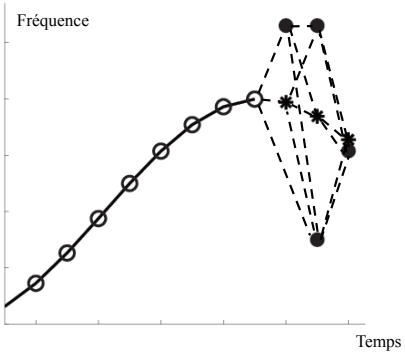
où le seuil en fréquence Δ_f est un paramètre du modèle, alors le partiel $\mathcal{P}_p$ est prolongé à trame $t+1$. Dans un principe d'allocation exclusif, l'atome $p_{t+1,c_{\min}}$ ne pourra pas être utilisé pour la continuation d'un autre partiel. Chaque atome non utilisé à la trame $t+1$ devient alors un partiel qui sera potentiellement prolongé aves des atomes de la trame $t+2$.

On constate ici que la structure de données se complexifie sensiblement par rapport à la plupart de celles introduites avec des outils de traitement du signal canoniques, amenant une approche plus « informatique » à une problématique de traitement du signal. Cette modélisation plus riche a permis une flexibilité de manipulation qui a donné lieu à la proposition d'un large ensemble d'heuristiques qui ne seront pas présentées ici, même si elles ont leur intérêt pour la qualité du rendu sonore de ce type d'approche.

La contrainte exprimée par l'équation 27 revient à avoir un prédicteur constant en fréquence. Dans le cas d'un modèle court terme de bonne qualité, par exemple dans le cas d'un signal monophonique à structure harmonique comme une note de flûte, cette heuristique peut être raisonnable. Dans le cas de signaux non stationnaires comme le glissando de trombone de la Figure 14, de signaux bruités et / ou polyphoniques, il est utile d'améliorer la finesse du prédicteur en considérant un prédicteur linéaire appliqué aux séries temporelles des paramètres long terme. Le faible échantillonnage de ces séries temporelles nécessite des estimateurs adaptés comme celui de Burg. [50]

Dans ce cas, à partir des fréquences du partiel de la trame $t-g$ à la trame t où g est la longueur de la série temporelle des fréquences pris en compte, on obtient par application du prédicteur

49. R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, August 1986



Une étape de continuation du partiel $\mathcal{P}_p$ dans les trames d'indices $t+1$ à $t+l$. La sélection de la meilleure continuation se fait parmi les continuations (traits pointillés) composées des atomes (point noirs) $p_{t+1\ldots l,c}$ de fréquence $f_{t+1\ldots l,c}$ proches des valeurs de fréquence prédites $\hat{f}_p[t+1\ldots l]$ obtenue par prédiction linéaire (étoiles) et ces valeurs prédites.

50. John Parker Burg. A new analysis technique for time series data. *NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands*, 1968

linéaire une prédiction des valeurs des fréquences du partiel pour les trames suivantes : $\hat{f}_p[t+1...l]$, où l est l'horizon temporel utilisé.

La contrainte d'évolution lente des paramètres dans un modèle d'évolution quasi stationnaire se traduit naturellement par un seuil sur le delta de fréquence, voir Equation 27. Dans le cas d'un modèle non stationnaire des paramètres long terme, on trouve un gain à considérer une analyse spectrale de ces évolutions possibles pour mieux déterminer les continuations à effectuer. Pour ce faire, parmi toutes les continuations possibles, la continuation qui engendre le moins de hautes fréquences est privilégiée. Là encore, du fait du faible échantillonnage des paramètres, des outils d'estimation spectrale spécifiques sont nécessaires.

Soit un partiel $\mathcal{P}_p$ que l'on cherche à prolonger dans les trames d'indices $t+1$ à $t+l$. On dispose des prédictions des valeurs de fréquence dans ces trames : $\hat{f}_p[t+1...l]$. Les atomes $p_{t+1...l,c}$ proches de ces prédictions sont sélectionnés. Toutes les successions de valeurs de fréquences sont alors considérées comme des signaux et analysées spectralement. La continuation engendrant le moins de contenu haute fréquence est sélectionnée.

Les bons résultats en terme de rendu sonore obtenu par cette chaîne d'analyse dans le cas de la modélisation long terme sous contrainte de débit [51] et l'interpolation de données manquantes, [52] montrent l'intérêt de la modélisation des modulations long terme, au moins du point de vue perceptif.

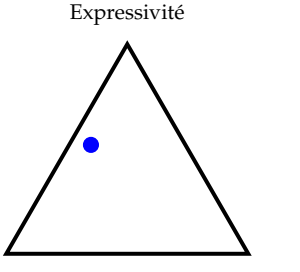
Fort de cette observation, Martin Raspaud a étudié un modèle long terme hiérarchique qui permet des manipulations riche comme par exemple de l'étirement temporel de notes vibrées. [53] Le principe est de considérer les paramètres long terme comme des signaux, eux mêmes modélisables comme des sinusoïdes dont les paramètres évoluent, cette fois ci, très lentement en fonction du temps. Il est alors aisé de contrôler les modulations de haut niveau comme le vibrato ou le tremolo.

D'une grande élégance formelle, d'une bonne expressivité, le modèle long terme n'a pas trouvé un large spectre d'application du fait d'un défaut clair de versatilité. Ce même handicap s'est retrouvé dans le domaine du codage où ce type de modèle n'a pas réussi à concurrencer les approches par transformée. [54] Certaines approches, dites hybrides, consistant à combiner un modèle long terme avec un modèle de bruit et / ou un modèle de transitoire ont été proposées pour tenter de palier à ce défaut. On tombe malheureusement alors dans un problème de sélection de modèles qui est à mon avis intractable. Je ne détaillerai donc pas ces approches.

51. M. Lagrange, S. Marchand, and J.B. Rault. Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds. *IEEE Transactions on Acoustics, Speech and Language Processing*, 2007

52. M. Lagrange, S. Marchand, and J. Rault. Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling. *Journal of the Audio Engineering Society*, 53(10):891–905, 2005

53. Martin Raspaud. *Modèles spectraux hiérarchiques pour les sons et applications*. PhD thesis, Bordeaux 1, 2007



Le modèle sinusoïdal à long terme est expressif et fidèle, mais pour un champ réduit de signaux sonores.

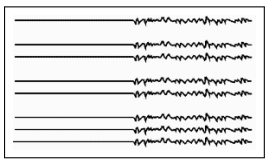
54. Albertus C Den Brinker, Erik Schuijers, and Werner Oomen. Parametric coding for high-quality audio. In *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002

Temps / fréquence / modulations

S'il est clair que la décomposition du signal sonore le long de l'axe fréquentiel est de première importance, la manière dont l'amplitude et la fréquence sont modulées au cours du temps au sein de ces bandes de fréquence l'est également. On peut citer pour

exemple le son éminemment désagréable d'un moteur thermique deux temps (type mobylette ou tronçonneuse).

L'importance perceptive de ce type de modulations dans notre jugement du réalisme d'un son est particulièrement bien exemplifiée par un son synthétique produit par John Chowning. Dans ce signal, des sinusoïdes d'amplitude et fréquence au début constantes sont ensuite modulées par un signal synthétique composé d'une composante sinusoïdale (pour le vibrato) auquel se sur-ajoute une composante stochastique. Ce signal modulant est ensuite mis à l'échelle pour chaque harmonique. Il est important de considérer que même si ce signal modulant a été judicieusement choisi, il n'est pas issu d'un processus d'estimation effectué à partir de signaux réels. L'écoute montre clairement que le caractère naturel, voisé, est sensiblement plus élevé avec l'ajout de ces modulations.



Ce son est disponible pour écoute : http://webpages.mcgill.ca/staff/Group2/abregm1/web/snd/Track24.mp3

Modèles perceptifs

Complétons nos connaissances en prenant un point de vue physiologique. L'étude des éléments physiologiques en charge du traitement de ces modulations est relativement récente car assez complexe à investiguer. On a néanmoins une bonne confiance dans le fait que le système auditif des mammifères dispose d'éléments de traitement dédiés à cette tâche. On peut notamment citer le modèle de Torsten Dau,[55] qui en plus de la traditionnelle analyse en fréquence effectuée au niveau de la cochlée ajoute une analyse de la modulation d'amplitude. On a donc un modèle temps / fréquence / taux de modulation. Celui de Shihab Shamma[56] étend le modèle de Torsten Dau en ajoutant une quatrième dimension : l'échelle, correspondant aux degrés de sélectivité des filtres de modulation.

Pour parvenir à ce modèle, Shihab Shamma a étudié certains neurones du cortex auditif primaire du furet, plus particulièrement ce que l'on appelle leur champs de réponse spectro / temporellle ou "spectro-temporal receptive fields" (strf). L'animal est contraint au niveau de la tête, et par opération chirurgicale, une aiguille permettant de mesurer l'activité électrique est insérée dans le crâne de l'animal et judicieusement placée. Par exposition d'un ensemble de stimuli comportant des modulations, il est alors possible de mesurer les strf de certains neurones du cortex auditif. Au vu de ce type de réponses, un modèle de traitement de signal est ensuite proposé.

Dans ce modèle, le signal est donc tout d'abord décomposé sur l'axe fréquentiel. La sortie est un équivalent du spectrogramme qui est ensuite décomposé une nouvelle fois fréquentiellement en temps et en fréquence grâce à des ondelettes bi-dimensionnelles paramétrées avec un facteur d'échelle. La dimension de l'échelle permet de définir le nombre de bandes qui sont conjointement utilisées pour effectuer cette seconde décomposition.

Ces modèles sont bien entendu sujet à débat dans la communauté de neurophysiologie. Par exemple, le facteur d'échelle n'est

55. Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5): 2892–2905, 1997

56. Jonathan Fritz, Shihab Shamma, Mounya Elhilali, and David Klein. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience*, 6 (11):1216, 2003

pas nécessaire dans le modèle de Torsten Dau pour expliquer les données recueillies grâce à son protocole expérimental.

Scattering d'ondelettes

En prenant un point de vue mathématique cette fois-ci, Stéphane Mallat a proposé un modèle conceptuellement proche du modèle computationel des strf proposé par Shihab Shamma. En mathématiques appliquées, et plus particulièrement dans le domaine des sciences des données, on cherche à construire des représentations qui aient de bonnes propriétés d'invariance et de stabilité aux déformations. C'est-à-dire que l'on souhaite que pour une représentation $\Phi(x)$ d'un signal x et d'un signal déformé $\tilde{x}$:

— $\Phi(\tilde{x}) = \Phi(x)$ (invariance)

— $|\Phi(\tilde{x}) - \Phi(x)| < \epsilon$ (stabilité)

où ϵ désigne une petite valeur.

Précisons ces notions en prenant quelques exemples. Lorsque l'on cherche à reconnaître l'instrument de musique qui a été utilisé pour jouer une note, on souhaite généralement disposer d'une représentation du signal de cette performance qui soit invariante à certains aspects de la performance comme :

1. un changement d'amplitude ;
2. un décalage en temps ;
3. un étirement en temps.

En effet, pour la tâche pré-citée, que la note ait été jouée plus ou moins proche du microphone,[57] à un instant donné ou quelques secondes plus tard, à une hauteur donnée ou une autre, pour une durée plus ou moins longue ne devrait pas modifier de manière trop conséquente notre représentation.

57. J'écarte ici pour la simplicité du discours les aspects de nuances qui ont une influence non négligeable sur le timbre.

Sous hypothèse de stationnarité, l'invariance au changement de volume est localement obtenue en normalisant les signaux que l'on souhaite comparer. L'invariance locale à la translation en temps se traduit dans le domaine de Fourier en une invariance à la translation de phase par la prise du module. On souhaite généralement que l'invariance au volume et à la translation soit totale. En ce qui concerne l'invariance à l'étirement temporel, on souhaite généralement plutôt que la représentation soit stable à la déformation.

