hal-00849736, version 1 - 17 Sep 2013

DÉCOUVERTE AUTOMATIQUE DE STRUCTURES MUSICALES EN TEMPS RÉEL PAR LA GÉOMÉTRIE DE L'INFORMATION

VINCENT LOSTANLEN

Rapport de fin de master 2 ATIAM

effectué au sein de l'équipe-projet MuTant,
sous la direction d'Arshia Cont et Arnaud Dessein,
du 1er mars au 31 juillet 2013.

Ircam, Paris, le 31 juillet 2013.

hal-00849736, version 1 - 17 Sep 2013

Vincent Lostanlen, *Découverte automatique de structures musicales en temps réel par la géométrie de l'information.* Rapport de fin de master 2 ATIAM, Paris, juillet 2013.

hal-00849736, version 1 - 17 Sep 2013

RÉSUMÉ

Ce rapport de fin de stage vise à explorer la découverte automatique de structures musicales dans un fichier audio. Notre contribution principale est de formuler ce problème dans le cadre de la géométrie de l'information computationnelle, une discipline émergente mêlant statistiques, géométrie différentielle, et fouille de données. Nous y rassemblons la segmentation en évènements musicaux et le calcul de leurs similarités en un seul schéma séquentiel de structuration. De plus, nous proposons une métrique originale entre segments temporels, qui combine plusieurs critères de ressemblance géométrique : divergence entre centroïdes, mais aussi rapports d'inclusion et d'intersection entre les boules informationnelles associées. Nous testons la découverte de structures sur deux extraits musicaux, différant par leur orchestration, leur genre et le descripteur sonore choisi. À l'issue des deux expériences, notre métrique se révèle plus performante que les autres mesures de dissimilarité, exclusivement fondées sur une comparaison de centroïdes. En raison de l'omniprésence des familles exponentielles dans les modèles d'estimation paramétriques, notre travail peut s'appliquer à un grand nombre de problèmes en apprentissage automatique en temps réel.

ABSTRACT

This master's thesis aims at exploring the challenge of automatically retrieving musical structures within an audio file. Our main contribution is to formulate this well-studied problem in the framework of computational information geometry, an emerging field at the frontier between statistics, differential geometry, and data mining. In this framework, we unify the fundamental tasks of event segmentation and similarity computing in a single sequential scheme for structure discovery. Furthermore, we propose an original metric on temporal segments, which combines several criteria of geometrical comparability : divergence between centroids, as well as inclusion and intersection ratios of corresponding information balls. We test audio-based music structure discovery on two samples, differing in orchestration, genre, and sound representation. In both experiments, our metric compares favorably with all centroid-based dissimilarity measures. Since exponential families are ubiquitous in parametric estimation, our work encompasses a broad range of applications in online machine learning.

hal-00849736, version 1 - 17 Sep 2013

à quelle distance sommes-nous ?
emler
tchamitchian
echampard

REMERCIEMENTS

Je remercie avant tout mes encadrants de stage, Arshia Cont et Arnaud Dessein. Depuis le choix d'une thématique de recherches jusqu'à la correction de ce manuscrit, il m'ont apporté une aide précieuse et constante.

Merci beaucoup à Joakim Andén pour m'avoir assisté personnellement dans l'utilisation de son *audio scattering toolbox*.

Geoffroy Peeters m'a aidé à constituer un protocole d'évaluation comparée des métriques de structuration. Je le remercie pour ses conseils avisés.

C'est en échangeant avec Stéphane Mallat que j'ai pu faire un pas décisif dans mes recherches. Je le remercie pour le temps qu'il a consacré à ces discussions fructueuses.

Toujours disposé à me conseiller en mathématiques pour le traitement du signal, Philippe Cuvillier a été un interlocuteur privilégié au sein de l'équipe-projet MuTant à l'Ircam. Je le remercie pour l'esprit critique dont il a su faire preuve au cours de mon stage.

Lors de la phase de rédaction, j'ai pu bénéficier de la LaTeXpertise de mon collègue Pierre Donat-Bouillud, que je remercie chaleureusement.

Enfin, pour leur aide à la relecture, leur soutien bienveillant et leur gentillesse, je suis très reconnaissant à mes parents ainsi qu'à Claire.

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

TABLE DES MATIÈRES

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

TABLE DES FIGURES

ABRÉVIATIONS, SIGLES ET ACRONYMES

AudioBIFS *Audio BInary Format for Scene description*

CNRS Centre national de la recherche scientifique

DSP *Digital Sound Processing*

GDIF *Gesture Description Interchange Format*

GLR *Generalised Likelihood Ratio*

Inria Institut national de recherche en informatique et automatique

Ircam Institut de recherche et de coordination acoustique/musique

KL Kullback-Leibler

LR *Likelihood Ratio*

MIDI *Musical Instrument Digital Interface*

MIG *Music Information Geometry*

MIR *Music Information Retrieval*

MIREX *Music Information Retrieval Evaluation eXchange*

MFCC *Mel-Frequency Cepstral Coefficients*

MPEG *Moving Picture Experts Group*

MP3 *MPEG-1/2 Audio Layer 3*

NLP *Natural Language Processing*

OSC *Open Sound Control*

PCM *Pulse Code Modulation*

RDF *Resource Description Format*

RWC *Real World Computing*

SDM *Self-Distance Matrix*

XML *eXtensible Markup Language*

NOTATIONS

$\mathbf{v}$, $\mathbf{M}$ vecteur, matrice

$\mathcal{E}$ ensemble

$\mathscr{P}$ famille de probabilités

$\mathbb{P}_\lambda$ loi de probabilité de paramère $\boldsymbol{\lambda}$

$\mathbf{M}^\top$ matrice transposée de $\mathbf{M}$

$[\![a; b]\!]$ ensemble des entiers compris entre a et b

$\|\mathbf{v}\|_1$ somme des composantes de $\mathbf{v}$ en valeur absolue

$\mathbf{e}_k^n$ vecteur de $\mathbb{R}^n$ dont tous les échantillons sont nuls, sauf celui d'indice k, égal à 1

$\delta(z)$ symbole de Kronecker : vaut 0 partout, sauf en 0 où il vaut 1

$a \underset{H_1}{\overset{H_0}{\lessgtr}} b$ test statistique. On souscrit à l'hypothèse « nulle » H_0 si et seulement si $a < b$.

$\overline{\mathcal{E}}$ adhérence de l'ensemble $\mathcal{E}$

∇z gradient du champ scalaire z

$\mathbb{1}_{\mathcal{E}}$ fonction indicatrice de l'ensemble $\mathcal{E}$

$L^2(\mathbb{R})$ espace des fonctions de carré intégrable sur $\mathbb{R}$

$D^1(\mathbb{R})$ espace des fonctions dérivables sur $\mathbb{R}$

$\hat{y}$ transformée de Fourier du signal y (section 4.2 seulement)

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

1 INTRODUCTION

Ce travail vise à montrer que les outils théoriques de la géométrie de l'information sont aptes à produire un dispositif efficace de structuration d'un flux audio. Dans ce cadre, un des avantages notables du système que nous développons est qu'il peut fonctionner en temps réel sur une « machine d'écoute », ce qui suscite des opportunités inédites en création musicale mixte. Après avoir défini le problème de la découverte de structures, nous le plaçons dans le contexte de la recherche en informatique et traitement du signal musical, dont nous dressons un bref état de l'art. Enfin, nous répertorions les principales contributions personnelles apportées au système de segmentation audio initialement conçu par nos encadrants, Arshia Cont et Arnaud Dessein.

1.1 QU'EST-CE QUE LA STRUCTURE MUSICALE ?

1.1.1 *Écoute et structure*

L'écoute musicale se caractérise par une propension à percevoir la structure d'une œuvre selon plusieurs échelles temporelles. En particulier, elle accède sans peine à la contexture de ses reprises, variations et renvois, et en produit une représentation régulière et cohérente. Dès lors, la mémoire de motifs passés génère des affects de sérénité ou de surprise, qui conditionnent la perception des évènements à venir. Dans l'expérience commune, le phénomène sonore est indissociable de sa structure temporelle sous-jacente.

À chacune de ces échelles, l'écoute décèle un plan de cohérence, strié par des contrastes, replié sur lui-même par la répétition de motifs, et dont le développement capte l'attention de l'auditeur. Nous définissons la structure musicale au sens large comme la mise en relation des plans de cohérence de l'œuvre écoutée ; et, partant, la découverte de structures comme l'estimation adéquate de cette mise en relation.

Dans ce travail, nous avons voulu donner une place prépondérante au temps, la seule donnée commune à toutes les manifestations de l'activité musicale. C'est à travers lui que nous tâchons d'équilibrer les qualités perçues pour concevoir une forme qui dépasse les représentations partielles. Ainsi, la découverte de structures s'insère doublement dans une problématique temporelle : d'une part, pour quantifier la ressemblance entre des instants divers ; d'autre part, pour produire le temps musical, celui du rythme.

Nous pensons que les facultés de coordination induites par la découverte de structures constituent, en association avec des capacités sensori-motrices de contrôle du corps et un comportement d'apprentissage persévérant, les éléments essentiels de la pratique musicale. Certes, l'oreille a évolué afin de développer d'abord des avantages sélectifs adaptés à son environnement, tels que la localisation de mouvement, la détection du danger et, chez l'homme, la compréhension de la parole. Mais ce processus morphogénétique spécifique nous a conduit à acquérir une faculté, polyvalente et abstraite, d'écoute, c'est-à-dire essentiellement de formation de structure. Nous défendons l'écoute comme l'entité centrale de la musique, et donc le point de départ de son analyse — voir à ce sujet Deliège (2010).

Or, une part croissante, et quasiment incontournable, de notre expérience musicale, transite aujourd'hui par l'ordinateur[1]. À la différence des lecteurs magnétiques ou optiques, ce dernier n'est pas qu'un simple transducteur de contenu multimédia, puisqu'il est en mesure d'exécuter des opérations logiques et arithmétiques complexes. Par conséquent, de même que le microphone incarne un analogue évident du tympan, on pourrait se donner l'objectif de simuler la réponse électrique de la cochlée humaine et les processus neuronaux ultérieurs par une « machine d'écoute » artificielle. Une telle machine contribuerait à rendre le musicien virtuel (*synthetic performer*) de Vercoe (1990) aussi fiable qu'un virtuose humain, même dans les contextes les plus ardus ; c'est en fait un rêve qui rappelle des spéculations publiées par Ada Lovelace en 1843, bien avant la construction des premiers ordinateurs.

Les avancées présentes en psychoacoustique et neurologie de l'audition sont beaucoup trop éloignées d'une description systématique pour permettre une imitation cybernétique exacte de la découverte de structures dans les sons. Cependant, les outils du traitement du signal et de l'apprentissage automatique (*machine learning*) donnent des modèles fonctionnels approximatifs de l'écoute, aptes à réaliser des tâches spécifiques avec plus ou moins d'acuité. Nous souhaitons poursuivre cet effort de recherche en présentant un nouveau système de visualisation des répétitions dans un flux audio quelconque.

1.1.2 *Niveaux de représentation de l'information musicale*

Les technologies actuelles stockent, traitent et diffusent les données multimédias selon un paradigme unique, hérité de Shannon : le codage binaire de l'information. Par conséquent, il est envisageable

1. Par *ordinateur*, nous entendons tout dispositif électronique doté d'un processeur, d'une mémoire vive et de ports de connectivité. Ainsi, les récentes générations de téléphones portables, de baladeurs numériques, de tablettes électroniques, de consoles de jeu et de nombreux autres systèmes embarqués peuvent être vus comme des ordinateurs au sens large, au même titre que les ordinateurs de bureau et les serveurs informatiques traditionnels.

hal-00849736, version 1 - 17 Sep 2013

Niveau de représentation		Bande passante	Exemple de format
Connaissance		0 Hz – 1 mHz	XML, RDF
Symbolique	Cellule	10 mHz – 1 Hz	MIDI, OSC
	Pulsation	1 Hz – 3 Hz	
	Note	0.5 Hz – 10 Hz	
Signal		20 Hz – 20 kHz	MP3, PCM
Physique		illimitée	AudioBIFS, GDIF

TABLE 1.1.1: Niveaux de représentation de l'information musicale (Vinet, 2003), bandes passantes typiques et exemples de formats numériques associés.

d'utiliser l'ordinateur pour associer des représentations musicales apparemment hétérogènes. À titre d'exemple, sur une puce de mémoire vive, un enregistrement stéréophonique peut parfaitement côtoyer sa transcription en fichier MIDI ainsi que des méta-données sémantiques RDF le concernant (titre, artiste, année, etc.), tandis que les noyaux de calcul du processeur opèrent conjointement sur ces trois documents.

Vinet (2003) définit quatre niveaux de représentation technique de l'information musicale : physique, signal, symbolique et connaissance (voir table 1.1.1). Notre définition de la découverte de structures chez l'humain s'adapte alors immédiatement à l'ordinateur : étant donné un flux d'information musicale, il s'agit de proposer des *conversions* adéquates entre ces niveaux de représentation. La discipline scientifique qui vise à produire de telles conversions porte le nom anglophone de *Music Information Retrieval* (MIR) (Downie, 2003). C'est un sujet vaste, abondamment étudié au cours des dix dernières années, et qui se ramifie en un grand nombre de problématiques d'ingénierie. L'une d'entre elles, baptisée *structure retrieval* en anglais, s'attache à identifier les répétitions et variations dans un signal musical, dont la durée est de l'ordre de quelques secondes à quelques minutes. Cette seconde définition, plus modeste que la précédente et à laquelle nous nous cantonnerons dans le reste de ce travail, n'englobe pas toutes les échelles temporelles intervenant dans l'écoute, mais isole et compare des segments temporels appelés *modèles* afin d'en extraire un regroupement hiérarchique. En sortie d'un système de *structure retrieval*, on s'attend à obtenir un parcours sur un graphe inconnu *a priori*, chaque état symbolisant une cellule musicale inédite. Les parties redondantes de l'œuvre — refrains, leitmotivs, ostinatos, etc. — correspondent alors aux états visités de nombreuses fois.

De la durée d'un échantillon numérique à la plus grande cellule structurelle, éventuellement longue de plusieurs minutes, le rapport d'échelle est de l'ordre du million. Or, les notions de répétition, de

hal-00849736, version 1 - 17 Sep 2013

contraste, de variation ou de monotonie ne prennent un véritable sens sur le signal que sur des durées dépassant le dixième de seconde ; par ailleurs, les cellules principales sont elles-mêmes composées d'évènements sonores plus petits, tels que des notes d'instruments de musique. Pour permettre une structuration pertinente, il est donc nécessaire de construire un ou plusieurs niveaux intermédiaires (*mid-level*) de représentation du son, chaque niveau réalisant une intégration de l'information au niveau précédent, selon des durées caractéristiques croissantes.

1.2 MOTIVATION

1.2.1 *L'ordinateur qui écoute comme un musicien*

C'est au niveau du signal que l'information musicale est la plus répandue parmi le grand public, sous forme de fichiers stéréophoniques compressés. Avec la massification des bases de données de signaux musicaux en ligne — certaines atteignant aujourd'hui plusieurs millions de titres — et la présence, à l'heure des téléphones portables, d'un microphone dans toutes les poches ou presque, le traitement du signal musical a de beaux jours devant lui, et peut viser une implémentation à très grande échelle. Plus particulièrement, la détection automatique de structures a indéniablement prouvé son utilité dans au moins deux applications pour les utilisateurs : la génération automatique de résumé (Peeters et al., 2002) et l'amélioration de la consultation de catalogues multimédias, en particulier depuis l'instauration de la norme MPEG-7 (Quackenbush et Lindsay, 2001). C'est toutefois dans un autre contexte, plus proche de la création musicale contemporaine, que s'inscrit ce travail de stage.

L'équipe-projet MuTant, fondée en 2012 selon une conjonction entre l'Inria, l'Ircam et le CNRS[2], vise à modéliser les facultés fondamentales de musiciens humains pour les incorporer dans des systèmes informatiques dédiés à la composition ou à la performance. MuTant est notamment en charge du développement d'Antescofo (Cont, 2008a), un musicien virtuel capable de réagir aux fluctuations de tempo de ses partenaires. Outre un système robuste de programmation synchrone, Antescofo repose sur une machine d'écoute artificielle qui analyse les évènements sonores au fur et à mesure qu'ils se produisent. Ce logiciel assouplit considérablement l'expressivité des œuvres mixtes, qui mélangent exécution instrumentale et réalisation informatique. En effet, bien qu'elle séduise souvent les compositeurs par son extrême précision et ses vertus de reproductibilité, la « bande magnétique » ne bénéficie pas de boucle de rétroaction, ce qui compromet

2. L'Inria est l'Institut national de recherche en informatique et automatique. L'Ircam est l'Institut de recherche et de coordination acoustique/musique. Le CNRS est le Centre national de la recherche scientifique.

hal-00849736, version 1 - 17 Sep 2013

l'alignement rythmique avec les musiciens (Cont, 2012). On fait donc appel à un réalisateur pour ajuster en permanence la partition sur le flux sonore, sujet notamment à des fluctuations de tempo et des points d'orgue non mesurés. Mais, dès les années 1980, le compositeur Philippe Manoury a pu constater que faire porter « manuellement » la contrainte de synchronicité sur un nombre restreint de points de la partition ne suffisait pas à unir durablement l'interprétation humaine et le traitement électronique (Manoury, 2007).

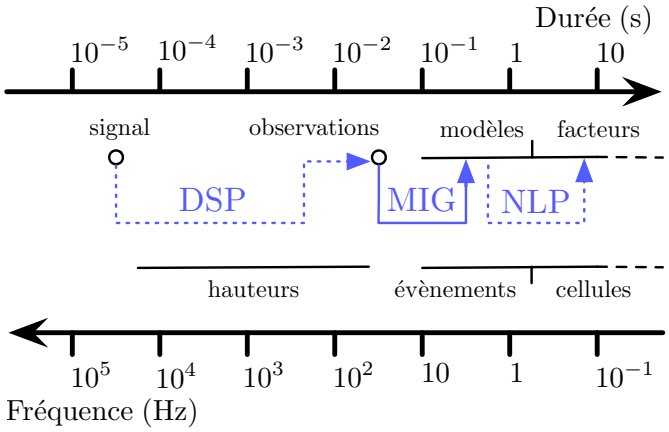
Aujourd'hui, Antescofo est utilisé sur un répertoire de plusieurs dizaines d'œuvres contemporaines, par des compositeurs d'envergure internationale tels que Marco Stroppa, Pierre Boulez, Philippe Manoury et Kaija Saariaho. Malgré ce plébiscite, il n'a pas encore atteint son stade de maturité, en particulier en ce qui concerne la modélisation du signal musical ; l'une des pistes d'amélioration réside dans l'extension du suivi à des contenus partiellement ou totalement improvisés, mais fondés sur des logiques structurelles longues. Notre stage au sein de l'équipe MuTant vise à proposer un module d'écoute artificielle capable de comprendre, en partie, ces logiques.

1.2.2 *Géométrie de l'information musicale*

Certains systèmes de structuration postulent l'existence de classes naturelles pour les signaux de musique, dans lesquelles les évènements n'ont plus qu'à être rangés : on parle de structuration *supervisée*. À l'inverse, dans un contexte non supervisé, on ne peut pas construire ces classes à l'avance puisque l'on ne se donne aucun *a priori* sur le flux sonore. Dans cette seconde approche, que nous poursuivons dans tout ce travail, l'objectif de la découverte de structures est double. D'une part, il faut extraire, au niveau du signal, ce que le musicien appelle « note » ou, pour employer un terme plus général, *évènement*, en tant qu'unité symbolique minimale d'expression. D'autre part, il faut une procédure permettant de déterminer si deux évènements sont égaux ou pas, en accord avec un critère de perception humaine. Pourvu que l'une et l'autre soient gérées efficacement, il n'est pas difficile de déterminer quelles combinaisons d'évènements sont des *facteurs* récurrents de l'œuvre (Cont et al., 2010).

Ces tâches, appelées *segmentation* et *classification*, opèrent sur des critères complémentaires. Tandis que la première vise à reconnaître les ruptures locales le long du signal, la seconde trouve des ressemblances globales entre des évènements non adjacents. Nous verrons que les outils du traitement du signal et des statistiques nous donnent une fonction de dissimilarité naturelle entre des observations du signal à des instants différents, chaque observation correspondant à une durée d'environ 20 ms. Ainsi, la segmentation se ramène à la détection des dépassements d'un seuil de dissimilarité intra-segment. Dans le cadre de son application au traitement de la musique, l'idée

hal-00849736, version 1 - 17 Sep 2013



DSP : *Digital Signal Processing* (traitement du signal)

MIG : *Music Information Geometry* (géométrie de l'information musicale)

NLP : *Natural Language Processing* (traitement du langage naturel)

FIGURE 1.2.1: Schéma des principales échelles temporelles du problème. Les objets de la partie supérieure correspondent au temps physique, tandis que ceux de la partie inférieure correspondent au temps musical.

de la géométrie de l'information est d'étendre cette mesure locale à l'échelle symbolique de l'évènement, dont la durée typique est de l'ordre de 400 ms. On peut dès lors associer à chaque segment un *modèle* de dimension raisonnablement faible, et utiliser la mesure étendue pour comparer les modèles. Si la dissimilarité entre deux modèles est en-dessous d'un certain seuil, on leur associe le même symbole. Puisqu'elle réutilise les calculs effectués pendant la segmentation pour discriminer les modèles, une telle entreprise est particulièrement intéressante dans une optique d'implémentation en temps réel [3].

Sur la figure 1.2.1, on a placé les différents niveaux de représentation de l'information musicale intervenant dans ce problème ainsi que leurs durées typiques respectives. On y voit que la géométrie de l'information (MIG pour *Music Information Geometry*) joue un rôle de charnière entre le traitement du signal (DSP pour *Digital Signal Processing*) et le traitement du langage naturel (NLP pour *Natural Language Processing*) qui opèrent respectivement au niveau du signal et des symboles.

Initiée en 1945 par Rao, et développée en tant que telle pendant les années 1980 par Chentsov, Eguchi, Amari et Nagaoka, la géométrie de l'information consiste en la formulation de problèmes statistiques

3. Au cours de ce stage, nous avons tâché de concevoir des programmes informatiques assez peu coûteux en ressources pour traiter des fichiers sonores en moins de temps que la durée de ceux-ci. Au sens strict du terme, pourtant, le temps réel est une contrainte beaucoup plus difficile à assurer formellement. Son étude a donné naissance au paradigme de programmation synchrone (Berry et al., 1986), dont se nourrit notamment Antescofo.

hal-00849736, version 1 - 17 Sep 2013

à l'aide de notions empruntées à la géométre différentielle. C'est dans sa thèse de doctorat (Cont, 2008b) qu'Arshia Cont entreprend, de façon quasi inédite, d'appliquer des méthodes de géométrie de l'information aux signaux de musique, en voyant les observations sur le signal comme les paramètres d'un processus aléatoire X sur un ensemble $\mathcal{X}$.

Étant donnée une séquence de réalisations aléatoires $\Xi = \left(x^1 \dots x^n\right)$, le modèle canonique associé à Ξ est l'estimateur $\widetilde{\boldsymbol{\lambda}}$ qui en maximise la vraisemblance, c'est-à-dire tel que la quantité

$$L_\Xi\left(\boldsymbol{\lambda}\right) = \mathbb{P}_{\boldsymbol{\lambda}}\left[\left(X^1 \dots X^n\right) = \Xi\right] \tag{1.2.1}$$

soit maximale en $\widetilde{\boldsymbol{\lambda}}$. Dès lors, la comparaison de portions temporelles observées se ramène à la comparaison géométrique de leurs estimateurs ; la détection de dans la séquence, à un test d'homogénéité statistique ; et la découverte de structures, à un algorithme de recherche par proximité (*proximity search*).

Un résultat important dû à Fisher, Darmois, Koopman et Pitman (1934–1936) donne une condition sur la famille $\mathscr{P} = \left(\mathbb{P}_{\boldsymbol{\lambda}}\right)$ pour assurer l'existence d'une statistique exhaustive $\mathbb{T}$ avec un nombre borné m de composantes scalaires, et ce pour un processus de variables aléatoires indépendantes et identiquement distribuées dont le nombre tend vers l'infini. Sous des hypothèses peu restrictives, $\mathscr{P}$ possède une telle statistique exhaustive si et seulement si elle s'exprime comme une famille exponentielle :

$$\forall t \in [\![1; n]\!],\ \forall x \in \mathcal{X},\ \mathbb{P}_{\boldsymbol{\lambda}^t}\left[X^t = x^t\right] = \exp\left(V\left(x\right)^\top \boldsymbol{\theta}^t - F\left(\boldsymbol{\theta}^t\right)\right) \tag{1.2.2}$$

où $\boldsymbol{\theta}^t$ est relié à $\boldsymbol{\lambda}^t$ de façon déterministe, et où V et F sont des fonctions connues.

Seules les familles exponentielles offrent cette statistique exhaustive, ce qui rend leur utilisation pratique, notamment sous une contrainte de temps réel. Par ailleurs, on verra que $\mathbb{T}$ permet de construire directement l'estimateur du maximum de vraisemblance du processus. Les chapitres 2 et 3 s'inscrivent dans ce cadre mathématique général, et leur portée dépasse la seule application à la découverte de structures en musique. Nous envisageons notamment, à l'issue de ce projet, d'en transposer les résultats au traitement des images, afin d'aboutir à une structuration des plans dans un signal vidéo.

1.3 ÉTAT DE L'ART

1.3.1 *Structuration de locuteurs*

Voici près d'un demi-siècle que la création musicale contemporaine confie à l'ordinateur la tâche de générer des sons en temps réel. En ce sens, la synthèse numérique poursuit le processus de diversification

hal-00849736, version 1 - 17 Sep 2013

des timbres initié au début du XX^{e} siècle par le recours à des modes de jeu étendus, à une nouvelle lutherie et aux outils de synthèse analogique. Par contre, l'analyse musicale de signal audio est une discipline relativement nouvelle : la comparaison de données de grande dimension sur de longues durées, *a fortiori* selon des contraintes de temps réel, représente un coût computationnel que seules les générations récentes d'ordinateurs peuvent affronter.