Il est clairement explicité par Joakim Andén[58] pourquoi le spectre de magnitude de Fourier d'un signal harmonique n'est pas stable à même un petit étirement temporel. En effet, l'impact de l'étirement selon l'axe des fréquences n'est pas le même. Si un petit étirement déplacera d'une petite quantité les fréquences basses, les fréquences plus élevées le seront beaucoup plus. Prenons l'exemple d'un signal harmonique de fréquence fondamentale 440 Hz, un étirement temporel va réduire la valeur de cette fréquence à disons 430 Hz pour un delta de 10 Hz. Ce delta sera n fois plus grand pour la $n^{\text{ième}}$ harmonique soit 100 Hz pour la dixième harmonique. Pour une transformée de Fourier avec une quantification

58. Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16): 4114–4128, 2014

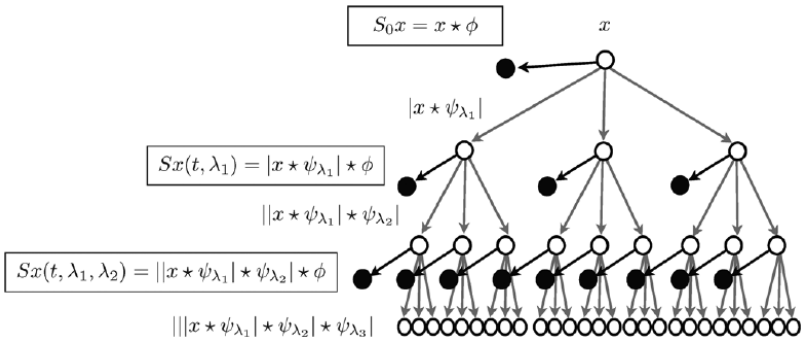


FIGURE 16: Propagation du signal au travers du scattering d'ondelettes au premier et second ordre (figure issue de la référence nº 59).

de l'axe fréquentiel d'une dizaine de Hertz, la différence entre les deux spectres va croître très rapidement en fonction du delta car l'énergie des harmoniques supérieures va très rapidement se placer sur des pas de quantification différents de ceux où l'énergie des harmoniques se plaçait initialement.

Pour être plus stable à ce type de déformation, il convient alors de « délocaliser » les hautes fréquences de manière plus conséquente que les basses. Suivant les communautés, on obtient cette délocalisation progressive sur l'axe des fréquences avec des transformées à Q constant, de type tiers d'octave (acoustique), des mels ou des Barks (parole).

En se plaçant dans le cadre du scattering d'ondelettes, dont le schéma fonctionnel est montré sur la Figure 16, on obtient ce type de représentation en délocalisant en temps toutes les bandes de fréquence du module de la transformée en ondelettes ψ_1 par l'application d'un filtre moyennant ϕ pour obtenir les coefficients de scattering au premier ordre S_1 :

Ce résultat est, à mon sens, d'importance, car il permet de placer un cadre mathématique pour mieux expliquer l'utilité d'un axe fréquentiel logarithmique qui a été empiriquement vérifié dans de nombreux domaines du traitement de l'audio et motivé jusqu'ici par des arguments essentiellement neuro mimétiques.

$$S_1 x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi(t). \tag{28}$$

Ceci induit une perte d'information qu'il convient de compenser. Le principe du scattering d'ondelettes permet de répondre élégamment à ce problème en cascadant des opérations de décomposition et en utilisant chaque niveau de décomposition pour produire une représentation compacte et informative. Ainsi, on obtient au second ordre :

$$S_2 x(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t). \tag{29}$$

Comme le montre la Figure 16, le scattering dissocie décomposition et représentation, ce qui permet de préserver certaines propriétés d'intérêt comme, par exemple, l'inversibilité pour la décomposition sous certaines conditions et la stabilité pour la représentation.

La sortie du filtre moyenneur au premier ordre est donc équivalente à une transformée à Q constant. En décomposant les sorties rectifiées du premier banc de filtres en ondelettes ψ_{λ_1} avec un autre banc de filtres en ondelettes ψ_{λ_2}, on capte alors l'information de

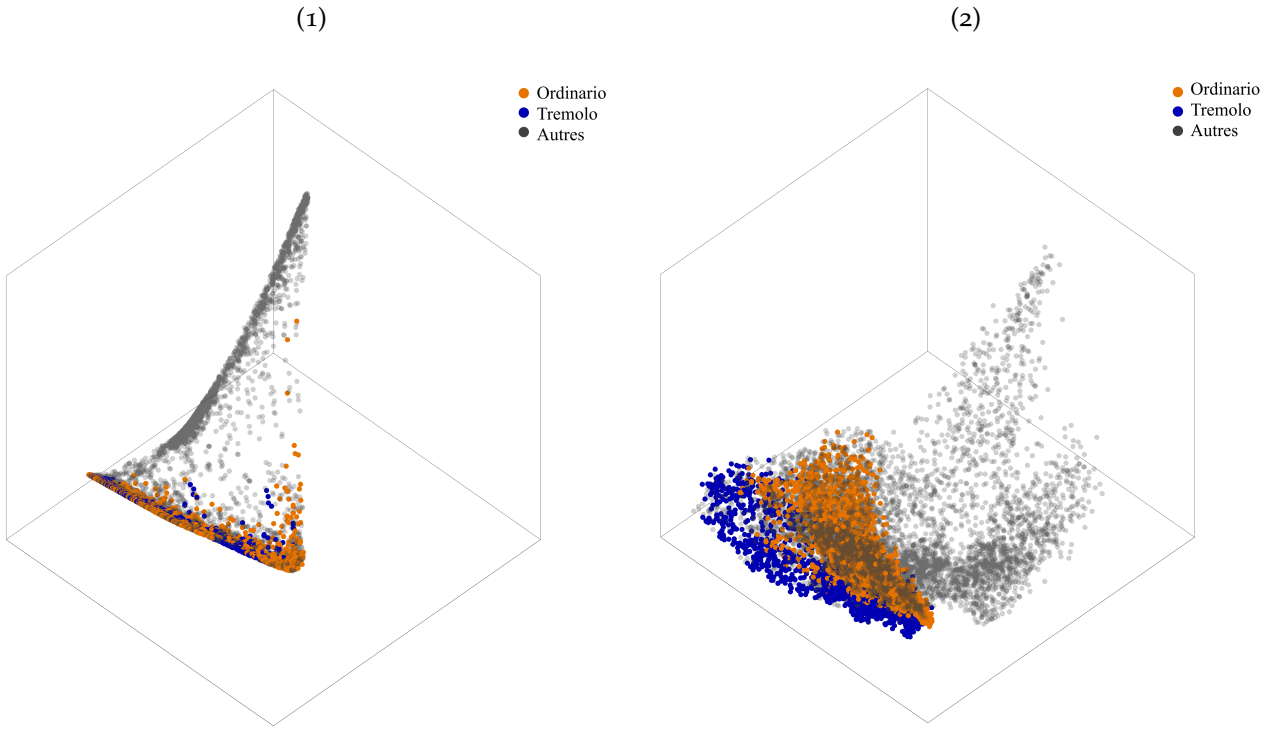


FIGURE 17: "Diffusion maps" obtenues à partir du scattering d'ondelettes à l'ordre 1 et 2.

modulation d'amplitude, *i.e.* comment l'énergie fluctue dans une même bande de fréquence. On a vu précédemment l'importance de ce type de modulation pour la perception. Il est donc opportun de d'étudier l'apport du second ordre pour une représentation riche du signal sonore.

Son apport a été démontré expérimentalement pour de nombreuses tâches de modélisation. Nous avons complété cet effort de recherche en considérant des sons d'instruments de musique du répertoire avec des modes de jeux étendus. [59] En effet, si la reconnaissance de l'instrument est considérée comme un problème résolu par une représentation à Q constant, la reconnaissance du mode de jeu nécessite d'être à même de modéliser des éléments de modulations qui ne sont pas aisément capturés par une représentation à Q constant ou un scattering d'ondelettes à l'ordre un. En considérant la méthode des "diffusion maps" [60] pour obtenir un espace de projection à 3 dimensions à partir respectivement d'une représentation à Q constant et du scattering d'ondelettes à l'ordre 2, on note par exemple que le mode de jeu *tremolo* est mieux distingué du mode de jeu ordinaire dans la seconde projection, voir Figure 17.

On est donc en mesure de correctement modéliser les modulations d'amplitude. Pour permettre de modéliser des modulations sur l'axe fréquentiel, il est nécessaire de considérer des ondelettes bi-directionnelles sur le plan temps / fréquence, ce qui rend le modèle conceptuellement plus proche du modèle perceptif de Shihab Shamma. [61]

En terme d'architecture, le scattering d'ondelettes suit une orga-

59. Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange. Extended playing techniques: the next milestone in musical instrument recognition. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, pages 1–10, 2018

60. Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005

61. J. Andén, V. Lostanlen, and S. Mallat. Joint time-frequency scattering. *IEEE Transactions on Signal Processing*, 67(14): 3704–3718, July 2019

nisation de type réceptive, 1) fréquence et 2) temps ou fréquence / temps ensuite. On verra dans la partie suivante un modèle prenant une approche de type source / filtre donc temps puis fréquence et qui est à ce titre plus proche d'un modèle de production.

Synthèse modale

En se référant plutôt aux processus physiques de production sonore, les approches modales permettent de proposer des modèles qui sont à la fois efficaces et effectifs. Ces modèles se basent sur un modèle de production de type source / filtre, où la source est généralement décrite en terme de propriétés temporelles, et le filtre en terme de propriétés fréquentielles. Pour un signal de parole, on considèrera que la source peut être soit un train d'impulsions résultant de l'ouverture périodique des plis vocaux pour la production des sons voisés (voyelles) soit un signal stochastique résultant de la compression du flux respiratoire pour la production des plosives ou sifflantes (consonnes).

Le modèle source / filtre décorrèle donc naturellement la contribution de l'excitateur et celle du résonateur, ce qui est particulièrement attrayant lorsque l'on recherche un modèle de son permettant une bonne capacité d'interaction. Pour peu que les modèles utilisés pour les deux parties (source et filtre) soient bien adaptés, on obtient des modèles de sons qui sont souvent fidèles et particulièrement expressifs. [62]

Nous avons étudié ce type de modèle pour la synthèse de sons de roulements d'une bille sur une plaque. Ces sons présentent un challenge pour ce type de synthèse car l'interaction entre la bille et la surface sur laquelle elle roule est complexe. Néanmoins, ce type de modèle adapté au problème [63] a donné des résultats encourageants. [64]

Dans cette contribution, l'étape d'analyse consiste à identifier les composantes résonantes paramétrées par leur fréquence de résonance f_k et leur amplitude a_k et les utiliser pour opérer une déconvolution du signal enregistré $s[n]$. Le résiduel $r[n]$ de cette déconvolution permet d'identifier une forme typique $i[n]$ de l'impact de la bille avec la plaque la position des impacts. Cette forme est alors utilisée pour déconvoluer $r[n]$ et ainsi obtenir un signal $d[n]$ essentiellement composé de pics indiquant la position et la puissance des impacts de la bille sur la plaque.

La source est modélisée comme une série d'impacts dont l'amplitude est modélisée par une distribution exponentielle paramétrée par λ_a et l'intervalle de temps entre deux impacts par une distribution gamma paramétrée par k_i et θ_i. Ces paramètres sont estimés à partir de $i[n]$.