Les premiers systèmes de découverte automatique de structures dans les sons étaient dévolus au traitement de la parole, et consistaient en une structuration non supervisée de locuteurs (*speaker diarisation*) dans l'enregistrement audio d'une conversation — voir notamment Gish et al. (1991). À l'écoute d'un débat radiophonique, l'oreille perçoit sans peine les arrêts et reprises dans le discours des participants, fussent-ils tous inconnus. C'est une faculté qui semble simple, mais qui peut être vue comme une forme rudimentaire d'écoute musicale, et qui constitue un bon exemple de découverte de structures temporelles. Par ailleurs, puisqu'elle constitue le fondement de nombreuses problématiques en traitement de la parole — transcription, indexation, décompte de temps, etc. — elle fait aujourd'hui l'objet d'efforts intensifs de recherche (Anguera Miró et al., 2012).

Comme notre procédé de découverte de structures musicales, la structuration non supervisée de locuteurs se décompose le plus souvent en une phase de segmentation et une phase de partitionnement. Mais tandis qu'un monologue présente des paramètres de timbre stationnaires sur son intégralité, la quasi-stationnarité en musique ne dure souvent qu'une fraction de seconde tout au plus, le temps d'une note. Les segments sont des entités beaucoup plus courtes et contiennent donc moins d'information, ce qui complique la tâche de reconnaissance. En contrepartie, les évènements musicaux commencent le plus souvent par une phase d'activation brutale appelée *attaque*, qui contraste fortement avec celle qui la précède. La segmentation du signal musical est donc une tâche relativement facile : d'ailleurs, dans beaucoup de systèmes d'extraction d'information musicale symbolique, celle-ci est remplacée par une détection des activations (*onset detection*).

1.3.2 *Annotation et visualisation de structures*

Ce furent Tzanetakis et Cook (1999) qui les premiers apportèrent une méthodologie pour l'annotation et l'évaluation de systèmes de découverte automatique de structure. Dans la lignée des travaux de Scheirer et Slaney (1998) sur la discrimination entre parole et musique, ils utilisèrent un ensemble de cinq descripteurs de forme — centroïde spectral, étalement spectral, flux spectral, taux de passage par zéro et intensité — pour décrire le timbre du signal sur des observations de 20 ms. Le calcul de la moyenne et de la variance de ces

hal-00849736, version 1 - 17 Sep 2013

observations sur des fenêtres longues d'une seconde aboutit à un flux de données $\widetilde{\lambda}^t$ de dimension 10 ; la segmentation reposait alors sur un dépassement de seuil portant sur la dérivée de la distance entre deux $\widetilde{\lambda}^t$ successives. Même si, dans le schéma de Tzanetakis et Cook, la segmentation intervenait en *aval* de la construction des modèles, on voit déjà, dans l'idée d'étudier l'évolution des observations sur une durée grande devant leurs variations transitoires, l'intention de passer du niveau signal au niveau symbolique. Pour comparer deux modèles $\widetilde{\lambda^{t_1}}$ et $\widetilde{\lambda^{t_2}}$, les auteurs employaient une distance de Mahalanobis

$$d_\Sigma\left(\widetilde{\lambda^{t_1}} \| \widetilde{\lambda^{t_2}}\right) = \widetilde{\lambda^{t_2}}^\top \Sigma^{-1} \widetilde{\lambda^{t_2}}, \tag{1.3.1}$$

où Σ est la matrice de covariance de $\widetilde{\lambda}$ sur l'intégralité du fichier audio. De l'aveu des auteurs, cette distance n'a pas de sens physique évident puisque chaque dimension y joue un rôle égal, ce qui est une hypothèse peu réaliste. Par ailleurs, la dérivation temporelle étant ici une soustraction d'éléments consécutifs, l'étape de segmentation manquait de robustesse. Toutefois, en confrontant les résultats obtenus aux performances de sujets humains, cet article prouve que la frontière entre autocorrélation physique et structure musicale n'est pas infranchissable.

L'autre article pionnier du domaine fut publié par Foote en 1999. Celui-ci ne segmentait pas le signal mais construisait des pseudo-modèles par moyenne glissante sur les observations, longs de 100 ms. Ensuite, il les comparait systématiquement en calculant une matrice d'auto-distance (SDM pour *Self-Distance Matrix*) entre chaque paire. Si la notion de matrice d'auto-distance pour des systèmes dynamiques, alors sous le nom de *recurrence plot*, était déjà introduite par Eckmann et al. (1987), Foote fut le premier à l'appliquer à des sons. La distance utilisée est le produit scalaire des deux vecteurs, normalisé par le produit de leurs normes respectives. Même si aucune démarche de structuration automatique n'est entreprise, l'auteur proposa de visualiser la matrice pour y découvrir « à la main » des zones rectangulaires de faible distance, c'est-à-dire les cellules structurelles de la musique. Pour un extrait musical d'une minute, la matrice de similarité de Foote comporte autour de 360 000 points. Ce nombre augmentant quadratiquement avec le temps, il est difficilement envisageable, même au vu des technologies actuelles, d'implémenter un tel protocole en temps réel. À la section 4.1, nous présentons une matrice d'auto-distance telle que l'aurait construite Foote à l'époque, afin de la comparer avec notre matrice de visualisation.

1.3.3 *Approche par séquences et approche par états*

Dans la perspective de générer automatiquement le « résumé audio » de chaque chanson d'une grande base de chansons, Peeters

hal-00849736, version 1 - 17 Sep 2013

	homogène	non homogène
répété	par états	par séquences
non répété	par états	par états (*états-poubelles*)

TABLE 1.3.1: Hypothèses d'homogénéité et de répétition sur les modèles et approches de structuration correspondantes.

(2004) dressa un premier état de l'art de la découverte de structures en musique. L'article distingue deux approches possibles. La première consiste à rechercher des répétitions de séquences dans la matrice d'auto-distance, sous forme de sous-diagonales de coefficients faibles ; d'où sa qualification d'*approche par séquence*s. La seconde se base sur une mesure d'innovation (*novelty measure*) portant sur les observations, que l'on suppose faible à l'intérieur des cellules structurelles et élevée à leurs extrémités (Foote, 2000). Si ces cellules présentent des disparités évidentes en termes d'instrumentation, de rythme ou de tonalité, on peut penser que cette mesure d'innovation pourra les délimiter. On vise alors à passer directement du niveau du signal aux états du graphe symbolique, sans faire appel à une échelle intermédiaire ; on parle donc d'*approche par état*s.

Comme le fait remarquer l'auteur, la différence entre ces deux types d'approche réside essentiellement dans la durée des modèles choisis : si cette dernière est faible devant la durée typique des cellules musicales à identifier, celles-ci paraîtront hétérogènes, d'où la nécessité d'une approche par séquence. À l'inverse, si la mesure de dissimilarité utilisée opère à plus long terme, les cellules prendront la forme de blocs homogènes. Depuis, l'auteur a proposé une méthode permettant de déterminer, au vu des descripteurs sonores choisis, laquelle des deux approches est la plus adaptée (Peeters, 2011). Par ailleurs, à l'occasion de son habilitation à diriger des recherches, il a récemment publié un large aperçu des méthodes en découverte de structures dans les flux sonores (Peeters, 2013). Nous renvoyons à ce mémoire pour une étude bibliographique plus approfondie.

Du fait de leur filiation avec la structuration de locuteurs, c'est l'approche par états qui est de loin la plus répandue dans la littérature scientifique. Son avantage majeur est que les sections non répétées, même hétérogènes, seront tout de même agglomérées en « états-poubelles », après que l'on aura posé les identités entre états homogènes (voir la case inférieure droite de la table 1.3.1). On s'épargne ainsi l'étape de détection de sous-diagonales, une tâche hasardeuse et peu robuste aux déformations temporelles telles que le swing ou les variations de tempo. Malheureusement, il n'est pas évident de définir une mesure d'innovation générique à l'échelle de la phrase musicale. Dans les répertoires rock ou pop, chaque cellule — couplet,

hal-00849736, version 1 - 17 Sep 2013

refrain, pont, etc. — possède un timbre homogène et suffisamment distinct des autres, ce qui facilite leur segmentation. C'est pourquoi les expériences de structuration par états se sont concentrées sur le traitement des chansons jusqu'à présent.

Les recherches récentes dans le domaine visent à rendre l'approche par états plus sensible à l'hétérogénéité intrinsèque des sons musicaux. D'après les résultats de la campagne d'évaluation MIREX 2012, le système actuel le plus performant (Serrà et al., 2012) opère sur un critère combiné d'homogénéité, de répétition et de contraste, inspiré par la typologie de Paulus et al. (2010). En deuxième position, Kaiser et Peeters (2013) proposent d'envisager plusieurs hypothèses concurrentes de transition entre états sur la mesure d'innovation, et ce à différentes échelles temporelles.

En ce qui nous concerne, nous pensons que l'échelle de l'évènement ($T \approx 400\,\text{ms}$) s'impose naturellement comme la limite d'homogénéité sonore, de même que l'échelle de l'observation ($T \approx 20\,\text{ms}$) s'impose comme la limite de stationnarité spectrale. Par conséquent, les « états » que nous dégageons ne sont pas les cellules structurelles de l'œuvre, mais plutôt leurs briques mélodiques fondamentales. Pour les agréger, il ne sera pas nécessaire d'entreprendre une détection de sous-diagonales : des méthodes de traitement du langage naturel, initialement prévues pour le texte ou le génome, se révèlent mieux adaptées. Avec l'équipe-projet MuTant, nous avons l'ambition de découvrir les structures, non seulement des chansons, mais aussi des œuvres de musique classique ou contemporaine.

1.4 PLAN ET CONTRIBUTIONS PRINCIPALES

1.4.1 *Plan du rapport de fin de stage*

Le chapitre 2, consacré à la détection séquentielle de rupture, est issu d'un cadre développé par Dessein à l'occasion de sa thèse de doctorat, soutenue en 2012. De portée moins générale que cette dernière, il peut toutefois être lu comme un tutoriel d'introduction à la géométrie de l'information à l'usage des étudiants et chercheurs en traitement du signal audionumérique. On y aborde, pas à pas, les notions de famille exponentielle standard, de statistique exhaustive et de rapport de vraisemblance généralisé, tout en soulignant à chaque fois la légitimité de leur emploi.

Au chapitre 3, nous introduisons les divergences de Bregman et explicitons leur lien fort avec les familles exponentielles. Nous présentons alors trois algorithmes, issus d'articles scientifiques récents en géométrie de l'information (Cayton, 2009; Nock et Nielsen, 2005) portant sur les boules de Bregman. Chacun de ces algorithmes est illustré par une expérience portant sur un extrait musical simple. Nous proposons enfin une métrique de comparaison entre les modèles qui

hal-00849736, version 1 - 17 Sep 2013

prend en compte plusieurs facteurs géométriques, dont notamment l'inclusion et l'intersection de boules informationnelles.

Le chapitre 4, enfin, décrit deux expériences de découvertes de structures, portant sur des extraits musicaux aux caractéristiques très différentes : l'un est une œuvre classique pour piano ; l'autre, un thème de jazz interprété par huit musiciens. Pour ce second extrait, on abandonne l'observation par spectre de Fourier au profit du spectre de scattering (Andén et Mallat, 2011), une représentation à plus long terme fondée sur la théorie des ondelettes. Dans les deux expériences, la métrique mixte améliore la découverte de structures par rapport à une comparaison de moyennes sur les observations.

1.4.2 *Contributions principales*

Ce travail de stage nous a donné l'occasion de développer un système de découverte de structures en temps réel. Bien qu'il demeure à l'état de prototype et s'avoue très loin d'une ambition industrielle, il incorpore des résultats puissants en géométrie de l'information, dont beaucoup sont appliqués aux signaux de musique pour la première fois. En particulier,

- la conception d'un algorithme itératif de projection sur la frontière d'une boule de Bregman, inspirée de la méthode de Newton et faisant intervenir la matrice d'information de Fisher (section 3.2),
- les notions de *rapports d'inclusion* et *d'intersection*, calculés grâce à l'algorithme itératif précédent (sections 3.2 et 3.3),
- la construction d'une *métrique mixte*, combinant les divergences de Bregman (voir section 3.1) avec les critères d'inclusion et d'intersection (section 3.4),
- la mise en place d'une évaluation des apports de la géométrie de l'information dans le cas, très répandu en traitement du signal, d'observations sur le spectre de Fourier (section 4.1), et
- la mise en place de cette évaluation dans le cas, moins fréquent mais plus compétitif, d'observations sur le spectre de scattering (section 4.3), confronté à un signal de musique plus complexe,

en sont les principales contributions nouvelles.

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

2

SEGMENTATION DE FLUX D'HISTOGRAMMES

Dans ce chapitre, nous présentons un algorithme qui détecte les changements brusques dans un flux de données multidimensionnel, et ce au fur et à mesure qu'ils se produisent. Celui-ci repose sur le rapport de vraisemblance généralisé, une notion statistique que nous revisitons avec le formalisme de la géométrie de l'information. Nous en explicitons les équations principales dans le cas particulier de données en histogrammes, et étudions son implémentation dans le cadre de la segmentation en évènements musicaux d'un spectrogramme sonore.