En remarquant fort justement que, lorsque la bille impacte fortement la plaque sur laquelle elle roule, l'intervalle entre cet impact et le suivant augmente, Simon Conan a proposé un modèle amélioré où ces variables sont modélisées de manière conjointe. [65]

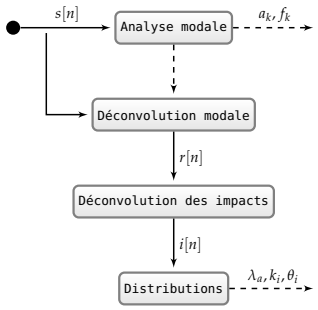


Schéma de traitement de la phase d'analyse d'un modèle de synthèse modale pour les sons de roulement. Les traits pleins désignent des signaux, les traits pointillés désignent des paramètres.

62. Mitsuko Aramaki and Richard Kronland-Martinet. Analysis-synthesis of impact sounds by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2): 695–705, 2006

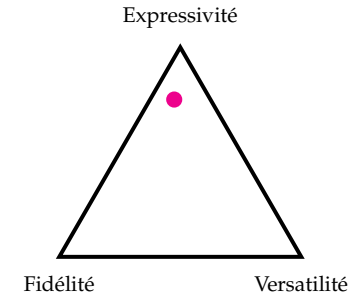
63. Mathieu Lagrange, Gary Scavone, and Philippe Depalle. Analysis / synthesis of sounds generated by sustained contact between rigid objects. *IEEE Transactions on Audio Speech and Language Processing*, 18-3:509–518, 2010

64. Emma Murphy, Mathieu Lagrange, Gary Scavone, Philippe Depalle, and Catherine Guastavino. Perceptual evaluation of rolling sound synthesis. *Acta acustica united with Acustica*, 5-97: 840–851, 2011

65. Simon Conan, Olivier Derrien, Mitsuko Aramaki, Solvi Ystad, and Richard Kronland-Martinet. A synthesis model with intuitive control capabilities for rolling sounds. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(8):1260–1273, 2014

Ce type d'approche se situant à l'intersection entre modélisation statistique et modélisation physique possède deux propriétés d'intérêt, avec en premier lieu, la causalité. Au niveau du modèle de synthèse, un échantillon est strictement considéré comme une composante des échantillons précédents. En second, cette approche met l'accent sur les propriétés de filtrage des composantes modales, il ne s'agit pas ici de synthétiser un signal riche à partir d'une représentation complète du signal, mais plutôt d'exploiter un modèle source / filtre avec une implantation dédiée de ces deux modules en fonction de la tâche visée.

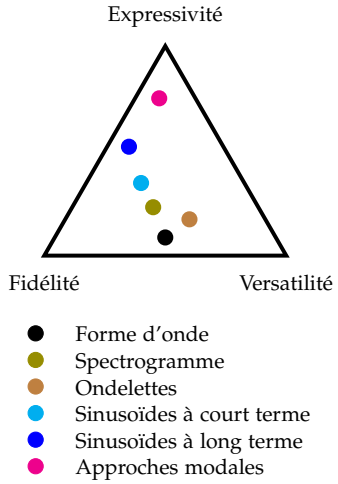
Malgré toutes ces bonnes propriétés, de manière peut-être encore plus marquée que pour les modèles sinusoïdaux à long terme, la versatilité reste malheureusement en deçà, car le changement de structures pour la source et le filtre ainsi que leurs modes d'interactions nécessitera le plus souvent des adaptations d'ordre computationel si ces changements sont conséquents. En particulier, les interactions non linéaires sont difficiles à modéliser en raison de manque d'outils d'estimation génériques.



Le modèle modal est particulièrement expressif et fidèle pour un champ assez réduit de signaux sonores.

Discussion

Si l'on prend une vue d'ensemble des différents compromis pris par les modèles de sons précédemment décrits, on voit qu'il existe des compromis équilibrés entre fidélité et expressivité au détriment de la versatilité et des compromis équilibrés entre fidélité et versatilité au détriment de l'expressivité, avec un « juste milieu » occupé par le spectrogramme. Le grand défi de la modélisation sonore est donc de proposer des approches qui puissent se placer dans un compromis entre versatilité et expressivité avec une fidélité qui reste raisonnable.



	fidélité	versatililité	expressivité
multirésolution	●	●	
causalité	●		●
non linéarité	●	●	
dimensionalité réduite			●
contrôle lent			●

Au vu des arguments discutés durant cette présentation, quelques éléments se révèlent incontournables pour espérer relever ce défi. Tout d'abord, le modèle doit inclure des aspects de **multirésolution** pour les raisons discutées dans ce chapitre. Il me paraît également incontournable que le modèle soit **causal**. La non causalité permet certes une plus grande flexibilité algorithmique et facilite souvent l'identification. Pour des applications de génération en revanche, la perte de la causalité amène malheureusement des approximations qui sont souvent détrimentaires à la fidélité, car le processus physique qui a produit le signal sonore est lui causal.

Ensuite, de nombreuses étapes de la propagation des ondes so-

nores peuvent être considérés comme des systèmes linéaires. Néanmoins, il s'avère critique pour la qualité perceptive des signaux produits que le modèle puisse également modéliser des processus **non linéaires** pour accéder à une grande versatilité. Cela pose bien entendu le problème du contrôle, les structures algorithmiques non linéaires étant plus complexes à manipuler. Pour des raisons d'expressivité, il est ensuite nécessaire que le contrôle soit aisé, ce qui amène d'autres contraintes en terme d'espace de description. Il est souhaitable qu'il soit d'une **faible dimensionalité** et que son **échantillonnage temporel** soit sensiblement plus **lent** que celui du signal.

La plupart des outils présentés dans ce chapitre relèvent du domaine du traitement du signal. Il est néanmoins remarquable que, pour la plupart des contributions récentes en sciences des données, les éléments déterminants au progrès des systèmes relèvent principalement de l'apprentissage automatique. Même si ces disciplines partagent de nombreux fondements mathématiques en terme de modélisation de l'information, il est important de bien comprendre les différences de méthodologie. Ce sujet passionnant nécessiterait une étude complète, je me contenterai donc ici d'une caricature pour les besoins du propos. Dans une approche traitement du signal, l'expert met en place une structure algorithmique avec un nombre de degrés de liberté réduit qui lui permet d'envisager un paramétrage manuel ou par le biais d'un plan expérimental relativement réduit. Dans l'approche apprentissage, le nombre de degrés de liberté devient tellement grand que les mécanismes d'optimisation automatique deviennent indispensables. Il est bien entendu que ces méthodes d'optimisation ont leurs paramètres, qui doivent être également choisis avec soin.

La place de l'apprentissage dans les chaînes de prise de décision n'a fait que grandir ces dernières années. Ce mouvement sonne-t-il le glas du traitement du signal ? En terme de composant algorithmique indépendant c'est en effet probable. En terme de discipline, de nombreuses opportunités sont ouvertes. En effet, les avancées obtenues grâce aux mécanismes d'apprentissage relèvent pour beaucoup de démarches inductives. A ce jour, la compréhension et la formalisation de la manière dont l'information est traitée dans ces structures algorithmiques complexes est donc largement incomplète.

Finalement, qu'il s'agisse de modèles de perception, de modèles d'apprentissage, ou de modèles de production, « tout est filtrage ». Il est pour cette raison prédictible que les contributions à l'interface de ces deux disciplines constitueront des avancées significatives aux défis relevant du domaine des sciences des données. [66]

Dans cette quête, l'étude des signaux temporels comme les signaux sonores permet de se confronter aux besoins pré-cités : multirésolution, causalité, non-linéarité. Ces verrous sont au cœur des problématiques identifiées par la communauté des sciences des données. Par exemple, les avancées récentes par réseaux de

66. Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016

neurones convolutionels récurrents pour la synthèse de signaux sonores basés échantillons [67] proposent des architectures qui possèdent de manière remarquable les conditions nécessaires pré-citées et devraient obtenir en théorie un excellent compromis entre fidélité et versatilité. La possibilité de conditionner le réseau de manière assez flexible donne en outre de bons espoirs en terme d'expressivité.

Est-ce que l'intérêt de ce type d'architecture se confirmera en amenant des réponses aux défis identifiés dans ce chapitre ? Répondre à cette question ouvre un champ d'expérimentation passionnant pour l'expérimentateur qui se doit d'être guidé par une approche méthodologique rigoureuse, et fondée sur des questionnements scientifiques clairement explicités.

67. Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499

La méthode scientifique en sciences des données

Le canon en matière de méthodologie pour l'étude d'un sujet particulier nous est offert par la méthode scientifique. Issue de pionniers comme Francis Bacon,[68] elle permet, par son bon usage, de faire une étape importante dans la qualité de notre appréhension du monde.

Si l'on cherche à remonter plus loin dans le passé, il est remarquable de constater que l'apport des penseurs grecs et romains qui est parvenu jusqu'à nous porte essentiellement sur la nature et la définition de la vérité et finalement assez peu sur la bonne manière de l'obtenir. On trouve en revanche des éléments en Orient sur cette période, comme le rapporte le *Kālāma Sutta*.

Ce texte discute de deux éléments d'intérêt : 1) son exposé simple du caractère fondamentalement mal posé de la quête de la vérité, 2) le caractère résolument pragmatique de sa résolution de la « tension sceptique ». Face à l'absence de référentiel non discutable permettant de confirmer ou d'infirmer une thèse, même après tout le soin que l'on peut apporter à notre argumentation, et seulement si l'on peut démontrer son intérêt pour la communauté, on peut choisir de vivre avec si tel est notre souhait, ce choix restant éminemment personnel.

C'est néanmoins l'observation minutieuse et l'analyse rigoureuse qui va fournir le fondement de cette prise de décision. Il est donc capital de disposer de protocoles solides pour réaliser nos expériences. Et il est pour cela important de prendre conscience que la manière dont on considère une pratique impacte nécessairement cette pratique. Cet aller et retour entre (i) pratique expérimentale et (ii) observation critique de cette pratique, constitue un facteur qualitatif déterminant pour le succès de cette quête.

Cela s'incarne pour moi très bien dans la relation encadré / encadrant, par exemple dans le cas d'une thèse de doctorat. Le doctorant, poussé par la force de la jeunesse, dispose d'un potentiel conséquent et d'un enthousiasme qui lui permet d'avancer malgré les nombreuses déconvenues qui sont le quotidien du chercheur. Sans vision globale de l'état des connaissances de la communauté et sans vision critique, de haut niveau, du travail effectué, il y a de fortes chances pour que ses « échecs » restent les conclusions de ses efforts. C'est la fonction importante de l'encadrant que de mettre en perspective le travail effectué et de le placer judicieusement pour apporter de la connaissance à la communauté visée.

68. Francis Bacon. *Novum organum*. Clarendon press, 1878

Question : « Maître, comment savoir si quelque chose est vrai ou faux ?
Comment savoir ce qu'il faut croire ? »

Réponse : « Ne croyez pas en quelque chose simplement parce que vous l'avez entendu.
Ne croyez pas aux traditions simplement parce que elles se sont transmises depuis de nombreuses générations.
Ne croyez pas en quelque chose simplement parce que tout le monde en parle.
Ne croyez pas en quelque chose simplement parce qu'on le trouve dans vos livres religieux.
Ne pas croire en quelque chose simplement en vous basant sur l'autorité de vos enseignants et de vos aînés.
Mais après une observation minutieuse et une analyse rigoureuse, Quand vous trouvez que quelque chose est en accord avec la raison, Et qui est propice au bien et au bénéfice de l'un et de tous, Alors vous pouvez l'accepter et vivre en fonction. »

Question : « Alors, Maître, cela signifie-t-il que nous ne devrions pas nécessairement croire ce que vous venez de répondre ? »

Kālāma Sutta, Gautama Bouddha, 500 avJc

Quels sont les savoirs nécessaires pour critiquer de manière constructive un cycle de la méthode scientifique ? C'est bien connu, acquérir des connaissances sur une pratique peut se faire en pratiquant, si possible avec des personnes reconnues dans cette pratique. Il est sans doute vain de vouloir formaliser complètement ces savoirs, néanmoins, ne serait ce que pour mieux les maîtriser et les transmettre, je me suis intéressé à ce problème pendant plusieurs années.