La problématique de la détection de rupture (*abrupt change detection*) dans un flux de données trouve son origine dans le domaine du contrôle de qualité. Afin de s'assurer du bon déroulement d'un processus industriel, il est important de pouvoir en déceler les pannes de façon rapide et efficace ; d'où l'intérêt d'étudier statistiquement l'évolution des paramètres de contrôle autour de leur valeur théorique. La plupart de ces approches est fondée sur l'estimation de rapports de vraisemblance (LR pour *Likelihood Ratio*), que celle-ci soit bayésienne (Girshick et Rubin, 1952) ou non (Page, 1954). Toutefois, s'agissant de signaux de musique, on ne dispose pas d'un régime « normal » de fonctionnement : si rupture il y a, il faut estimer sa position en même temps que la valeur estimée des paramètres avant et après celui-ci. C'est pour aborder ce type de contexte que Lorden (1971) a introduit le rapport de vraisemblance *généralisé*. Bien qu'André-Obrecht (1986) ait fait appel aux rapports de vraisemblance généralisés pour segmenter un signal de parole en syllabes, c'est Dessein (2012, chapitres 2 et 3) qui propose de diversifier leur champ d'application à de nombreuses classes de problèmes en traitement du signal audionumérique, et fournit un cadre unificateur pour les résoudre sous des contraintes de temps réel. En ce qui nous concerne, nous ne visons pas à la plus grande généralité, mais à définir progressivement les outils de géométrie de l'information nécessaires au dispositif de segmentation en évènements. Nombre de ces outils se révèleront essentiels au chapitre suivant, qui sera consacré à une comparaison entre les segments. Nous renvoyons à l'article de Polunchenko et Tartakovsky (2012) pour un état de l'art récent sur la détection séquentielle de rupture.

2.1 MODÉLISATION ALÉATOIRE

2.1.1 *Espace des paramètre sources*

Considérons un flux d'histogrammes $\boldsymbol{h}^t$ de dimension $m > 1$, où les instants temporels t sont des entiers consécutifs. Ces histogrammes peuvent, par exemple, quantifier la distribution de l'énergie d'un signal sur m bandes de fréquences, selon une fenêtre temporelle glissante. Dans les deux premières sections de ce chapitre, t est un instant particulier arbitraire. On note $\mathcal{X} = [\![1; m]\!]$ l'ensemble des entiers de 1 à m. On appelle *paramètre source* le vecteur $\boldsymbol{\lambda}^t$ strictement positif résultant d'une normalisation de $\boldsymbol{h}^t$ à 1 :

$$\forall x \in \mathcal{X},\ \lambda_x^t = \frac{h_x^t + \epsilon}{\left\|\boldsymbol{h}^t\right\|_1 + m\epsilon}. \tag{2.1.1}$$

Le réel $\epsilon > 0$ est choisi relativement petit devant l'amplitude des composante h_x^t en présence d'un signal. Il a notamment pour rôle d'éviter les indéterminations numériques dues à la division par zéro dans le cas où $\boldsymbol{h}^t$ s'annule. L'ajout d'une constante sur chaque composante de $\boldsymbol{\lambda}^t$ est une opération classique en traitement du signal, qui porte le nom de *blanchiment spectral*. Par ailleurs, on impose à chaque λ_x^t d'être dans l'ouvert $]0; 1[$, ce qui va permettre, dès la section 2.2, de définir les logarithmes naturels $\ln \lambda_x^t$ et $\ln\left(1 - \sum_{k=1}^{m-1} \lambda_x^t\right)$, et donc une famille exponentielle $\mathscr{P}$ homéomorphe au domaine

$$\mathcal{S} = \left\{\boldsymbol{\lambda} \in]0; 1[^m \,|\, \|\boldsymbol{\lambda}\|_1 = 1\right\}, \tag{2.1.2}$$

appelé l'*espace des paramètres sources*. Dans $\mathcal{S}$, tout histogramme normalisé à m classes, strictement positif, trouve une représentation univoque.

2.1.2 *Simplexe des moments*

Bien que $\mathcal{S}$ soit de dimension m, le nombre de degrés de liberté de la famille $\boldsymbol{\lambda}^t = \left(\lambda_x^t\right)_{x\in\mathcal{X}}$ n'est que de $(m-1)$, en raison de la condition de normalisation

$$\|\boldsymbol{\lambda}^t\|_1 = \sum_{x=1}^{m} \lambda_x^t = 1 \tag{2.1.3}$$

résultant de l'équation 2.1.1. Le paramètre source est donc redondant d'un point de vue informationnel. Soit $\boldsymbol{\eta}^t$ le vecteur composé des $(m-1)$ premières composantes de $\boldsymbol{\lambda}^t$. Grâce à la relation linéaire 2.1.3, l'intégralité du vecteur $\boldsymbol{\lambda}^t$ est reconstruite sans ambigüité à partir de $\boldsymbol{\eta}^t$:

$$\begin{cases} \forall k \in [\![1; m-1]\!], & \lambda_x^t = \eta_x^t \\ \text{et} & \lambda_m^t = 1 - \|\boldsymbol{\eta}^t\|_1 \end{cases} \tag{2.1.4}$$

hal-00849736, version 1 - 17 Sep 2013

L'ensemble de définition $\mathcal{M}$ de $\boldsymbol{\eta}^t$ est le simplexe ouvert construit sur le repère orthonormé canonique de l'espace $\mathcal{N} = \mathbb{R}^{m-1}$.

$$\mathcal{M} = \left\{ \boldsymbol{\eta} \in]0;\, 1[^{m-1} \mid \|\boldsymbol{\eta}\|_1 < 1 \right\} \tag{2.1.5}$$

Le vecteur $\boldsymbol{\eta}^t$ est appelé *paramètre des moments*, car il est étroitement lié à l'espérance — le moment d'ordre 1 — de la variable aléatoire X^t ; cette dénomination est explicitée à l'équation 2.3.7. L'espace $\mathcal{M}$ est un moyen, non unique, de représenter l'observation à l'instant t tout en minimisant la quantité d'information requise. Tandis que $\mathcal{S}$ présente une dimension « dégénérée », il est beaucoup plus commode de faire de la géométrie différentielle sur $\mathcal{M}$.

2.1.3 *Distribution catégorique*

Il est possible de voir $\boldsymbol{\lambda}^t$ comme le paramètre de la loi $\mathbb{P}_{\boldsymbol{\lambda}^t}$ d'une variable aléatoire discrète X^t portant sur les entiers de $\mathcal{X}$.

$$\forall x \in \mathcal{X},\ \mathbb{P}_{\boldsymbol{\lambda}^t}\left[X^t = x\right] = \lambda_x^t \tag{2.1.6}$$

Puisque les entiers de $\mathcal{X}$ numérotent m catégories qualitatives, X^t est une variable aléatoire dite *catégorique*. Grâce à l'équation 2.1.4, la loi de X^t est définie à l'aide du paramètre des moments $\boldsymbol{\eta}^t$. La relation précédente peut donc être notée comme suit :

$$X^t \sim \mathfrak{C}\left(\eta_1^t \dots \eta_{m-1}^t, 1 - \|\boldsymbol{\eta}^t\|_1\right). \tag{2.1.7}$$

Soit $(\mathbf{e}_k^{m-1})$ la base orthonormale canonique de $\mathcal{N}$, et soit $\mathbf{0}^{m-1}$ son vecteur nul. Ces éléments forment l'ensemble $\mathcal{V}$ des sommets du simplexe $\mathcal{N}$.

$$\mathcal{V} = \left\{ \mathbf{e}_1^{m-1} \dots \mathbf{e}_{m-1}^{m-1},\, \mathbf{0}^{m-1} \right\} \tag{2.1.8}$$

On définit, à l'aide du symbole delta de Kronecker, une application canonique injective $V : \mathcal{X} \longrightarrow \mathcal{V}$. Puisque $\operatorname{card} \mathcal{X} = \operatorname{card} \mathcal{V} = m$, V est une bijection, en vertu du principe des tiroirs.

$$\forall x \in \mathcal{X},\ V(x) = \delta(x - m) \times \mathbf{0}^{m-1} + \sum_{k=1}^{m-1} \delta(x - k) \times \mathbf{e}_k^{m-1} \tag{2.1.9}$$

Bien que $\mathcal{V}$ ne soit pas inclus dans $\mathcal{M}$ — c'est en fait un sous-ensemble de sa frontière $\partial\mathcal{M}$ — les vecteurs $V(x)$ vérifient tous une condition de normalisation analogue à 2.1.3, ce qui permet de les voir aussi comme des distributions de probabilité catégoriques :

$$\forall x \in \mathcal{X},\ \sum_{k=1}^{m-1} \left(V(x)\right)_k = \left(V(x)\right)_x = 1. \tag{2.1.10}$$

hal-00849736, version 1 - 17 Sep 2013

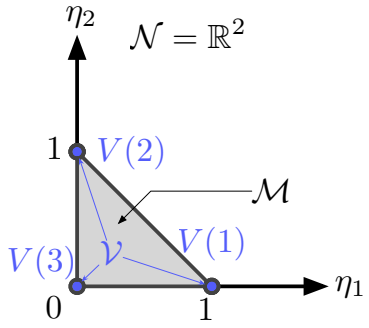


FIGURE 2.1.1: Visualisation du simplexe $\mathcal{M}$ pour $m = 3$.

Pour $x \in \mathcal{X}$, la loi $\mathfrak{C}\left((V(x))_1 \dots (V(x))_{m-1}, 1 - \|V(x)\|_1\right)$ est un cas limite de $\mathbb{P}_\lambda$ où toute la masse de probabilité est concentrée sur la catégorie x.

Sur la figure 2.1.1, on a représenté $\mathcal{M}$ et $\mathcal{V}$ dans le cas $m = 3$. L'espace $\mathcal{M}$ a alors la forme d'un triangle rectangle isocèle ouvert.

2.2 FAMILLE EXPONENTIELLE STANDARD

C'est à ce stade que nous voulons, à l'aide de V, exprimer $\mathbb{P}_{\lambda^t}$ comme un élément d'une famille exponentielle $\mathscr{P}$ de lois de probabilité sur $\mathcal{N}$. Pour ce faire, on tâche de définir un paramètre dit *naturel*, c'est-à-dire issu de $\boldsymbol{\eta}^t$ — et donc de $\boldsymbol{\lambda}^t$ — comme un logarithme naturel.

2.2.1 *Paramètre naturel*

Puisque les $\left(\eta_k^t\right)_{k \in [\![1; m-1]\!]}$ sont tous strictement positifs et que leur somme est strictement inférieure à 1, le vecteur

$$\boldsymbol{\theta}^t = \ln\left(\frac{\boldsymbol{\eta}^t}{1 - \|\boldsymbol{\eta}^t\|_1}\right) = \ln\left(\frac{\boldsymbol{\lambda}^t}{\lambda_m^t}\right) \tag{2.2.1}$$

est défini sur $\mathcal{M}$. En posant de même $\theta_m^t = \ln\left(\frac{\eta_m^t}{1 - \|\boldsymbol{\eta}^t\|_1}\right) = \ln 1 = 0$, on peut écrire l'identité suivante :

$$\forall x \in \mathcal{X},\ \theta_x^t = V(x)^\top \boldsymbol{\theta}^t. \tag{2.2.2}$$

Dans $\mathcal{M}$, et sous la transformation V, les composantes du modèle peuvent être extraites par un simple produit scalaire. Le paramètre naturel $\boldsymbol{\theta}^t \in \mathcal{N}$ peut donc être vu comme le noyau d'une forme linéaire sur $\mathcal{V}$. L'intérêt de cette approche est que cette forme linéaire s'étend canoniquement à l'adhérence $\overline{\mathcal{M}}$ de $\mathcal{M}$, comme on le verra à la section suivante.

hal-00849736, version 1 - 17 Sep 2013

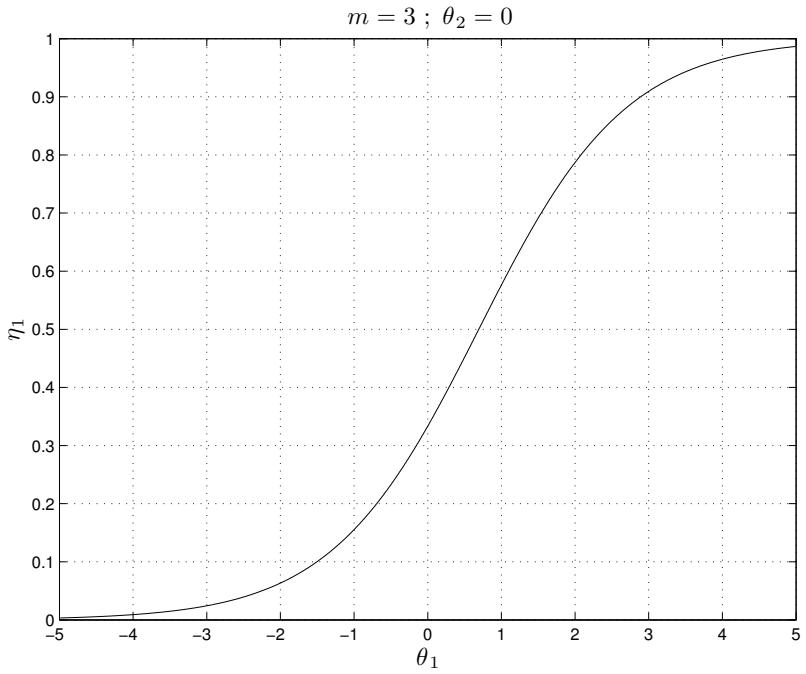


FIGURE 2.2.1: Sigmoïde partielle $\theta_1 \mapsto \eta_1$ (voir équation 2.2.3) pour $m = 3$, en fixant θ_2 à 0.