Les solutions que j'évoque dans ce chapitre m'ont été particulièrement utiles. Je pense qu'elles pourront profiter aux communautés scientifiques auxquelles j'ai eu le plaisir de contribuer et plus largement dans ce que l'on appelle aujourd'hui « les sciences des données », où j'ai pu observer des phénomènes récurrents qui sont, à mon sens, néfastes pour le bon accroissement de la connaissance. Je citerai par exemple :

1. l'absence de formalisme expérimental canonique et rigoureux ; ce qui permet d'orienter et de justifier son effort de recherche à peu de frais ;
2. l'ajustement aux données, par exemple par méta paramétrisation ah hoc des approches algorithmiques proposées ou de l'utilisation de composants algorithmiques hétérogènes pour améliorer les performances ; ce qui peut entraîner la production de données quantitatives coïncidentes et des conclusions qualitatives dont les bases expérimentales sont fragiles.

Il n'est pas question d'accuser l'ensemble de la communauté de « piper les dés » de manière ouverte. Ce sont des problèmes profonds qui ont leurs sources dans le fait que la recherche est effectuée par des être humains dans un contexte social donné, avec des moyens qui sont ce qu'ils sont et qui sont alloués de la manière dont ils sont alloués. Par exemple, la façon dont un chercheur doit gérer sa carrière dans le contexte actuel peut poser des problèmes qui relève de l'éthique et peuvent avoir une influence majeure sur sa production. J'aborderai cette question dans la dernière section de ce chapitre dédiée à la revue par les pairs.

J'aimerai prendre ici pour exemple le travail remarquable de Y-Lan Boureau dans le domaine du traitement de l'image. [69] Cette étude analyse rigoureusement l'importance de différentes composants d'une chaîne de traitement d'image, en particulier les étapes de codage et d'union. La conclusion de cette étude est la suivante : "While much research has been devoted to devising the best possible coding module, our results show that with linear classification, switching from average to max pooling increases accuracy more than switching from hard quantization to sparse coding. These results could serve as guidelines for the design of future architectures."

Pour bien comprendre l'importance de cette conclusion, il est important de placer un peu de contexte. Nous sommes en 2010, le terme de parcimonie fait couler beaucoup d'encre et de nombreuses

69. Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2559–2566. Citeseer, 2010

« Bien que de nombreuses recherches aient été consacrées à la conception du meilleur module de codage possible, nos résultats montrent qu'avec la classification linéaire, passer d'une étape d'union par la moyenne à une étape d'union par le maximum augmente davantage la précision que le passage de la quantification classique au codage parcimonieux. Ces résultats pourraient servir de lignes directrices pour la conception de futures architectures. »

approches utilisant ce terme comme étendard sont proposées. Cet engouement est le résultat d'une combinaison complexe de biais cognitifs, de structures sociales et idéologiques, qui amènent incidemment à concentrer l'effort de recherche d'une communauté vers ce qui semble être la bonne marche à suivre. Ce fait étant considéré comme acquis, combien d'études se comparent à un algorithme de référence ? Qui se pose la question de savoir si les gains obtenus à cette étape de la chaîne de traitement sont potentiellement intéressants par rapport à d'autres éléments ?

Le fait est que les travaux expérimentaux de synthèse comme celui discutés ici sont rares, car ils sont coûteux en temps et finalement assez peu « excitants », pour celle ou celui qui conduit l'étude, mais surtout pour la communauté à qui ces travaux sont destinés. Avoir à remettre en question ses orientations de recherche, ses inclinaisons, ses biais, n'est pas une tâche facile, ni agréable. [70]

On peut donc supposer qu'il est nécessaire de pouvoir comparer simplement et efficacement de nombreuses approches différentes ; d'utiliser un même formalisme expérimental qui soit rigoureux et validé par la communauté. Les deux questions qui se posent alors, auxquelles je vais essayer de répondre à la lumière de ma participation à l'organisation de challenges de comparaison d'algorithmes, sont les suivantes : 1) est-ce suffisant ? 2) quelles sont les précautions à prendre ?

Pour répondre rapidement à la première question, je dirais que non, ce n'est pas suffisant, car la question du sens est toujours présente et doit être traitée avec des outils méthodologiques qui relèvent du questionnement scientifique. Sans prendre conscience de ce dernier point, on risque de perdre toute opportunité de faire croître la connaissance.

Les challenges tels qu'ils sont pratiqués aujourd'hui sont bien entendu une avancée incontestable par rapport à l'approche « mon problème, ma base de données, mon algorithme, ma métrique » qui peut avoir son intérêt dans des phases de défrichage thématique mais qui ne peuvent être raisonnablement considérée une fois la problématique bien identifiée. Malheureusement, la mise en compétition met en exergue l'approche « la fin justifie les moyens ». Au vu des objectifs le plus souvent fixé par la plupart des challenges, cela est tout à fait de bon aloi.

Pour répondre à la deuxième question, je prendrais l'exemple du "music information retrieval" (mir) où il est troublant de constater qu'une infime partie des publications font état de travaux effectués conjointement avec des musicologues. Or, qui connait mieux la matière que ceux qui la questionnent depuis des centaines d'années ? Certes, un dialogue est nécessaire car les deux communautés sont assez disjointes, autant en terme d'objectifs, que de vocabulaire et de temporalité.

Il est à mon avis néanmoins critique pour l'avenir de toute discipline portant sur la science des données de ne pas oublier d'où viennent ces données, que des communautés scientifiques dis-

70. Cette difficulté à se remettre en question, que ce soit au niveau individuel qu'au niveau de la communauté peut amener des difficultés considérables pour l'expression du chercheur. Je citerai ici pour exemple la réponse d'un éditeur associé d'une revue de référence dans le domaine du traitement du signal audio-numérique à la soumission d'une étude de réplication effectuée par Bob Sturm et ses collègues ; réplication d'une étude publiée dans cette même revue. La réponse disait en substance que cette étude de réplication était rigoureuse et non criticable sur le fond, mais que la revue ne souhaitait pas publier des études qui mettaient en défaut des études publiées dans cette même Revue.

La différence de temporalité inter communauté constitue dans mon expérience la contrainte plus difficile à mitiger lorsque q'un chercheur en ingéniérie cherche à collaborer avec des spécialistes de la perception ou de la cognition.

posent de savoirs vivants sur le sujet et qu'entrer en contact avec les membres de ces communautés peut nous permettre de poser des questions qui ont du sens, et éviter de ne pas incidemment redécouvrir des évidences « oubliées ».

Au vu de ces arguments, si la question est de savoir si la conduite d'un protocole expérimental rigoureux suffit à produire de la connaissance de qualité, la réponse est négative. Elle constitue néanmoins à mon sens une condition nécessaire pour laquelle un bagage solide de connaissances est à notre disposition pour améliorer la qualité de nos productions.

Il m'est donc apparu comme important, en parallèle de mon implication à l'interface sciences des données / perception, cognition, de revenir aux bases de la méthode scientifique pour étudier la faisabilité d'une recherche expérimentale qui soit plus rigoureuse et plus efficace.

Méthodologie expérimentale

Plaçons nous dans le cadre de la méthode scientifique moderne. Elle constitue un processus itératif et cyclique dans lequel nos connaissances sont continuellement révisées à l'aide d'un algorithme constitué des composants suivants :

— **Caractérisations** : observations, définitions et mesures ;

— **Hypothèses** : explications théoriques ou hypothèses d'observations ;

— **Prédictions** : raisonnement inductif et déductif à partir de l'hypothèse ou de la théorie ;

— **Expériences** : évaluation de la pertinence de chacun des éléments précédents.

— **Revue** : mise à disposition de la communauté des ressources nécessaires à la compréhension et la capacité de réplication des expériences.

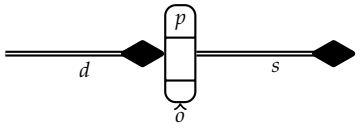
Il est difficile de produire des recommandations qui soient pertinentes pour l'intégralité des étapes de la méthode scientifique. En effet, les trois premiers composants relèvent de l'état des connaissances et des « directions » thématiques prises par une thématique donnée. Chaque communauté scientifique doit réfléchir aux axes à privilégier, à l'allocation des ressources, etc. J'ai donc fait le choix de consacrer ma réflexion à la partie expérimentale de la méthode scientifique, les autres relevant plus d'un choix ou d'une inclinaison culturelle et politique.

Pour la partie expérimentale de la méthode, on dispose en revanche d'un ensemble fourni de méthodes et d'outils pour mener à bien cette étape. Ces connaissances nous proviennent essentiellement de la science du vivant (médecine et biologie) où les conclusions tirées des études effectuées ont, depuis longtemps, un potentiel important. L'impact probable qu'auront les sciences des don-

nées sur l'humanité dans les décennies qui viennent nous incitent fortement à faire de même.

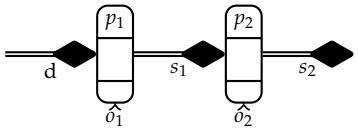
Même si la tâche de systématisation est moins ambitieuse pour le composant expérimental, il est bien entendu illusoire de pouvoir proposer un environment d'expérimentation qui soit suffisamment flexible pour permettre toutes les expériences utiles à la communauté. C'est néanmoins par ce biais que j'ai approché la question, en suivant le paradigme du *faire pour apprendre*. Durant plusieurs années, j'ai donc développé une plateforme expérimentale basée sur un formalisme que je présenterai ici. Elle n'est probablement pas la plus élégante qui puisse être envisagée, mais elle a l'avantage d'être rompue à la pratique, où les besoins en expressivité sont dominants.

En sciences des données comme dans d'autres domaines, le principe est d'étudier la sortie s d'un processus p en fonction de certaines données d'entrée d. On souhaite que s soit la meilleure possible. Cette performance est mesurée en fonction d'un ensemble de critères, que l'on appellera ici des *observations* (o).



Un processus de traitement de donnees.

À des fins de reproductibilité, on supposera que d est pérenne, *i.e.* n'est pas modifié par p. s et o sont non volatils, *i.e.* stockés sur disque. p peut éventuellement être scindé en plusieurs étapes de traitement, qui appliquées de manière séquentielle, permettent de produire le même résultat ($s_2 = s$ et $o_2 = o$).



Séparation de ce processus en deux étapes.

Ce découpage peut se faire de manière analytique en considérant l'architecture algorithmique de p. En pratique, ce découpage se fait de manière à optimiser le temps nécessaire à la production de s et o. En effet, le but de la découpe en sous tâches est de réduire le temps de calcul associé à la production de s et o. Choisir de scinder un processus en deux étapes de traitement est donc fonction du compromis entre le temps de calcul nécessaire à la production de s_1 et o_1 et leurs accès. En effet, si le temps nécessaire pour stocker et rapatrier s_1 et o_1 est supérieur au temps de calcul nécessaire à leur production, ce découpage n'est pas pertinent.

Le but de l'expérimentation est la confrontation de plusieurs alternatives concernant l'architecture de p. Nous nommerons *facteur*, un élément de conception de p. Chacun de ces facteurs peut prendre plusieurs valeurs, que nous nommerons *modalités*. Nous nommerons également *condition expérimentale*, un ensemble des modalités nécessaires à la paramétrisation complète de p. Pour déterminer la condition menant à la meilleure solution, l'exploration d'un plan expérimental est pour cela nécessaire. Néanmoins, dès que l'étude dépasse la trivialité, le coût d'exécution du plan complet, *i.e.* comprenant toutes les conditions expérimentales, devient prohibitif.

Une première technique consiste à effectuer l'intégralité du plan expérimental mais en simplifiant le problème à résoudre, souvent en réduisant la taille de d (nombres d'exemples ou de dimensions). Cette approche est risquée car elle peut amener à des choix de conception algorithmique qui ne généralisent pas correctement si le

jeu de données réduit n'est pas représentatif.

Une seconde approche, plus sûre, consiste à judicieusement choisir les conditions qui nous permettent de maximiser notre connaissance de l'influence des différents facteurs, et en particulier leurs interactions. En effet, supposer que tous les facteurs sont indépendants permet de linéariser l'exploration des conditions en considérant un plan expérimental en étoile. Malheureusement, cette indépendance est rarement complètement vérifiée de manière expérimentale. Il convient alors de pouvoir juger de ce degré d'interaction pour faire des choix éclairés. On peut pour cela utiliser des plans à facteurs réduits suivi d'une analyse de corrélations pour identifier l'importance de chaque facteur ainsi que les corrélations éventuelles entre facteurs.[71]

71. Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017

Considérons un plan expérimental avec 4 facteurs f_1, f_2, f_3, et f_4 ayant chacun 10 modalités $f_i \in \{1, 2, ..., 10\}$ et un processus ayant une sortie $s = f_1 * f_2 + f_3 + \mathcal{N}(0, 1)$, où $\mathcal{N}(0, 1)$ désigne un échantillon d'une loi normale de moyenne nulle et de variance unitaire. On considère pour simplifier ici que la sortie constitue notre observation $o = s$.

Déterminer « à l'aveugle » la meilleure condition expérimentale, *i.e.* en considérant le processus computationel étudié comme une boîte noire, se fait par l'exploration du plan complet qui va nécessiter l'évaluation de $10^4 = 10000$ conditions. On cherche donc à réduire ce nombre d'évaluations nécessaire. Pour cela, on va réduire le nombre de modalités pour chaque facteur aux deux extrema. L'exploration de ce plan expérimental va donc nécessiter l'évaluation de $2^4 = 16$ conditions expérimentales. Une analyse factorielle multivariée va permettre d'identifier l'importance des différents facteurs ainsi que leurs interactions. Comme on peut le voir sur la Table 1, l'analyse indique une interaction entre les facteurs f_1 et f_2 et une influence nulle de f_4 sur notre observation. Cette connaissance nous permet de déterminer la meilleure condition avec $2^4 + 10^2 + 10 = 126$ évaluations, ce qui amène un gain d'un facteur 100 en coût de calcul environ par rapport au plan complet. On voit ici l'intérêt de pouvoir facilement manipuler tous les éléments du plan expérimental. Il faut bien entendu nuancer une telle performance, car dans des cas où les courbes de réponses des observations sont très peu régulières, les outils statistiques utilisés ici peuvent montrer leurs limites.

	f_1	f_2	f_3	f_4
f_1	+	+		
f_2	+	+		
f_3			+	
f_4				

TABLE 1: Analyse de variance du plan d'expérience réduit. Le signe + indique une valeur-p inférieure à .05.

La poursuite d'une expérimentation peut se décomposer en trois grandes étapes :

1. **la conception** : conception et implantation de p, mise en place du plan expérimental, paramétrisation de l'environnement d'exécution, exécution de tests de conformité ;
2. **l'exécution** : exécution des conditions expérimentales choisies ;
3. **la réduction** : conversion des observations générées par chaque condition expérimentale en données exploitables.

Chacune de ces étapes a ses besoins propres en terme de res-

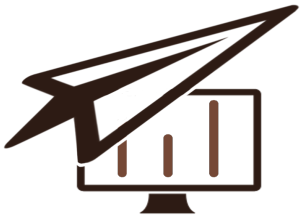
sources. L'étape de conception nécessite une capacité élevée d'interaction avec l'implantation de p. Il est souhaitable que p puisse être exécuté localement, éventuellement avec une version réduite de d pour assurer au maximum la qualité du développement de p avant l'étape d'exécution et un cycle d'essai / modification qui soit le plus réduit possible.

L'étape d'exécution a pour but de réaliser la plus grande partie possible du plan expérimental demandé. L'interaction avec p étant alors minimale, on peut déporter le calcul sur des machines distantes. Des possibilités d'analyse *a posteriori* du déroulement de l'exécution de cette étape sont cruciales. En effet, même avec une implantation de p vérifiée et validée, il est rare que cette implantation ait le comportement souhaité pour toutes les conditions expérimentales. Il est donc crucial que cela n'impacte pas le bon déroulement des autres conditions et de pouvoir accéder facilement aux conditions expérimentales qui mettent en défaut l'implantation courante de p.

L'étape d'exécution va produire un ensemble généralement grand d'observations qu'il sera nécessaire de réduire pour produire un ensemble de données facilement manipulable, dans le but de nous informer sur le comportement de p lorsqu'il est soumis à d. Même si le nombre d'observations est trop grand pour être appréhendé directement pour un être humain, on supposera que le volume de données est suffisamment faible pour pouvoir être manipulé localement.

ExpLanes

Le projet expLanes est né d'une frustration grandissante de l'auteur au sujet 1) du coût de développement nécessaire à la mise en place d'un protocole expérimental de qualité, 2) l'absence de canons dans ce domaine, et 3) du risque élevé d'erreurs durant l'implantation de ce protocole expérimental compromettant sérieusement la validité des résultats obtenus. La version actuelle, écrite pour l'environnement *MATLAB*®, a nécessitée trois ans de développement et est maintenant stabilisée et utilisée pour un bon nombre de mes contributions.



http://mathieulagrange.github.io/explanes

Architecture

Le principe d'expLanes est de fournir un environnement complet permettant de mettre en place un protocole expérimental destiné à répondre à des questionnements relevant des sciences des données. En particulier, il allège et systématise certaines étapes incontournables de ce type de protocole de manière à accélérer le processus de production, réduire les risques d'erreurs, et faciliter la maintenance à long terme.

Son architecture logicielle se compose de quatre composants principaux :

1. un module de **commande**,
2. un module de **planification**,
3. un module d'**exécution**,
4. un module de **réduction**.

Le module de **commande** permet de configurer l'environnement logiciel et matériel dans lequel sera exécuté la commande transmise par l'utilisateur. Toutes les commandes disponibles sont présentes à l'état statique dans un fichier de configuration qui est spécifique à chacun des utilisateurs. On peut, par l'intermédiaire de l'interpréteur modifier les valeurs de ces commandes ou en ajouter d'autres. Au sein du code de chaque étape de l'implantation du processus, cette configuration est disponible via l'objet `config`.

Le module de **planification** permet de gérer le plan d'expérience. A partir d'un fichier de description de ce plan, il génère efficacement une version manipulable. L'utilisateur peut alors aisément sélectionner les conditions expérimentales qu'il souhaite exécuter ou visualiser grâce à un mécanisme de masquage. Supposons un plan expérimental composé de 4 facteurs de 4 modalités chacune, le masque $\{2, 0, 1\}$ spécifie que la commande associée doit s'appliquer aux conditions expérimentales avec la seconde modalité du premier facteur, toutes les modalités du second facteur et la première modalité du troisième facteur. Le plan expérimental correspondant comprend alors 16 conditions au lieu de 256 pour le plan complet dont le masque implicite est $\{0, 0, 0\}$.

Le module d'**exécution** se charge, en fonction de la commande d'exécution donnée par l'utilisateur, des tâches suivantes :

— séquencement des conditions expérimentales ;
— chargement des données nécessaires et stockage des données produites ;
— enregistrement de la progression des calculs ;
— enregistrement des conditions expérimentales générant des erreurs.

Une fois les calculs terminés, le module de **réduction** permet, pour les conditions expérimentales sélectionnées par l'utilisateur, de :

— récupérer les observations correspondantes ;
— visualiser, par de nombreux moyens graphiques, ces observations ;
— produire, à partir de ces visualisations, un rapport d'étude.

Usage

Prenons un exemple d'étude qui nous permettra d'apprécier les fonctionnalités décrites précédemment du point de vue de l'expérimentateur. Nous cherchons de manière empirique à obtenir des connaissances sur l'aire de base et le volume de certaines formes géométriques tridimensionelles.

Le code de cet exemple est disponible`https://github.com/mathieulagrange/expLanes/tree/master/demo/geometricShape`

La création de l'expérimentation *geometricShape* se fait comme suit :

```
expCreate('geometricShape');
```

Nous pouvons ensuite instantier deux étapes de traitement, une qui se chargera du calcul de l'aire de la base :

```
geometricShape('addStep', 'base');
```

et une autre du volume :

```
geometricShape('addStep', 'space');
```

Nous nous intéressons dans cette expérimentation à l'influence potentielle des différents attributs (forme, couleur, rayon, largeur, hauteur) sur l'aire de sa base et son volume. Nous considérons donc chacun de ces attributs comme des facteurs expérimentaux avec des modalités particulières.

Les formes étudiées sont le cylindre, la pyramide et le cube, nous ajoutons donc un facteur nommé `'shape'` de trois modalités :

```
geometricShape('addFactor', ...
    {'shape', {'cylinder', 'pyramid', 'cube'}});
```

La couleur peut être bleu ou rouge, le rayon peut être de 2, 4, ou 6 mètres, la largeur de la pyramide et du cube peut être de 1, 2, et 3 mètres et la hauteur de 2, 4 ,et 6 mètres. On note ici que certains facteurs ne devront être explorés que pour une modalité d'un autre facteur. Il est en effet très utile en pratique de pouvoir exprimer ce type d'exclusion, qui est courante dans tout plan expérimental de complexité raisonnable mettant en compétition plusieurs approches algorithmiques de paramétrisation différentes.

Le fichier statique, nommé `geshFactors.txt` pour cette expérimentation, permet de stocker les informations nécessaires à la création du plan expérimental :

```
Factors:
1    shape === {'cylinder', 'pyramid', 'cube'}
2    color === {'blue', 'red'}
3    radius ==1/1= [2, 4, 6]
4    width ==1/[2 3]= 1:3
5    height =2=1/[1 2]= 2:2:6
```

La ligne 5 peut se lire comme suit, le facteur `'height'` n'est considéré que pour l'étape de traitement 2 et les modalités 1 et 2 du facteur 1.

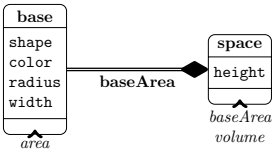
Les deux étapes de traitement sont implantées comme suit. Chaque fonction responsable de chaque étape dispose de la même signature. La variable *config* expose la configuration de l'expérimentation, la variable *setting* la condition expérimentale courante, et la variable *data* expose les données produite par l'étape précédente ou les données d'entrée pour la première étape.

La première étape se charge du calcul de l'aire de la base de la forme géométrique. Pour les besoins de l'exemple, on suppose ici que π n'est pas connu de manière précise, mais que nous disposons de 100 observations bruitées de cette valeur, ce qui est simulé ici par l'ajout d'une certaine incertitude sur sa mesure :

```
function [config, store, obs] = gesh1base(config, setting, data)

uncertainty = randn(1, 100);
switch setting.shape
    case 'cylinder'
        baseArea = (pi+uncertainty)*setting.radius^2;
    otherwise
        baseArea  = setting.width^2;
end
store.baseArea = baseArea;
obs.area = baseArea;
```

La seconde étape complète le calcul effectué par la première pour calculer le volume de la forme géométrique :

```
function [config, store, obs] = gesh2space(config, setting, data)

switch setting.shape
    case 'cube'
       volume = data.baseArea*setting.width;
    case 'cylinder'
        volume = data.baseArea*setting.height;
    case 'pyramid'
        volume = data.baseArea*setting.height/3;
end

obs.baseArea = data.baseArea;
obs.volume = volume;
```



La commande `geometricShape('f')` génère un diagramme où les éléments principaux de l'expérimentation sont visibles (étapes de traitement, facteurs, données sauvegardées, observations).