2.2.2 *Sigmoïdes*

Il s'agit d'écrire l'identité 2.2.2 en fonction de $\boldsymbol{\lambda}$. On commence par inverser la relation 2.2.1 :

$$\boldsymbol{\eta}^t = \frac{\exp \boldsymbol{\theta}^t}{1 + \| \exp \boldsymbol{\theta}^t \|_1}. \tag{2.2.3}$$

L'application $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}$ est un homéomorphisme de $\mathcal{M}$ vers $\mathcal{N}$. Sur chacune de ses composantes, sa courbe représentative prend la forme d'un S (voir figure 2.2.1), d'où sa qualification de *sigmoïde*.

2.2.3 *Fonction log-normalisatrice*

Le dénominateur de la fraction 2.2.3 est une fonction scalaire sur $\mathcal{M}$. Grâce à l'identité 2.2.2, il est possible de l'exprimer simplement à l'aide de produits scalaires :

$$f : \boldsymbol{\theta} \longmapsto \sum_{x \in \mathcal{X}} \exp \left(V(x)^\top \boldsymbol{\theta} \right) = \frac{1}{\lambda_m}. \tag{2.2.4}$$

Ainsi, la relation linéaire 2.2.2 devient une relation log-linéaire entre $\boldsymbol{\eta}^t$ et $\boldsymbol{\theta}^t$. On comprend pourquoi la famille $\mathscr{P}$ est qualifiée d'*exponentielle*, et la fonction f de *normalisatrice*.

$$\forall x \in \mathcal{X},\ \lambda_x^t = \frac{\exp \left(V(x)^\top \boldsymbol{\theta}^t \right)}{f\left(\boldsymbol{\theta}^t\right)} \tag{2.2.5}$$

hal-00849736, version 1 - 17 Sep 2013

Puisque f est strictement positive, il est fréquent de prendre son logarithme népérien F, afin d'être en cohérence avec les produits scalaires de la forme $V(x)^\top \boldsymbol{\theta}$. Selon les contextes, F est appelée *log-normalisatrice*, fonction de log-partition ou cumulant. Nous choisissons, de façon un peu arbitraire, la première de ces trois dénominations.

$$\forall \boldsymbol{\theta} \in \mathcal{N},\ F(\boldsymbol{\theta}) = \ln f(\boldsymbol{\theta}) = \ln(1 + \|\exp \boldsymbol{\theta}\|_1) \tag{2.2.6}$$

2.2.4 *Famille exponentielle standard*

En résumé, les transformations $\boldsymbol{\eta} \mapsto \boldsymbol{\theta}$ et V nous ont permis d'exprimer $\mathbb{P}_{\boldsymbol{\lambda}^t}$ comme la loi d'une famille exponentielle dite *standard*, c'est-à-dire sous la forme canonique suivante [1] :

$$\forall x \in \mathcal{X},\ \mathbb{P}_{\boldsymbol{\lambda}^t}\left[V\left(X^t\right) = V(x)\right] = \exp\left(V(x)^\top \boldsymbol{\theta}^t - F\left(\boldsymbol{\theta}^t\right)\right). \tag{2.2.7}$$

Afin d'étudier X^t, initialement défini sur un ensemble discret $\mathcal{X}$, on a d'abord construit un ensemble $\mathcal{V}$, canoniquement isomorphe à $\mathcal{X}$, d'éléments de l'espace vectoriel topologique $\mathcal{N} = \mathbb{R}^{m-1}$. Ensuite, on a plongé $\mathcal{V}$ dans une sous-variété différentielle $\mathcal{M}$ de $\mathcal{N}$, sur laquelle la loi $\mathbb{P}_{\boldsymbol{\lambda}^t}$ s'étend de façon continûment dérivable. Enfin, on a défini une bijection entre les coordonnées locales $\boldsymbol{\eta}$ de $\mathcal{M}$ et les formes linéaires $\boldsymbol{\theta}$ de l'espace $\mathcal{N}$. C'est à partir de l'expression 2.2.7, en l'appliquant à une réunion d'instants de la série temporelle X, que l'on va pouvoir construire un estimateur du paramètre source $\boldsymbol{\lambda}$ et détecter les changements brusques dans les observations sur le signal.

2.3 MAXIMUM DE VRAISEMBLANCE

2.3.1 *Vraisemblance d'une réalisation aléatoire*

Soit $n \in \mathbb{N}^*$; considérons une séquence $\left(X^1 \dots X^n\right)$ de variables aléatoires indépendantes, distribuées selon des lois respectives $\mathbb{P}_{\boldsymbol{\lambda}^1} \dots \mathbb{P}_{\boldsymbol{\lambda}^n}$. Etant donnée une trajectoire $\left(x^1 \dots x^n\right)$ de $\left(X^1 \dots X^n\right)$, la *vraisemblance* de $\left(x^1 \dots x^n\right)$ est la fonction $L_{\left(x^1 \dots x^n\right)} : \mathcal{S} \longrightarrow [0;1]$ associant à tout paramètre source $\boldsymbol{\lambda}$ la probabilité conditionnelle que la réalisation $\left(x^1 \dots x^n\right)$ soit obtenue à partir de n variables aléatoires i.i.d. de loi $\mathfrak{C}(\boldsymbol{\lambda})$.

$$L_{\left(x^1 \dots x^n\right)} : \boldsymbol{\lambda} \in \mathcal{S} \longmapsto \mathbb{P}_{\boldsymbol{\lambda}}\left[\left(X^1 \dots X^n\right) = \left(x^1 \dots x^n\right)\right] \tag{2.3.1}$$

1. Pour des raisons de lisibilité, nous ne rappelons pas la dépendance de $\boldsymbol{\theta}^t$ en $\boldsymbol{\lambda}^t$. Dans la suite, plus généralement, on se permettra de faire intervenir les vecteurs $\boldsymbol{\theta}$ et $\boldsymbol{\eta}$ sans les relier explicitement au paramètre source $\boldsymbol{\lambda}$. En effet, grâce aux équations bijectives 2.1.4, 2.2.1 et 2.2.3, on peut, selon la convenance, faire référence au même objet statistique en le voyant tantôt comme un point de $\mathcal{S}$, de $\mathcal{M}$ ou de $\mathcal{N}$.

hal-00849736, version 1 - 17 Sep 2013

Puisque les variables $X^1 \dots X^n$ sont indépendantes, toute loi jointe $\mathbb{P}_\lambda\left(X^1 \dots X^n\right)$ est égale au produit des lois marginales $\mathbb{P}_\lambda\left(X^1\right) \dots \mathbb{P}_\lambda\left(X^n\right)$.

$$\forall \boldsymbol{\lambda} \in \mathcal{S},\ L_{(x^1 \dots x^n)}(\boldsymbol{\lambda}) = \prod_{t=1}^{n} \mathbb{P}_\lambda\left[X^t = x^t\right] \tag{2.3.2}$$

Grâce à l'équation 2.2.7, ce produit devient une exponentielle de somme :

$$\forall \boldsymbol{\lambda} \in \mathcal{S},\ L_{(x^1 \dots x^n)}(\boldsymbol{\lambda}) = \exp\left(\sum_{t=1}^{n} V\left(x^t\right)^\top \boldsymbol{\theta} - n \times F(\boldsymbol{\theta})\right) \tag{2.3.3}$$

2.3.2 *Statistique exhaustive*

On définit

$$\mathbb{T} : \Xi = \left(\mathbf{v}^1 \dots \mathbf{v}^n\right) \in \mathcal{V}^n \longmapsto \frac{1}{n} \sum_{t=1}^{n} \mathbf{v}^t \tag{2.3.4}$$

sur le produit cartésien $\mathcal{V}^n$, prenant ses valeurs dans l'adhérence $\overline{\mathcal{M}}$ de l'ensemble $\mathcal{M}$. Afin d'alléger les notations, on appelle $\mathbb{T}(\Xi)$ la moyenne arithmétique de la suite finie Ξ quel que soit son nombre n d'échantillons. Dans le cas particulier $n = 1$, $\mathbb{T}$ est réduite à l'identité sur $\mathcal{V}$. La vraisemblance L_Ξ s'écrit sous une forme analogue à l'équation 2.2.7 :

$$\forall \boldsymbol{\lambda} \in \mathcal{S},\ L_\Xi(\boldsymbol{\lambda}) = \exp\left(n \times \left(\mathbb{T}(\Xi)^\top \boldsymbol{\theta} - F(\boldsymbol{\theta})\right)\right). \tag{2.3.5}$$

Dans l'équation précédente, la dépendance en Ξ est contenue exclusivement dans le vecteur $\mathbb{T}(\Xi)$. Autrement dit, afin de construire la vraisemblance d'une réalisation, il est suffisant d'enregistrer en mémoire son image par l'application $\mathbb{T}$. C'est pourquoi cette dernière porte le nom de *statistique exhaustive*. Au passage, on constate que $\mathbb{T}(\Xi)$ est invariant par permutation des éléments de Ξ, en raison de l'hypothèse d'indépendance des variables aléatoires X^t. C'est afin de respecter cette hypothèse que l'on a plaidé, à la section 1.3.3, pour une limitation de la durée du modèle Ξ à la durée d'un évènement musical.

2.3.3 *Extension de la vraisemblance*

En pratique, à tout instant t, les réalisations aléatoires de X^t nous sont cachées ; on les observe par l'intermédiaire de leur statistique exhaustive $\mathbb{T}\left(X^t\right)$. Plutôt que la vraisemblance de trajectoires particulières, on s'intéresse donc à celle de l'espérance mathématique $\mathbb{E}_\lambda\left[V\left(X^1\right) \dots V\left(X^n\right)\right]$ du processus $V(X)$. On a par définition (voir équation 2.1.9) :

$$\mathbb{E}_\lambda\left[V\left(X^t\right)\right] = \lambda_m^t \times \mathbf{0}^{m-1} + \sum_{x=1}^{m-1} \lambda_x^t \times \mathbf{e}_x^{m-1}. \tag{2.3.6}$$

hal-00849736, version 1 - 17 Sep 2013

Le terme $\lambda_m \times \mathbf{0}^{m-1}$ est nul et peut donc être supprimé. Par conséquent, l'expression précédente se réduit à

$$\mathbb{E}_{\boldsymbol{\lambda}}\left[V\left(X^t\right)\right] = \boldsymbol{\eta}^t, \tag{2.3.7}$$

ce qui explique pourquoi le vecteur $\boldsymbol{\eta}^t$ porte le nom de *paramètre des moments.* On pose :

$$\Xi = \left(\mathbb{E}_{\boldsymbol{\lambda}^1}\left[V\left(X^1\right)\right] \dots \mathbb{E}_{\boldsymbol{\lambda}^n}\left[V\left(X^n\right)\right]\right) = \left(\boldsymbol{\eta}^1 \dots \boldsymbol{\eta}^n\right) \in \overline{\mathcal{M}}^n. \tag{2.3.8}$$

On étend canoniquement la moyenne arithmétique $\mathbb{T}$ sur $\mathcal{V}^n$ à $\overline{\mathcal{M}}^n$, et l'on continue de noter $\mathbb{T}$ l'extension obtenue. La vraisemblance étendue L_Ξ associe alors à tout paramètre source $\boldsymbol{\lambda} \in \mathcal{S}$ la densité de probabilité de la moyenne empirique de n variables aléatoires indépendantes et identiquement distribuées de loi $\mathfrak{C}(\boldsymbol{\lambda})$ autour de la valeur $\mathbb{T}(\Xi)$.

2.3.4 *Estimation par maximum de vraisemblance*

On cherche à maximiser L_Ξ par rapport à $\boldsymbol{\theta}$, ce qui revient, par stricte positivité de la vraisemblance — assurée par l'équation 2.1.1, relative au blanchiment spectral — à maximiser son logarithme naturel $\ln L_\Xi$. Celui-ci est différentiable sur $\mathcal{N}$ en tant que fonction de $(m-1)$ variables. On calcule son gradient selon $\boldsymbol{\theta}$, c'est-à-dire le vecteur de ses dérivées partielles selon chacune des composantes θ_k :

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln L_\Xi(\boldsymbol{\lambda}) = n \times \left(\mathbb{T}(\Xi) - \boldsymbol{\nabla} F(\boldsymbol{\theta})\right). \tag{2.3.9}$$

Par stricte concavité de $\ln L_\Xi$ sur $\mathcal{N}$, il y a équivalence entre l'existence d'un maximum global $\tilde{\boldsymbol{\theta}}$ et l'annulation du gradient $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln L_\Xi$ en ce point, ce qui mène à la condition

$$\boldsymbol{\nabla} F\left(\tilde{\boldsymbol{\theta}}\right) = \mathbb{T}(\Xi). \tag{2.3.10}$$

Or, en appliquant la règle de dérivation en chaîne sur la log-normalisatrice F, et en faisant appel aux équations 2.2.3 et 2.2.5, on obtient :

$$\forall \boldsymbol{\theta} \in \mathcal{N},\ \boldsymbol{\nabla} F(\boldsymbol{\theta}) = \frac{\boldsymbol{\nabla} f(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} = \frac{\exp(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} = \boldsymbol{\eta}. \tag{2.3.11}$$

Autrement dit, l'application $\boldsymbol{\nabla} F$ a pour ensemble d'arrivée $\mathcal{M}$. Dès lors, on a existence d'un maximum de vraisemblance pour la séquence Ξ si et seulement si $\mathbb{T}(\Xi)$ appartient à l'ouvert $\mathcal{M}$, et pas à sa frontière. Par convexité de l'ouvert $\mathcal{M}$, une condition exhaustive pour être assuré de cette appartenance est d'avoir $\boldsymbol{\eta}^t \in \mathcal{M}$ à tout instant t. C'est pourquoi l'on a fait en sorte, à l'équation 2.1.1, que les mesures d'énergie spectrale $\boldsymbol{\lambda}^t$ aient des composantes toutes non nulles. Dans la suite, on supposera que la condition $\mathbb{T}(\Xi) \in \mathcal{M}$ est

hal-00849736, version 1 - 17 Sep 2013

systématiquement vérifiée. Ainsi, on est certain que la statistique exhaustive $\mathbb{T}$, initialement de dimension $m-1$, ne peut pas dégénérer en un vecteur de plus basse dimension.

En somme, afin de trouver le paramètre source $\tilde{\boldsymbol{\lambda}}$ maximisant la vraisemblance de la séquence Ξ, il suffit de poser

$$\tilde{\boldsymbol{\eta}} = \mathbb{T}(\Xi) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\eta}^t \tag{2.3.12}$$

et d'appliquer l'équation 2.1.4. L'équation 2.3.12 est bien connue en traitement du signal : elle rend légitime l'estimation d'une densité spectrale de puissance $\tilde{\boldsymbol{\eta}}$ en calculant une moyenne arithmétique de périodogrammes $\boldsymbol{\eta}^t$.

2.4 DÉTECTION SÉQUENTIELLE DE RUPTURE

S'agissant de la séquence $\left(X^1 \dots X^n\right)$ introduite à la section précédente, le problème de la détection de rupture consiste à déterminer, à partir de l'observation $\Xi_0 = \left(x^1 \dots x^n\right)$, si les variables aléatoires $X^1 \dots X^n$ sont identiquement distribuées ou si elles présentent une rupture temporelle.

2.4.1 *Définition*

En empruntant le vocabulaire des tests statistiques, l'hypothèse nulle H_0 signifie ici qu'en scindant Ξ_0 en deux sous-séquences $\Xi_{\mathrm{P}}^i = \left(x^1 \dots x^i\right)$ et $\Xi_{\mathrm{F}}^i = \left(x^{i+1} \dots x^n\right)$, celles-ci sont de même loi, et ce pour tout instant $i \in [\![1; n-1]\!]$. A l'inverse, l'hypothèse alternative H_1 signifie qu'il existe un instant i, appelé *point de changement*, tel que les sous-séquences Ξ_{P}^i et Ξ_{F}^i suivent des lois $\mathfrak{C}(\boldsymbol{\lambda}_{\mathrm{P}}^i)$ et $\mathfrak{C}(\boldsymbol{\lambda}_{\mathrm{F}}^i)$ différentes. Dans toute la suite, les indices P et F — pour « passé » et « futur » — correspondent à ces notations. Afin de confronter H_0 et H_1 et, le cas échéant, de déterminer le point de changement, on estime les paramètres $\tilde{\boldsymbol{\eta}}_{\mathrm{P}}^i$, $\tilde{\boldsymbol{\eta}}_{\mathrm{F}}^i$ et $\tilde{\boldsymbol{\eta}}_0$, maximisant respectivement la vraisemblance des observations $\mathbb{T}\left(\Xi_{\mathrm{P}}^i\right)$, $\mathbb{T}\left(\Xi_{\mathrm{F}}^i\right)$ et $\mathbb{T}(\Xi_0)$. Dans un second temps, on compare la somme des log-vraisemblances partielles à la log-vraisemblance de l'intégralité de la séquence. Soit $i \in [\![1; n-1]\!]$; afin d'obtenir $\tilde{\boldsymbol{\eta}}_{\mathrm{P}}^i$, $\tilde{\boldsymbol{\eta}}_{\mathrm{F}}^i$ et $\tilde{\boldsymbol{\eta}}_0$, il suffit d'écrire l'équation 2.3.12 pour les séquences Ξ_{P}^i, Ξ_{F}^i et Ξ_0 :

$$\tilde{\boldsymbol{\eta}}_{\mathrm{P}}^i = \frac{1}{i}\sum_{t=1}^{i} \boldsymbol{\eta}^t \quad ; \quad \tilde{\boldsymbol{\eta}}_{\mathrm{F}}^i = \frac{1}{n-i}\sum_{t=i+1}^{n} \boldsymbol{\eta}^t \quad ; \quad \tilde{\boldsymbol{\eta}}_0 = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\eta}^t. \tag{2.4.1}$$

2.4.2 *Rapport de vraisemblance généralisé*

Le *rapport de vraisemblance généralisé* (Lorden, 1971; Brandt, 1983) de Ξ à l'instant i est le réel $\Gamma^i(\Xi)$ défini par :

hal-00849736, version 1 - 17 Sep 2013

$$\tilde{\Gamma}^i(\Xi) = \frac{L_{\Xi_{\mathrm{P}}^i}\left(\tilde{\boldsymbol{\lambda}}_{\mathrm{P}}^i\right) \times L_{\Xi_{\mathrm{F}}^i}\left(\tilde{\boldsymbol{\lambda}}_{\mathrm{F}}^i\right)}{L_{\Xi_0}\left(\tilde{\boldsymbol{\lambda}}_0\right)}. \tag{2.4.2}$$

On parle de rapport *généralisé* car les paramètres sources $\tilde{\boldsymbol{\lambda}}_{\mathrm{P}}^i$, $\tilde{\boldsymbol{\lambda}}_{\mathrm{F}}^i$ et $\tilde{\boldsymbol{\lambda}}_0$ ne sont pas les vraies lois des séquences considérées mais des estimateurs empiriques. Plus l'hypothèse H_1^i d'un changement à l'instant i est plausible, plus la vraisemblance $L_{\Xi_0}\left(\tilde{\boldsymbol{\lambda}}_0\right)$ est petite devant le produit $L_{\Xi_{\mathrm{P}}^i}(\tilde{\boldsymbol{\lambda}}_{\mathrm{P}}^i) \times L_{\Xi_{\mathrm{F}}^i}(\tilde{\boldsymbol{\lambda}}_{\mathrm{F}}^i)$, et plus $\tilde{\Gamma}^i(\Xi)$ est grand. Il est commode de prendre le logarithme naturel du rapport de vraisemblance généralisé — souvent écrit GLR pour *Generalised Likelihood Ratio*.

$$\ln \tilde{\Gamma}^i(\Xi) = \ln L_{\Xi_{\mathrm{P}}^i}\left(\tilde{\boldsymbol{\lambda}}_{\mathrm{P}}^i\right) + \ln L_{\Xi_{\mathrm{F}}^i}\left(\tilde{\boldsymbol{\lambda}}_{\mathrm{F}}^i\right) - \ln L_{\Xi_0}\left(\tilde{\boldsymbol{\lambda}}_0\right) \tag{2.4.3}$$

2.4.3 *Transformée de Fenchel*

D'après l'équation 2.3.5, les log-vraisemblances maximales présentes dans le membre de droite sont proportionnelles aux images respectives de $\tilde{\boldsymbol{\eta}}_{\mathrm{P}}^i$, $\tilde{\boldsymbol{\eta}}_{\mathrm{F}}^i$ et $\tilde{\boldsymbol{\eta}}_0$ par la fonction $G = F^*$, définie sur $\mathcal{M}$ par

$$G : \boldsymbol{\eta} \in \mathcal{M} \longmapsto \sup_{\boldsymbol{\eta}^* \in \mathcal{N}} \left[\boldsymbol{\eta}^\top \boldsymbol{\eta}^* - F\left(\boldsymbol{\eta}^*\right)\right]. \tag{2.4.4}$$

G est appelée la *fonction conjuguée*, ou la *transformée de Fenchel*, de F. A l'aide des équations 2.2.1 et 2.2.6, il est possible d'en donner une formule explicite dans le cas particulier d'une distribution catégorique.

$$\forall \boldsymbol{\eta} \in \mathcal{M},\ G(\boldsymbol{\eta}) = \boldsymbol{\eta}^\top \ln\left(\frac{\boldsymbol{\eta}}{1 - \|\boldsymbol{\eta}\|_1}\right) - \ln\left(1 + \frac{\|\boldsymbol{\eta}\|_1}{1 - \|\boldsymbol{\eta}\|_1}\right) \tag{2.4.5}$$

L'expression ci-dessus se simplifie aisément, jusqu'à aboutir à

$$\forall \boldsymbol{\eta} \in \mathcal{M},\ G(\boldsymbol{\eta}) = \boldsymbol{\eta}^\top \ln \boldsymbol{\eta} + (1 - \|\boldsymbol{\eta}\|_1) \times \ln(1 - \|\boldsymbol{\eta}\|_1). \tag{2.4.6}$$

La fonction G possède alors des liens forts avec la théorie de l'information. On connaît l'expression classique de l'entropie de Shannon d'une loi discrète :

$$E(\boldsymbol{\lambda}) = -\sum_{x=1}^{m} \lambda_x \log_2 \lambda_x. \tag{2.4.7}$$

Après une conversion de logarithme, et en se souvenant que $\lambda_m = 1 - \|\boldsymbol{\eta}\|_1$ (voir équation 2.1.4), on aboutit tout simplement à

$$G(\boldsymbol{\eta}) = -E(\boldsymbol{\lambda}) \times \ln 2. \tag{2.4.8}$$

Dans le domaine des communications numériques, il est habituel de mesurer l'entropie de Shannon en *bits*. Puisque G n'est pas défini avec un logarithme en base 2 mais avec un logarithme naturel, on dit parfois qu'elle exprime une grandeur en *nats*.

hal-00849736, version 1 - 17 Sep 2013

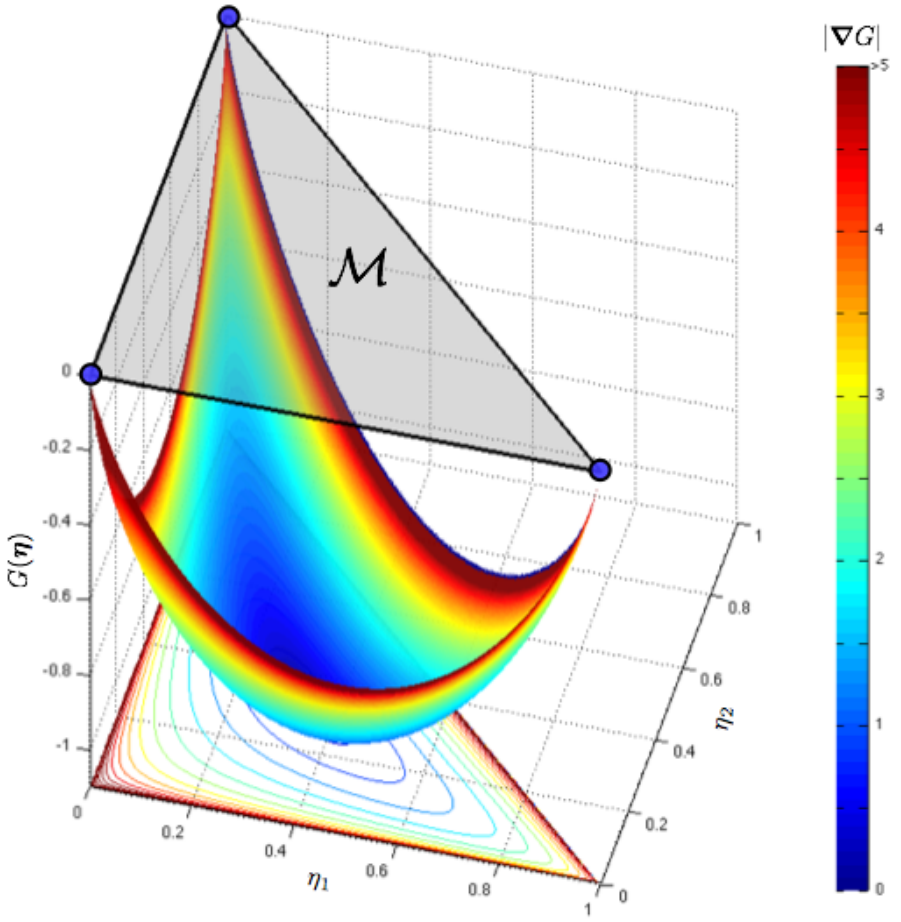


FIGURE 2.4.1: Visualisation de la fonction G en dimension $m = 3$. L'échelle de couleur est proportionnelle au module du gradient de G, borné par la valeur 5.

hal-00849736, version 1 - 17 Sep 2013

Sur la figure 2.4.1, on a représenté G comme une nappe paramétrée sur l'intérieur du triangle rectangle isocèle $\mathcal{M}$, dans le cas particulier $m = 3$.