Une fois l'implantation effectuée, le traitement est requis avec la commande `'do'`. Par exemple, la commande

```
geometricShape('do', 1);
```

demande le calcul de l'étape 1 pour toutes les conditions expérimentales, ici au nombre de 18. Il est souvent nécessaire de ne pas effectuer le plan expérimental complet (étude spécifique, reprise sur erreur, ...). Le mécanisme de masquage est alors utilisé. Par exemple, la commande

```
geometricShape('do', 0, 'mask', {[1 2] 0 1})}
```

demande le calcul successif des deux étapes pour les cylindres de rayon 2 et toutes les pyramides.

A la suite de la réalisation du plan complet, le répertoire dédié au stockage des données de calcul et d'observations contient :

```
base/color: blue, radius: 2, shape: cylinder_data.mat
base/color: blue, radius: 2, shape: cylinder_obs.mat
base/color: blue, radius: 4, shape: cylinder_data.mat
base/color: blue, radius: 4, shape: cylinder_obs.mat
base/color: blue, radius: 6, shape: cylinder_data.mat
base/color: blue, radius: 6, shape: cylinder_obs.mat
base/color: blue, shape: cube, width: 1_data.mat
base/color: blue, shape: cube, width: 1_obs.mat
...
```

```
space/color: blue, height: 2, shape: pyramid, width: 2_obs.mat
space/color: blue, height: 2, shape: pyramid, width: 3_obs.mat
space/color: blue, height: 4, radius: 2, shape: cylinder_obs.mat
space/color: blue, height: 4, radius: 4, shape: cylinder_obs.mat
space/color: blue, height: 4, radius: 6, shape: cylinder_obs.mat
space/color: blue, height: 4, shape: pyramid, width: 1_obs.mat
space/color: blue, height: 4, shape: pyramid, width: 2_obs.mat
space/color: blue, height: 4, shape: pyramid, width: 3_obs.mat
...
```

Ce n'est pas explicite dans cet exemple, mais les fichiers `_data` dédiés au stockage des données de calculs sont souvent de taille plus conséquente que les observations qui elles ont vocation à rester de taille raisonnable pour pouvoir être transférées des machines de calcul aux machines de développement et de visualisation pour pouvoir être aisément analysées. Pour toute expérimentation de taille raisonnable, les noms de fichiers peuvent également être

compactés pour ne pas dépasser les 250 caractères requis par la plupart des systèmes de gestions de fichiers.

À la fin du traitement, les observations de la dernière étape de traitement sont affichées :

```
Loaded data files dates are in the range: | 01-Apr-2019 10:15:55 || 01-Apr-2019 10:15:56 |
    'shape'       'color'    'radius'    'width'    'height'    'baseArea'          'time'    'volume'
    '---'         '---'      '---'       '---'      '---'       '---'               '---'     '---'
    'cylinder'    'blue'     '2'         ''         '2'         '  12.55 (4.32)'    '0.22'    '   25.11 (8.63)'
    'cylinder'    'blue'     '2'         ''         '4'         '  12.55 (4.32)'    '0.09'    '  50.22 (17.27)'
    'cylinder'    'blue'     '2'         ''         '6'         '  12.55 (4.32)'    '0.00'    '  75.32 (25.90)'
    'cylinder'    'blue'     '4'         ''         '2'         ' 50.85 (16.54)'    '0.00'    ' 101.70 (33.08)'
    'cylinder'    'blue'     '4'         ''         '4'         ' 50.85 (16.54)'    '0.00'    ' 203.41 (66.15)'
    'cylinder'    'blue'     '4'         ''         '6'         ' 50.85 (16.54)'    '0.00'    ' 305.11 (99.23)'
    'cylinder'    'blue'     '6'         ''         '2'         '109.79 (33.75)'    '0.00'    ' 219.57 (67.50)'
    'cylinder'    'blue'     '6'         ''         '4'         '109.79 (33.75)'    '0.01'    '439.14 (135.00)'
    'cylinder'    'blue'     '6'         ''         '6'         '109.79 (33.75)'    '0.00'    '658.72 (202.50)'
    'cylinder'    'red'      '2'         ''         '2'         '  11.92 (4.33)'    '0.00'    '   23.84 (8.66)'
    'cylinder'    'red'      '2'         ''         '4'         '  11.92 (4.33)'    '0.00'    '  47.68 (17.32)'
```

L'accès à une grande diversité de mises en forme de l'exposition des observations utiles au discours est un élément nécessaire pour une gestion efficace de l'expérimentation. La plateforme expLanes dispose d'un grand nombre de fonctionnalités conçues pour permettre une certaine flexibilité dans leurs usages.

Par exemple, ce dernier affichage peut être ré-obtenu grâce à la commande suivante :

```
geometricShape('display', 2, 'expose', '>', 'mask', {[1 2] 0 1});
```

La commande :

```
geometricShape('display', 2, 'mask', {1 0 1},...
 'expose', {'t', 'obs', 2});
```

affiche les volumes (observations 2) de l'étape 2 pour chaque cylindre de rayon 2 triés par la première observation sélectionnée. La fonte rouge indique la plus haute valeur, et les valeurs en gras, celles qui ne sont pas statistiquement différentes, au sens d'un t-test par paires effectué entre les observations de la condition expérimentale et celle correspondant à la plus haute valeur.

color	height	volume
blue	2	26.12±9.30
blue	4	52.23±18.60
blue	6	**78.35±27.90**
red	2	24.61±7.54
red	4	49.23±15.07
red	6	**73.84±22.61**

Visualisation sous forme de table LaTeX.

Dans notre exemple, même si le cylindre bleu est plus volumineux que le rouge, cette différence n'étant pas significative, on peut en conclure expérimentalement que la couleur n'influe pas sur le volume [72]. Une fois la phase d'analyse des résultats effectuée, nous pouvons alors communiquer les résultats grâce à la production de rapports d'expérimentation. Quelques exemples de rapports d'expérimentation pour des expérimentations plus ambitieuses sont disponibles sur le site du projet *expLanes*.

72. Ce qui est rassurant.

Je me suis attaché à décrire ici une contribution personnelle à une problématique d'importance qui, sans prétendre viser à l'universalité, propose une systématisation de l'étape d'expérimentation de la méthode scientifique, adaptée aux problématiques relevant des sciences des données.

La revue par les pairs

Dernière étape de la méthode scientifique, la revue par les pairs garantit la pérennité du cycle d'acquisition de connaissances par

cette méthode. La revue par les pairs est à ce titre un outil indispensable à la communauté scientifique. Elle comporte des risques pour le chercheur, car c'est en effet à ce moment que les fruits de son travail sont exposés, à la critique, à l'utilisation abusive. Dans le même temps ; c'est aussi une formidable opportunité pour lui de voir son travail servir de base pour la recherche future.

Dans une tradition issue de la physique expérimentale, la revue par les pairs se fonde sur la production d'une description aussi précise que possible du protocole expérimental et d'une discussion plus ou moins approfondie sur l'interprétation que font les auteurs sur les résultats obtenus. Ce format « article » est motivé par le fait que l'environnement expérimental est très coûteux à mettre en place et difficilement transposable. Par accumulation d'expériences avec différents protocoles expérimentaux, la communauté s'approche de ce que l'on peut être raisonnable de penser être la vérité. Un exemple particulièrement illustratif de ce processus est montré sur la Figure 18 montrant une série de mesures de la conductivité du cuivre en fonction de sa température. Par l'accumulation et la réduction de mesures effectuées de manière indépendante, notre connaissance du phénomène observé s'affine.

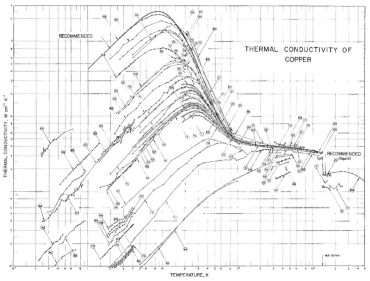


Figure 18: Mesures de la conductivité du cuivre en fonction de sa température. Chaque ligne pointée par une bulle numérotée désigne les mesures publiées dans un article donné.

Il faut néanmoins noter que l'on assiste actuellement à une « crise » de la publication d'articles, principalement à cause d'une normalisation d'une expression anglo saxonne "publish or perish" qui est, dans un nombre toujours plus grand de pays, devenue loi. Les nouveaux moyens numériques de diffusion facilitant cette inflation, le chercheur doit passer beaucoup de temps à écrire, mais aussi à relire, des articles dont le contenu s'allège en conséquence. Selon Bornmann,[73] le nombre de publications par an est passé de 650 000 en 1980 à presque 2 000 000 en 2012.

En considérant que le nombre de personnes dédiant leur carrière à la recherche scientifique est resté stable durant ces 50 dernières années, ces estimations de l'évolution du nombre d'articles publiés sont préoccupantes car elle posent de nombreuses inquiétudes quand à la solidité du principe de publication d'articles comme fondement de la revue par les pairs.

A ceci s'ajoute la position actuellement indéfendable des éditeurs qui, par abus de position historique, ne facilitent pas l'évolution nécessaire. Comment expliquer au contribuable les sommes demandées au chercheur pour l'archivage de ses articles (actuellement environ 1500 euros pour un fichier d'une dizaine de mégaoctets) ? Il est malheureusement difficile pour un chercheur de prendre la décision de manière unilatérale d'utiliser des dispositions alternatives.

Comme tout moment de crise, de nouvelles opportunités émergent également qui pourraient amener à une nouvelle implémentation de cette dernière étape de la méthode scientifique qui soit plus respectueuse du chercheur et également plus effective pour la communauté.

Dans le domaine de la sciences des données, on peut considérablement accélérer le processus de reproduction et d'extension,

73. Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66 (11):2215–2222, 2015

car il est plus aisé de mettre en place le paradigme de la recherche reproductible. En suivant la définition séminale de Donoho et en l'étendant au cadre des sciences des données, on peut considérer qu'une contribution scientifique doit pour cela être composée :

1. des données servant de base à l'expérimentation ;
2. de l'implémentation qui, en fonction de ces données, produit les éléments quantitatifs discutés dans le compte rendu ;
3. d'un compte rendu d'expérience ou article détaillant les hypothèses, le protocole expérimental et les conclusions.

"An article about computational science in a scientific publication is not the scholarship itself, it is merely an advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."
—D. Donoho

Concernant la pérennité des données et de l'article, ces données numériques dites « statiques » ne posent pas de challenge particulier. Ce n'est malheureusement pas le cas du code informatique car il dépend d'une pérennité de l'interprétation de ce code sur les machines contemporaines. L'archivage et la maintenance du matériel qui a servi aux expérimentations n'est pas dans le cas général envisageable. Les difficultés de réplication pour cause d'obsolescence matérielle et logicielle sont endémiques de la recherche en sciences des données car cette activité se situe généralement dans des secteurs de pointe évoluant très rapidement.

Je citerai pour exemple l'évolution actuelle des bibliothèques de calcul sur unité de calcul graphique comme la bibliothèque *cuda*® développée par *nvidia*®. Indispensable pour tout traitement efficace d'architectures profondes, cette bibliothèque est propriétaire et les versions antérieures ne sont plus maintenues après seulement quelques années.