En définitive, on peut calculer le logarithme de $\tilde{\Gamma}^i$ comme une combinaison linéaire d'images de G. Notre algorithme de détection séquentielle de rupture, emprunté à Dessein et Cont (2013), repose donc essentiellement sur la transformée de Fenchel.

$$\ln \tilde{\Gamma}^i (\Xi) = i \times G\left(\tilde{\boldsymbol{\eta}}_{\mathrm{P}}^i\right) + (n-i) \times G\left(\tilde{\boldsymbol{\eta}}_{\mathrm{F}}^i\right) - n \times G\left(\tilde{\boldsymbol{\eta}}_0\right) \tag{2.4.9}$$

2.4.4 *Règle de décision*

Soit γ un réel arbitraire strictement positif. La *règle de décision non bayésienne* associée à la détection de rupture est définie par

$$\max_{1 \leq i \leq n-1} \ln \tilde{\Gamma}^i (\Xi) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{2.4.10}$$

La segmentation de flux par détection séquentielle de rupture repose sur deux hypothèses fondamentales. La première stipule que les observations sont, de part et d'autre du point de changement, assez distinctes — et l'on verra à la section 3.1, avec l'introduction des divergences de Bregman, que cette distinguibilité possède un sens géométrique précis. Par chance, dans la plupart des signaux de musique, le début des évènements est marqué par une attaque brutale, qui se traduit par une hausse de l'énergie dans toutes les bandes de fréquence. Ce comportement transitoire est en rupture avec leur partie stationnaire, au cours de laquelle l'énergie est concentrée sur quelques partiels. Par ailleurs, dans le cadre statistique développé dans cette section, la détection séquentielle ne porte que sur un seul indice i. Cette limitation mène à une seconde hypothèse sur les observations : les points de changement doivent être suffisamment espacés dans le temps pour être détectés un par un. De toute évidence, ces deux hypothèses sont intrinsèquement liées : plus les changements seront abrupts, plus ils pourront être détectés rapidement, et mieux l'on sera capable de segmenter des évènements très brefs.

2.4.5 *Algorithme*

L'initialisation des variables, ainsi que la boucle synchrone de l'algorithme, sont décrites ci-après. L'entier p représente le premier point de rupture possible. À tout instant, les points de rupture examinés sont compris dans un tampon de mémoire vive (*buffer*) dont les indices vont de p à $t-1$. À chaque détection d'un nouveau segment, p est incrémenté de la longueur de ce segment ; les observations antérieures à p deviennent obsolètes, et peuvent être libérées de la mémoire vive. Il n'est pas difficile d'adapter cet algorithme afin de limiter la taille du modèle à une valeur maximale. Ainsi, l'on est assuré

hal-00849736, version 1 - 17 Sep 2013

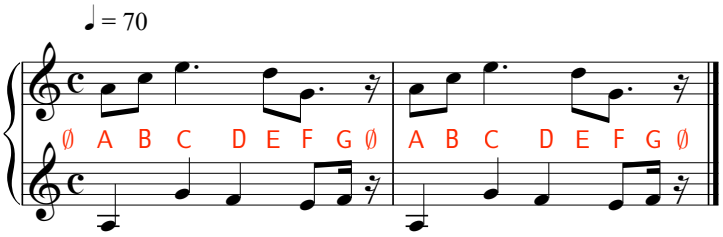


FIGURE 2.4.2: Transcription musicale du début de *Les Entretiens de la Belle et de la Bête*, issu de l'œuvre pour piano à quatre mains *Ma Mère l'Oye*, de Maurice Ravel. On a représenté les sept évènements sonores par les lettres de A à G. Le silence est représenté par le symbole ∅.

d'opérer en mémoire bornée. Par ailleurs, à chaque instant t, le temps de calcul étant environ proportionnel au nombre d'observations en mémoire, cette limitation permet aussi d'assurer que la détection de rupture est traitée plus vite que la cadence à laquelle arrivent les données. En pratique, les évènements musicaux font rarement plus d'une seconde, ce qui représente quelques centaines d'instructions à traiter en quelques millisecondes. Un ordinateur personnel n'a donc aucun mal à exécuter ce programme en temps réel.

Après que le dernier point de changement a été détecté dans le fichier, on produit un modèle du dernier évènement avec les observations restantes.

2.4.6 *Application à un ostinato de Ravel*

Afin de tester l'algorithme de détection séquentielle de rupture, on l'a appliqué à un enregistrement audio très simple : un ostinato de piano d'une mesure, répété deux fois. Il s'agit du début de *Les Entretiens de la Belle et de la Bête*, le quatrième mouvement de la célèbre suite *Ma Mère l'Oye*, composée par Maurice Ravel (1875–1937) autour de 1909. Nous avons transcrit cette ostinato à la figure 2.4.2.

L'enregistrement de piano considéré provient de la base de données RWC, pour *Real World Computing*, une association japonaise d'informatique. Cette base a été construite par Goto et al. (2002), et est maintenant devenue un standard parmi la communauté scientifique.

hal-00849736, version 1 - 17 Sep 2013

Détection séquentielle de rupture : initialisation

```
set λ % seuil de détection
t ← 0
p ← 1
η̃_0 ← 0
```

Détection séquentielle de rupture : boucle synchrone

```
t ← t + 1 % coup d'horloge
input λ^t
η^t ← λ^t_{1:m−1}
η̃_F^{t−1} ← η^t
for p ≤ j < t do
    η̃_F^j ← ((t − 1 − j) × η̃_F^j + η^t) / (t − j) % mise à jour barycentrique
    Γ̃^j ← 2 ( (j − p + 1) × φ(η̃_P^j) + (t − j) × φ(η̃_F^j) − (t − p + 1) × φ(η̃_0) ) %
        rapport de vraisemblance généralisé
end
i ← arg max_j Γ̃^j
if Γ̃^i > γ % règle de décision à l'instant i
    η̃_P^{t−1} ← ((t − p) × η̃_P − i × η̃_P^i) / (t − p − i)
    for i < j < (t − i) do
        η̃_P^j ← (j × η̃_P^j − i × η̃_P^i) / (j − i) % mise à jour barycentrique
    end
    η̃_0 ← ((t − p) × η̃_0 + η^t − i × η̃_P^i) / (t − p + 1 − i)
    p ← p + i % décalage temporel
    return η̃_P^i
else % pas de rupture détectée
    η̃_P^{t−1} ← η̃_0
    η̃_0 ← ((t − p) × η̃_0 + η^t) / (t − p + 1)
    return NULL
end
```

hal-00849736, version 1 - 17 Sep 2013

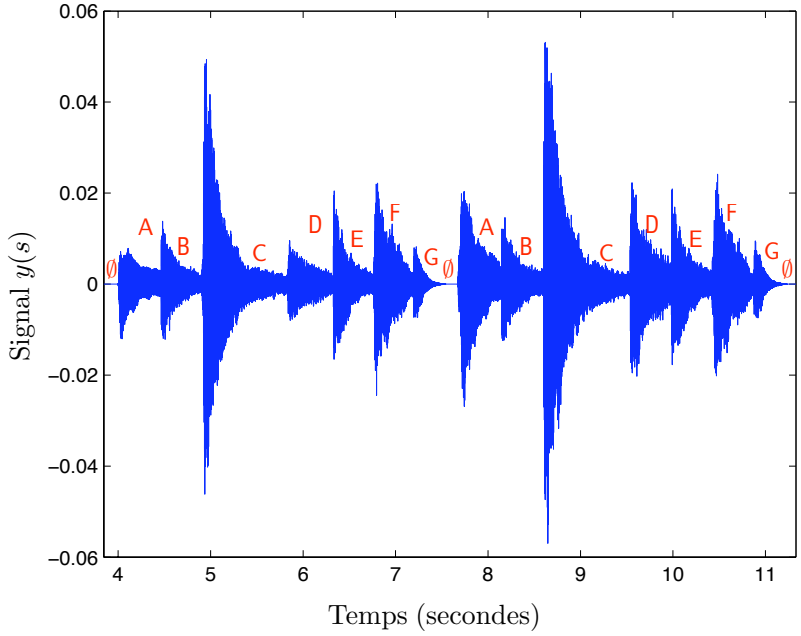


FIGURE 2.4.3: Extrait du fichier RWC-MDB-C-2001, n°23, Tr. 08. On a représenté les évènements avec la même convention qu'à la figure précédente.

Le fichier audio porte le nom suivant : RWC-MDB-C-2001, n°23, Tr. 08. Nous avons représenté l'évolution du signal monophonique $y(s)$, demi-somme des canaux gauche et droit, à la figure 2.4.3. On y voit clairement les attaques soudaines suivies d'une décroissance progressive de l'amplitude, phénomènes caractéristiques des instruments impulsifs tels que le piano. Les quatre premières secondes, correspondant à du silence, ne sont pas représentées.

La figure 2.4.4 décrit les résultats de segmentation audio en temps réel pour l'ostinato de Ravel. L'extrait 2.4.3 est représenté par le module de sa transformée de Fourier à court terme, avec une fenêtre d'analyse longue de $T = 23{,}2\,\text{ms}$, soit $m = 256$ bandes, et un pas d'analyse (*hop size*) quatre fois moindre. Le seuil de détection est fixé manuellement à $\gamma = 20$.

On constate que le délai de détection d'un point de changement, représenté par les intervalles entre barres rouges et barres bleues, et de l'ordre de 200 ms en moyenne. C'est beaucoup plus qu'un logiciel de suivi de partition tel qu'Antescofo, qui vise plutôt un temps de réaction de 15 ms ; mais satisfaisant dans un contexte « aveugle », c'est-à-dire sans information symbolique sur laquelle s'appuyer. Il existe une littérature scientifique abondante pour calculer, à partir d'un certain taux de fausse alarme, le seuil γ qui le respecte tout en minimisant le délai de détection moyen — voir notamment Poor et Hadjiliadis (2009) pour un aperçu récent de la question. En ce qui nous concerne, nous sommes avant tout vigilants à trouver un bon compromis entre

hal-00849736, version 1 - 17 Sep 2013

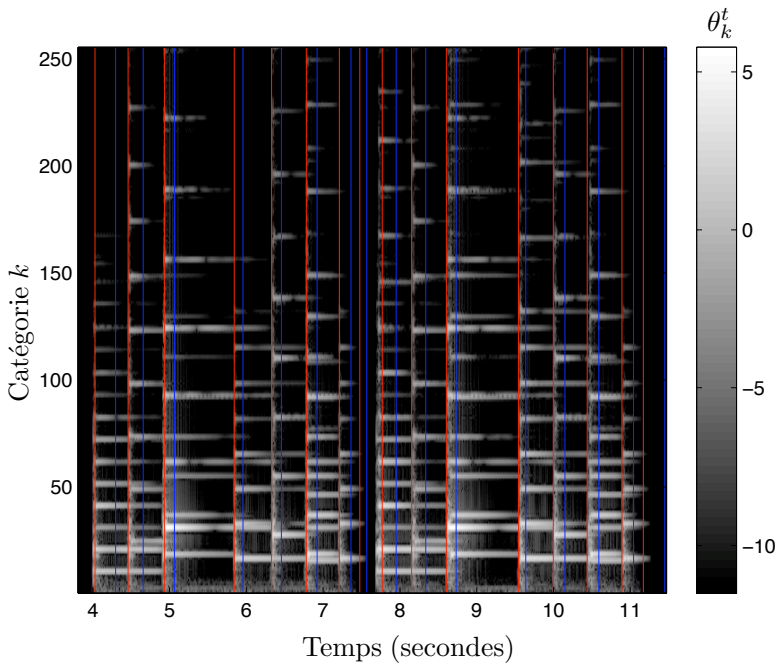


FIGURE 2.4.4: Résultat de segmentation pour l'ostinato de la figure 2.4.2. Les barres rouges sont les points de changement ; les barres bleues sont les instants où le changement est détecté.

taux de fausse alarme et taux de faux positifs. L'optimum serait de ne faire aucune erreur par rapport à la segmentation « vraie » ; mais l'existence de celui-ci n'est pas garantie, et correspond souvent à une plage très restreinte de seuils possibles. Dans l'exemple de l'ostinato de Ravel, les notes de piano y étant assez percussives, stationnaires et espacées, notre système parvient tout de même à détecter correctement les seize points de changement de l'extrait considéré, sans commettre aucune erreur. Cependant, dans l'optique d'une implémentation à grande échelle, il serait nécessaire d'améliorer la règle de décision 2.4.10 pour la rendre robuste à de nombreux types de signaux.