Il reste que, bien souvent, une maintenance régulière du code par les auteurs permet de préserver un certain niveau d'équivalence des résultats. Il est alors pour cela très utile de disposer d'un code qui soit bien structuré et facile à maintenir. Cette maintenance sera également grandement facilitée par le choix judicieux des dépendances du code informatique publié. Si la phase d'expérimentation nous incite généralement à expérimenter avec de nombreuses approches, dont certaines très novatrices mais potentiellement instables et faiblement maintenues, il convient au moment de la publication de réfléchir à un bon équilibre entre apport technique et pérennité.

Ceci étant dit, garantir une reproducibilité complète pour un temps long de toutes ses productions est un défi probablement trop coûteux pour le chercheur. Il convient néanmoins de réfléchir en amont à cette problématique et d'allouer l'énergie nécessaire aux projets qui trouvent intérêt dans la communauté.

Les plateformes de développement collaboratif comme GitHub sont à ce titre particulièrement adaptées car elles permettent d'être informé de l'intérêt que porte la communauté aux travaux publiés et d'assister les collègues lors des tentatives de réplication.

Même si l'implémentation particulière n'a pas vocation à être pérenne, il est bien évident que si les auteurs ont suivi un protocole expérimental canonique avec un effort de clarté et de concision dans l'implémentation, la réplication sera bien plus aisée. L'utilisation de cadres computationnels comme `expLanes` pour mettre en place ce protocole expérimental est alors d'intérêt car ils permettent d'encourager de bonnes pratiques en facilitant leurs mises en place.

Concernant les données nécessaires à la production des résultats, il est techniquement parlant aisé que les données expérimentales soient stockées de manière pérenne sur des plateformes publiques[74]. La contrainte ici n'est pas d'ordre technique mais légale. L'usage de données publiques doit donc être encouragé par la communauté, quitte à s'éloigner du champ applicatif immédiat, où les enjeux économiques, légaux, ou encore de respect de la vie privée

74. Les plateformes de type *archive* (`https://archive.org`) ou *zenodo* (`https://zenodo.org`) sont de bons candidats.

entrent en conflit avec les principes mêmes du questionnement scientifique.

Parcours académique

Statut actuel : Chargé de recherche CNRS classe normale
- recruté en 2009 par la commission interdisciplinaire (CID) 44
« Cognition, langage, traitement de l'information, systèmes naturels et artificiels »
- rattaché en 2012 à la section 07 « Sciences de l'information »

CNU : - section 27 « Informatique »
- section 61 « Génie informatique, automatique et traitement du signal »

Diplômes

2001 Master Informatique : « Accélération de la synthèse sonore »
encadré par Sylvain Marchand,
soutenu à l'Université de Bordeaux 1

2004 Doctorat Informatique : « Modélisation long terme des signaux polyphoniques »
dirigé par Myriam Desainte-Catherine, Sylvain Marchand et Jean-Bernard Rault,
soutenu à l'Université de Bordeaux 1
https://www.theses.fr/2004BOR12917
https://tel.archives-ouvertes.fr/tel-00009550

Carrière

2001-2004 **Doctorant Université Bordeaux 1** et Ingénieur de Recherche
à France Télécom R&D Rennes
TECH/IRIS (équipe codage et multimédia)

2004-2005 **Enseignant chercheur (ATER)** au LaBRI (Université Bordeaux 1)

2005-2006 **Enseignant chercheur (ATER)** à l'Enseirb (Université Bordeaux 1)

2006-2007 **Post-doctorant** au sein du département d'informatique
Université de Victoria, BC, Canada

2007-2008 **Post-doctorant** au sein du département de "Music Technology"
Université de McGill, QC, Canada

2008- 2009 **Post-doctorant** au sein de l'équipe Acoustique Audio et Ondes (Aao)
Télécom ParisTech

2009-2013 **Chercheur CNRS** au sein de l'équipe Analyse / Synthèse
Ircam (Umr 9912), Paris

2013- – **Chercheur CNRS** au sein de l'équipe
Signal, Images et Son (Sims)
Ls2n (Umr 6004), Ecole Centrale de Nantes

Participation à des projets financés

— 2003 - 2004 : "Adaptive rate-distortion optimized sound coder", projet Européen (Eu grant no. IST-2001-34095)

— 2006 - 2007 : "From the laboratory to the concert : applications of gesture research to live performance", projet Sshrc, Canada

— 2006 - 2007 : "Graph algorithms for audio analysis", projet Nserc, Canada

— 2007 - 2008 : "Sound and haptic synthesis", projet Européen Enactive et financement Nserc, Canada

— 2008 - 2009 : "Décomposition en éléments sonores et application à la musique", projet Anr

— 2009 - 2012 : "Quaero", Projet Oseo

— 2012 - 2014 : "Computational auditory scene analysis", projet fondation sciences et technologie, Portugal

— 2012 - 2015 : "Hierarchical object based learning", projet Anr jeune chercheur / jeune chercheuse (investigateur principal)

— 2015 - 2017 : Projet "Traité instrumental collaboratif en ligne" (Ticel), Projet Paris sciences et lettres (Psl) en collaboration avec le Conservatoire National de Musique et de Danse de Paris (Cnsmdp)

— 2017 - – : "Caractérisation des environnements sonores urbains : vers une approche globale associant données libres, mesures et modélisations" (Cense), projet Anr

Encadrement de doctorats

— Rémi Foucard (2010 - 2013) : "Fusion multi-niveaux par boosting pour le tagging automatique", taux d'encadrement 40 %

— Grégoire Lafay (2013 - 2016) : "Simulation de scènes sonores environnementales : application à l'analyse sensorielle et à l'analyse automatique", taux d'encadrement 40 %

— Jean-Rémy Gloaguen (2015 - 2018) : "Estimation du niveau sonore de sources d'intérêt au sein de mélanges sonores urbains : application au trafic routier", taux d'encadrement 30 %

— Félix Gontier (2017 - –) : "Modélisation de signaux sonores par approches neuronales profondes", taux d'encadrement 30 %

— Tom Souaille (2019 - –) : "Conception interactive en design sonore", taux d'encadrement 30 %

Indices bibliométriques

— 21 revues internationales à comité de lecture

— 62 conférences internationales à comité de lecture

— citations : 1615 (source Google Scholar, Mai 2019)

— indice h : 19 (source Google Scholar, Mai 2019)

Notice bibliographique

Les publications référencées ici sont disponible au format pdf à cette adresse :
https://mathieulagrange.github.io

Revues à comité de lecture

Jean-Rémy Gloaguen, Arnaud Can, Mathieu Lagrange, and Jean-François Petiot. Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization. *Applied Acoustics*, 143 :229–238, 2019.

Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange. Fourier at the heart of computer music : From harmonic sounds to texture. *Comptes Rendus de Physique de l'Académie des Sciences*, in press, 2019.

Grégoire Lafay, Mathias Rossignol, Nicolas Misdariis, Mathieu Lagrange, and Jean-François Petiot. Investigating soundscapes perception through acoustic scenes simulation. *Behavior Research Methods*, 2(51), 2018.

Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D. Plumbley. Detection and Classification of Acoustic Scenes and Events : Outcome of the DCASE 2016 Challenge. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(2) :379–393, 2018.

Mathias Rossignol, Mathieu Lagrange, and Arshia Cont. Efficient similarity-based data clustering by optimal object to cluster reallocation. *PloS one*, 13(6) :e0197450, 2018.

Emmanouil Benetos, Grégoire Lafay, Mathieu Lagrange, and Mark D Plumbley. Polyphonic Sound Event Tracking using Linear Dynamical Systems. *IEEE Transactions on Audio, Speech and Language Processing*, 25(6) :1266–1277, 2017.

Félix Gontier, Mathieu Lagrange, Pierre Aumond, Arnaud Can, and Catherine Lavandier. An Efficient Audio Coding Scheme for Quantitative and Qualitative Large Scale Acoustic Monitoring Using the Sensor Grid Approach. *Sensors*, 17(12) :2758, 2017.

Grégoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, and Axel Roebel. A morphological model for simulating acoustic scenes and its application to sound event detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(10) :1854–1864, 2016.

Grégoire Lafay, Nicolas Misdariis, Mathieu Lagrange, and Mathias Rossignol. Semantic browsing of sound databases without keywords. *Journal of the Audio Engineering Society*, 64(9) :628–635, 2016.

Mathieu Lagrange, Hervé Andrieu, Isabelle Emmanuel, and Gerard Busquets. Classification of rainfall radar images using the scattering transform. *Journal of Hydrology*, 2016.

Mathieu Lagrange, Grégoire Lafay, Boris Défréville, and Jean-Julien Aucouturier. The bag-of-frames approach : A not so sufficient model for urban soundscapes. *The Journal of the Acoustical Society of America*, 138(5) :487–492, 2015.

Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, 17 (10), 2015.

Alexey Ozerov, Mathieu Lagrange, and Emmanuel Vincent. Uncertainty-based learning of acoustic models from noisy data. *Computer Speech & Language*, 27(3) :874–894, 2013.

Mathieu Lagrange, Roland Badeau, Bertrand David, Nancy Bertin, Olivier Derrien, Sylvain Marchand, Laurent Daudet, et al. Décompositions en éléments sonores et applications musicales. *Traitement du Signal*, 28(6) :665–689, 2011.

Emma Murphy, Mathieu Lagrange, Gary Scavone, Philippe Depalle, and Catherine Guastavino. Perceptual evaluation of rolling sound synthesis. *Acta acustica united with Acustica*, 5-97 :840–851, 2011.

Mathieu Lagrange, Gary Scavone, and Philippe Depalle. Analysis / synthesis of sounds generated by sustained contact between rigid objects. *IEEE Transactions on Audio Speech and Language Processing*, 18-3 :509–518, 2010.

M. Lagrange and M. Raspaud. Spectral similarity metrics for sound source formation based on the common variation cue. *ACM Multimedia Tools and Applications Journal on Content-Based Multimedia Indexing*, 2009.

M. Lagrange, M. Raspaud, R. Badeau, and G. Richard. Explicit modeling of temporal dynamics within musical signals for acoustical unit similarity. *Pattern Recognition Letters*, 2009.

M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized Cuts for Predominant Melodic Source Separation. *IEEE Transactions on Acoustics, Speech and Language Processing*, 2008.

M. Lagrange and S. Marchand. Estimating the instantaneous frequency of sinusoidal components using phase-based methods. *Journal of the Audio Engineering Society*, 2007.

M. Lagrange, S. Marchand, and J.B. Rault. Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds. *IEEE Transactions on Acoustics, Speech and Language Processing*, 2007.

M. Lagrange, S. Marchand, and J. Rault. Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling. *Journal of the Audio Engineering Society*, 53(10) :891–905, 2005.

Actes de colloques à comité de lecture

Félix Gontier, Pierre Aumond, Mathieu Lagrange, Catherine Lavandier, and Jean-François Petiot. Towards perceptual soundscape characterization using event detection algorithms. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Surrey, United Kingdom, 2018.

Mathieu Lagrange, Mathias Rossignol, and Grégoire Lafay. Visualization of audio data using stacked graphs. In *19th International Society for Music Information Retrieval Conference*, Paris, France, 2018.

Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange. Extended playing techniques : the next milestone in musical instrument recognition. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, pages 1–10, 2018.

Jean-Rémy Gloaguen, Arnaud Can, Mathieu Lagrange, and Jean François Petiot. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. In *173rd Meeting of the Acoustical Society of America and the 8th Forum Acusticum (Acoustics)*, volume 25, pages 1266–1277, Boston, MA, United States, 2017.

Grégoire Lafay, Emmanouil Benetos, and Mathieu Lagrange. Sound Event Detection in Synthetic Audio : Analysis of the DCASE 2016 Task Results. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, United States, 2017.

Judicaël Picaut, Arnaud Can, Jérémy Ardouin, Pierre Crepeaux, Thierry Dhome, David Ecotiere, Mathieu Lagrange, Catherine Lavandier, Vivien Mallet, Christophe Mietlicki, and Marc Paboeuf. Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling. In *Acoustics '17*, page 4 p, Boston, United States, 2017.

Emmanouil Benetos, Grégoire Lafay, Mathieu Lagrange, and Mark D. Plumbley. Detection of Overlapping Acoustic Events using a Temporally-Constrained Probabilistic Model. In *ICASSP*, Shanghai, China, 2016.

Jean-Rémy Gloaguen, Arnaud Can, Mathieu Lagrange, and Jean-François Petiot. Estimating traffic noise levels using acoustic monitoring : a preliminary study. In *DCASE 2016, Detection and Classification of Acoustic Scenes and Events*, page 4p, Budapest, Hungary, 2016.

Mathieu Lagrange and Mathias Rossignol. Computational experiments in Science : Horse wrangling in the digital age. In *Research workshop on "Horses" in Applied Machine Learning*, London, United Kingdom, 2016.

Gerard Busquets, Mathieu Lagrange, Isabelle Emmanuel, and Hervé Andrieu. Classification of rainfall radar images using the scattering transform. In *European Conference on Signal Processing (EUSIPCO)*, Nice, France, 2015.

Axel Roebel, Jordi Pons Puig, Marco Liuni, and Mathieu Lagrange. On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence. In *Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 414 – 418, Brisbane, Australia, 2015.

Mathias Rossignol, Mathieu Lagrange, Grégoire Lafay, and Emmanouil Benetos. Alternate Level Clustering for Drum Transcription. In *European Conference on Signal Processing (EUSIPCO)*, Nice, France, 2015.

Grégoire Lafay, Mathias Rossignol, Nicolas Misdariis, Mathieu Lagrange, and Jean François Petiot. A new experimental approach for urban soundscape characterization based on sound manipulation ; a pilot study. In *International Symposium on Musical Acoustics (ISMA)*, Le Mans, France, 2014.

Mathias Rossignol, Gregoire Lafay, Mathieu Lagrange, and Nicolas Misdariis. Simscene : a web-based acoustic scenes simulator. In *First Web Audio Conference (WAC)*, 2014.

Carlo Baugé, Mathieu Lagrange, Joakim Andén, and Stéphane Mallat. Representing environmental sounds using the separable scattering transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

Emmanouil Benetos, Mathieu Lagrange, and Simon Dixon. Characterisation of acoustic scenes using a temporally-constrained shift-invariant model. In *International Conference on Digital Audio Effects*, York, United Kingdom, 2013.

Benjamin Cauchi, Mathieu Lagrange, Nicolas Misdariis, and Arshia Cont. Saliency-based modeling of acoustic scenes using sparse

non-negative matrix factorization. In *International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, 2013.

Rémi Foucard, Slim Essid, Mathieu Lagrange, and Gaël Richard. Etiquetage automatique de l'audio : une approche de boosting régressif basée sur une fusion souple d'annotateurs. In *Coresa*, 2013.

Rémi Foucard, Slim Essid, Mathieu Lagrange, and Gaël Richard. Etiquetage automatique de l'audio : une approche de boosting régressif basée sur une fusion souple d'annotateurs. In *International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, 2013.

Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley. Detection and classification of acoustic scenes and events : an ieee aasp challenge. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley. A database and challenge for acoustic scene classification and event detection. In *European Conference on Signal Processing (EUSIPCO)*, 2013.

Benjamin Cauchi, Mathieu Lagrange, Nicolas Misdariis, and Arshia Cont. Sparse representations for modeling environmental acoustic scenes, application to train stations soundscapes. In *International Conference on Acoustics*, Nantes, France, 2012.

Rémi Foucard, Slim Essid, Mathieu Lagrange, and Gaël Richard. A regressive bossting approach to automatic audio tagging based on soft annotator fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

Mathieu Lagrange, Luis Gustavo Martins, and George Tzanetakis. Cluster aware normalization for enhancing audio similarity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

Mathieu Lagrange, Alexey Ozerov, and Emmanuel Vincent. Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2012.

Remi Foucard, Slim Essid, Mathieu Lagrange, and Gael Richard. Multi-scale temporql fusion by boosting for music classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

Mathieu Lagrange and George Tzanetakis. Adaptive n-normalization for enhancing music similarity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, République Tchèque, 2011.

Alexey Ozerov, Mathieu Lagrange, and Emmanuel Vincent. Gmm-based classification from noisy features. In *1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011.

Francois Rigaud, Mathieu Lagrange, Axel Roebel, and Geoffroy Peeters. Drum extraction from polyphonic music based on a spectro-temporal model of percussive sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–384, Prague, Czech Republic, 2011.

Rémi Foucard, Jean-Louis Durrieu, Mathieu Lagrange, and Gaël Richard. Multimodal similarity between musical streams for cover version detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

Mathieu Lagrange and Joan Serra. Unsupervised accuracy improvement for cover song detection using spectral connectivity network. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Utrecht, Netherlands, 2010.

Mathieu Lagrange, Roland Badeau, Bertrand David, Nancy Bertin, Jose Echeveste, Olivier Derrien, and Sylvain Marchand. The desam toolbox : Spectral analysis of musical audio. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, Gratz, Austria, 2010.

Mathieu Lagrange, Roland Badeau, and Gaël Richard. Robust similarity metrics between audio signals based on asymmetrical spectral envelope matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

M. Lagrange and M. Raspaud. Spectral similarity metrics for sound source formation based on the common variation cue. In *Seventh International Workshop on Content-Based Multimedia Indexing (CBMI)*, Chania, Greece, 2009.

M. Robine, M. Lagrange, and P. Hanna. Meter Class Profile for Music Similarity and Retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2009.

M. Lagrange and B. Scherrer. Two-Step Modal Identification for Increased Resolution Analysis of Percussive Sounds. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, 2008.

M. Lagrange, L. G. Martins, and G. Tzanetakis. A computationally efficient scheme for dominant harmonic source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, 2008.

M. Lagrange, G. Scavone, and P. Depalle. Time-domain analysis / synthesis of the excitation signal in a source / filter model of contact sounds. *Proceedings of the International Conference on Auditory Display (ICAD)*, 2008.

M. Lagrange, N. Whetsell, and P. Depalle. On the control of the phase of resonant filters with applications to percussive sound modelling. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, 2008.

Emma Murphy, Mathieu Lagrange, Philippe Depalle, Gary Scavone, and Catherine Guastavino. Perceptual evaluation of a real-time synthesis technique for rolling sounds. In *5th International Conference on Enactive Interfaces*, Pisa, Italy, 2008.

Luis F. Teixeira, Luis G. Martins, Mathieu Lagrange, and George Tzanetakis. MarsyasX : Multimedia Dataflow Processing with Implicit Patching. In *ACM International Conference on Multimedia*, 2008.

G. Tzanetakis, R. Jones, C. Castillo, L. G. Martins, L. F. Teixeira, and M. Lagrange. Interoperability and Marsyas Runtime. In *Proceedings of the International Computer Music Conference (ICMC)*, Belfast, Ireland, 2008.

R. Jones, M. Lagrange, and A. Schloss. A Hand Drumming Dataset for Physical Modelling. In *Proceedings of the International Computer Music Conference (ICMC)*, Copenhagen, Denmark, 2007.

A. Kapur, G. Percival, M. Lagrange, and G. Tzanetakis. Pedagogical Transcription for Multimodal Sitar Performance. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2007.

M. Lagrange and G. Tzanetakis. Sound source tracking and formation using normalized cuts. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, USA, 2007.

M. Lagrange, L. G. Martins, and G. Tzanetakis. Semi-Automatic Mono to Stereo Up-mixing using Sound Source Formation. In *Audio Engineering Society (AES)*, 2007.

M. Lagrange, G. Percival, and G. Tzanetakis. Adaptive Harmonization and Pitch Correction of Polyphonic Audio using Spectral Clustering. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, 2007.

L. G. Martins, J. J. Burred, M. Lagrange, and G. Tzanetakis. Polyphonic Instrument Recognition using Spectral Clustering. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2007.

M. Robine, G. Percival, and M. Lagrange. Analysis of Saxophone Performance for Computer-Assisted Tutoring. In *Proceedings of the International Computer Music Conference (ICMC)*, Copenhagen, Denmark, 2007.

G. Tzanetakis, M. Lagrange, P. Spong, and H. Symonds. ORCHIVE : Digitizing and Analyzing Orca Vocalizations. In *Proceedings of the RIAO - Large-Scale Semantic Access to Content Conference*, 2007.

M. Lagrange and S. Marchand. Assessing the Quality of the Extraction and Tracking of Sinusoidal Components : Towards an Evaluation Methodology. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, 2006.

M. Lagrange, L. G. Martins, L. F. Teixeira, and G. Tzanetakis. Speaker segmentation of interviews using integrated video and audio change detections. In *Fifth International Workshop on Content-Based Multimedia Indexing (CBMI)*, Bordeaux, France, 2006.

Sylvain Marchand and Mathieu Lagrange. On the equivalence of phase-based methods for the estimation of instantaneous frequency. In *European Conference on Signal Processing*, 2006.

M. Robine and M. Lagrange. Evaluation of the Technical Level of Saxophone Performers by Considering the Evolution of Spectral Parameters of the Sound. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2006.

M. Lagrange. A New Dissimilarity Metric For The Clustering Of Partials Using The Common Variation Cue. In *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, 2005.

M. Lagrange, S. Marchand, and J. B. Rault. Improving Sinusoidal Frequency Estimation using a Trigonometrical Approach. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, 2005.

M. Lagrange, Sylvain Marchand, and J. B. Rault. Improving the Tracking of Partials for the Sinusoidal Modeling of Polyphonic Sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 241–244, 2005.

M. Lagrange, Sylvain Marchand, and J. B. Rault. Partial Tracking Based on Future Trajectories Exploration. In *116th Convention of the Audio Engineering Society*, Berlin, 2004.

M. Lagrange, Sylvain Marchand, and J. B. Rault. Using Linear Prediction to Enhance the Tracking of Partials. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 642–645, 2004.

M. Lagrange, Sylvain Marchand, Martin Raspaud, and J. B. Rault. Enhanced Partial Tracking Using Linear Prediction. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 141–146, 2003.

M. Lagrange, Sylvain Marchand, and J. B. Rault. Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 59–64, 2002.

M. Lagrange and Sylvain Marchand. Real-Time Additive Synthesis of Sound by Taking Advantage of Psychoacoustics. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 5–9, 2001.

Chapitre d'ouvrage

Luis Martins, Mathieu Lagrange, and George Tzanetakis. Modeling grouping cues for auditory scene analysis using a spectral clustering formulation. In *Machine Audition : Principles, Algorithms and Systems*. IGI Global, 2011.

Logiciels

Mathieu Lagrange. Explanes : a framework facilitating computational experiments (gpl). http://mathieulagrange.github.io/expLanes, 2014.

Mathias Rossignol, Grégoire Lafay, and Mathieu Lagrange. simscene : synthesis of sound scenes for the creation of evaluation corpora in audio event detection (gpl). https://bitbucket.org/mlagrange/simscene, 2013.

Mathias Rossignol and Mathieu Lagrange. kaverages : efficient clustering algorithms for arbitrary similarity matrices (gpl). https://bitbucket.org/mlagrange/kaverages, 2012.

Brevets

M. Lagrange and J. B. Rault. Procédé de restauration de partiels d'un signal sonore. PCT/FR 04/00619, France Télécom R&D, 2004.

M. Lagrange and J. B. Rault. Procédé de décision pour la restauration de partiels d'un signal sonore, dispositif et programme d'ordinateur pour sa mise en œuvre. PCT/FR 04/01415, France Télécom R&D, 2004.