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

3

COMPARAISON DE BOULES INFORMATIONNELLES

Dans ce chapitre, on développe un procédé de comparaison géométrique entre des nuages de points. L'approche statistique initiée au chapitre 2 mène naturellement à une mesure de dissimilarité sur les observations, appelée divergence de Bregman. Les segments détectés étant, par construction, des entités homogènes, on choisit de les modéliser par des boules informationnelles, c'est-à-dire des voisinages convexes de leur observation exhaustive $\widetilde{\boldsymbol{\eta}} = \mathbb{T}(\Xi)$. Dès lors, plutôt que de mesurer deux à deux la dissimilarité entre les trames du flux temporel, on se propose de quantifier la relation entre les boules de Bregman associées. Faute d'inégalité triangulaire, il est nécessaire de faire appel à des outils d'optimisation convexe pour y définir des rapports d'inclusion et d'intersection. En raison de l'omniprésence des divergences de Bregman, ce travail se généralise à une large classe de problèmes en apprentissage statistique.

Dans notre dispositif de découverte de structures, l'étape de segmentation menée au chapitre précédent visait à distinguer géométriquement des régions adjacentes du signal. On peut dire qu'à l'inverse, l'étape de reconnaissance vise à trouver des ressemblances géométriques entre des régions temporellement éloignées. L'idée de la géométrie de l'information est de les aborder toutes deux avec un formalisme unique. Mais ce chapitre ne se limite pas à transférer les outils statistiques de la détection séquentielle de rupture à la comparaison d'observations deux à deux : grâce à deux résultats récents de géométrie computationnelle (Cayton, 2009), on propose une « métrique mixte » qui, nous l'espérons, se révèlera pertinente à l'échelle des évènements musicaux.

3.1 DIVERGENCES DE BREGMAN

3.1.1 *Dualité de type Legendre*

Étant données une statistique exhaustive $\mathbb{T}$ et une log-normalisatrice F, on rappelle l'expression de la log-vraisemblance d'un paramètre source $\boldsymbol{\lambda} \in \mathcal{S}$ au vu d'une séquence d'observations $\Xi \in \overline{\mathcal{M}}^n$ (voir section 2.3) :

$$\ln L_{\Xi}(\boldsymbol{\lambda}) = \mathbb{T}(\Xi)^{\top} \boldsymbol{\theta} - F(\boldsymbol{\theta}) \tag{3.1.1}$$

Le paramètre naturel de la famille exponentielle associée est noté $\boldsymbol{\theta}$. Outre la stricte convexité de la fonction conjuguée $G = F^*$ et le fait

que $\mathcal{M}$ soit non vide, le gradient de G tend vers $+\infty$ en tout point de $\partial\mathcal{M}$. Rockafellar (1970) a montré que ces trois conditions mènent à une dualité entre $(F, \mathcal{N})$ et $(G, \mathcal{M})$ dite *de type Legendre*. Plus forte que la dualité de Fenchel, qui portait essentiellement sur les fonctions F et G, la dualité de type Legendre instaure un homéomorphisme entre les espaces eux-mêmes, comme en atteste le diagramme ci-dessous.

$$\boldsymbol{\theta} \in \mathcal{N} \underset{\boldsymbol{\nabla} G}{\overset{\boldsymbol{\nabla} F}{\rightleftarrows}} \boldsymbol{\eta} \in \mathcal{M}$$

On a par conséquent, pour tout couple de coordonnées duales $(\boldsymbol{\theta}, \boldsymbol{\eta})$:

$$F(\boldsymbol{\theta}) + G(\boldsymbol{\eta}) = \boldsymbol{\theta}^\top \boldsymbol{\eta} \tag{3.1.2}$$

3.1.2 *Définition*

En exploitant la relation précédente et en remplaçant $\boldsymbol{\theta}$ par $\boldsymbol{\nabla} G(\boldsymbol{\eta})$, on réécrit la log-vraisemblance $\ln L_\Xi(\boldsymbol{\lambda})$ en fonction du paramètre des moments $\boldsymbol{\eta}$:

$$\ln L_\Xi(\boldsymbol{\lambda}) = G(\boldsymbol{\eta}) + (\mathbb{T}(\Xi) - \boldsymbol{\eta})^\top \boldsymbol{\nabla} G(\boldsymbol{\eta}) \tag{3.1.3}$$

Si l'on fixe $\boldsymbol{\eta}$ et que l'on fait varier la statistique exhaustive $\mathbb{T}(\Xi)$, $\ln L_\Xi(\boldsymbol{\lambda})$ peut être vue comme la forme affine sur $\mathcal{M}$ passant par $G(\boldsymbol{\eta})$ et de vecteur directeur non nul $\boldsymbol{\nabla} G(\boldsymbol{\eta})$. En termes géométriques, il s'agit de l'ordonnée en $\mathbb{T}(\Xi)$ de l'hyperplan tangent à la fonction G en $\boldsymbol{\eta}$. Par stricte convexité, on a

$$G(\mathbb{T}(\Xi)) \geq G(\boldsymbol{\eta}) + (\mathbb{T}(\Xi) - \boldsymbol{\eta})^\top \boldsymbol{\nabla} G(\boldsymbol{\eta}) = \ln L_\Xi(\boldsymbol{\lambda}) \tag{3.1.4}$$

avec égalité si et seulement si $\boldsymbol{\eta} = \mathbb{T}(\Xi)$. On a vu à la section 2.4 que $G(\mathbb{T}(\Xi))$ est égal au maximum de log-vraisemblance de la séquence d'observations Ξ. Dès lors, afin de quantifier la dissimilarité entre la statistique exhaustive $\mathbb{T}(\Xi)$ et un modèle quelconque $\boldsymbol{\lambda}$, il est légitime de considérer l'écart entre ce maximum et la valeur $\ln L_\Xi(\boldsymbol{\lambda})$. Cet écart peut être vu comme l'image par $(\mathbb{T}(\Xi), \boldsymbol{\eta})$ de la fonction réelle

$$d_G : \boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathcal{M} \longmapsto G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}') - (\boldsymbol{\eta} - \boldsymbol{\eta}')^\top \boldsymbol{\nabla} G(\boldsymbol{\eta}') \tag{3.1.5}$$

définie sur le produit cartésien $\mathcal{M} \times \mathcal{M}$. d_G est appelée la *divergence de Bregman* associée à G (Bregman, 1967). A l'instar d'une distance, une divergence est une quantité positive et définie — elle s'annule si et seulement si ses arguments sont égaux. Par contre, elle n'est évidemment pas symétrique — on a $d_G(\boldsymbol{\eta} \| \boldsymbol{\eta}') \neq d_G(\boldsymbol{\eta}' \| \boldsymbol{\eta})$ en général — et ne vérifie pas l'inégalité triangulaire, ce qui requiert des algorithmes pointus pour la comparaison de boules (voir à ce sujet les sections 3.2 et 3.3). Le problème de la maximisation de la log-vraisemblance se

hal-00849736, version 1 - 17 Sep 2013

ramène à une minimisation de la divergence de Bregman à droite de $\mathbb{T}(\Xi)$:

$$\ln L_{\Xi}(\boldsymbol{\lambda}) = G(\mathbb{T}(\Xi)) - d_G(\mathbb{T}(\Xi) \| \boldsymbol{\eta}) \tag{3.1.6}$$

3.1.3 *Exemples*

La plupart des dissimilarités utilisées par la communauté de traitement du signal et d'apprentissage statistique sont en fait des divergences de Bregman. Pour $G_{\text{Euc}} : \boldsymbol{\eta} \mapsto \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2$, la divergence associée

$$d_{\text{Euc}}(\boldsymbol{\eta} \| \boldsymbol{\eta}') = \frac{1}{2}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2 \tag{3.1.7}$$

est proportionnelle au carré de la distance euclidienne entre les paramètres sources. La fonction $\boldsymbol{\nabla} G$ est réduite à l'identité, ce qui rend d_{Euc} symétrique. Un autre exemple de divergence de Bregman fréquemment rencontrée est la divergence d'Itakura-Saito

$$d_{\text{IS}}(\boldsymbol{\eta} \| \boldsymbol{\eta}') = \left\| \frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda}'} - \ln \frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda}'} - 1 \right\|_1, \tag{3.1.8}$$

relative à $G : \boldsymbol{\eta} \mapsto -\|\log \boldsymbol{\lambda}\|_1$, et qui présente la particularité d'être invariante par homothétie sur les paramètres sources. Ainsi, avec la divergence d'Itakura-Saito, il est facile de définir des *seuils différentiels* de perception, une notion importante en physiologie et en psychoacoustique.

Pour revenir au cas d'une distribution catégorique $\mathfrak{C}(\boldsymbol{\lambda})$, la divergence de Bregman associée s'appelle *fonction de perte logistique* :

$$d_G : \boldsymbol{\eta}, \boldsymbol{\eta}' \longmapsto \left\| \boldsymbol{\eta} \times \ln \frac{\boldsymbol{\eta}}{\boldsymbol{\eta}'} \right\|_1 + (1 - \|\boldsymbol{\eta}\|_1) \times \ln \left\| \frac{\boldsymbol{\eta}}{\boldsymbol{\eta}'} \right\|_1 \tag{3.1.9}$$

Exprimée en fonction des paramètres sources, d_G possède une forme plus simple. On reconnaît alors la *divergence de Kullback-Leibler*.

$$d_G(\boldsymbol{\eta} \| \boldsymbol{\eta}') = d_{\text{KL}}(\boldsymbol{\lambda} \| \boldsymbol{\lambda}') = \left\| \boldsymbol{\lambda} \times \ln \frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda}'} \right\|_1 \tag{3.1.10}$$

3.1.4 *Commentaires sur la divergence de Kullback-Leibler*

Très populaire parmi la communauté MIR, la divergence de Kullback-Leibler (KL) est souvent présentée comme « le bon outil pour comparer des distributions de probabilité », sans aucune autre forme de procès. C'est l'argument de relation canonique entre la famille exponentielle correspondant à la distribution catégorique et la divergence d_G qui est invoqué, de façon plus ou moins explicite, pour en justifier l'intérêt. Nous validons cet argument dans le cas où la distribution est modélisée en pratique par un histogramme normalisé et strictement positif. Mais si l'on estime $\boldsymbol{\lambda}$ de façon fréquentiste, c'est-à-dire

hal-00849736, version 1 - 17 Sep 2013

en mesurant le nombre d'occurrences de chaque catégorie, il est possible que certains de ses coefficients soient nuls, *a fortiori* lorsque le nombre d'expériences indépendantes répétées est modeste. Pour tout autre paramètre source λ', la divergence $d_{\mathrm{KL}}(\lambda \| \lambda')$ est infinie, sauf dans le cas où λ' possède exactement les mêmes coefficients nuls. Par conséquent, en raison de sa « sévérité » aux abords de zéro, le choix de la divergence KL pour comparer des vecteurs creux risque de se solder par un échec : avec une dissimilarité infinie entre de nombreuses paires de vecteurs, la pertinence d'un regroupement hiérarchique devient compromise. En revanche, puisque les histogrammes que nous considérons procèdent de spectres de Fourier mesurés physiquement, il est raisonnable de penser qu'ils sont partout non nuls. En effet, le bruit thermique du microphone suffit à donner à $\hat{y}(\omega)$ un minimum strictement positif, de l'ordre de 10^{-8} en pratique. Ce bruit thermique joue tout à fait le rôle du nombre ϵ requis à l'équation 2.1.1.

Le second risque à considérer dans l'utilisation d'une divergence telle que Kullback-Leibler réside dans son absence de symétrie : puisque la dissimilarité entre λ et λ' n'est pas égale à celle entre λ' et λ, on craint que la notion de similarité, celle qui nous intéresse en fin de compte, soit ambiguë. On peut rétorquer que d'un point de vue perceptif, la comparaison entre un objet de référence et un objet examiné est intrinsèquement *asymétrique*[1]. Afin de retrouver quelles sont les répétitions exactes dans une matrice de dissimilarité KL, il est donc indifférent d'effectuer un seuillage sur la partie triangulaire supérieure ou inférieure. En pratique, les seuls cas d'asymétrie concernent les pauses, c'est-à-dire les segments où aucune note n'est jouée ; ceux-ci ont alors un spectre normalisé très plat, composé de valeurs proches de $\frac{1}{m}$. Dans le but d'accroître la robustesse de la détection de structures, nous suggérons donc, en amont de la segmentation entre évènements polyphoniques, de réaliser une pré-segmentation entre zones de silence et zones d'activité. Outre des techniques traditionnelles issues du traitement de la parole (Lynch et al., 1987), il est possible d'y réemployer la géométrie de l'information. Dessein (2012, sous-section 4.3.1) a montré qu'en modélisant l'énergie à court terme par une distribution de Rayleigh, il est possible de détecter en temps réel les ruptures entre activité et silence.

Une troisième objection à l'utilisation de la divergence de Kullback-Leibler pour représenter la dissimilarité entre sons musicaux réside dans le découpage du spectre sonore en catégories discrètes, lorsque celui-ci présente une bande passante uniforme. Comme on l'a évoqué plus haut, il est critique que les partiels de deux sons jugés similaires soient exactement identiques en fréquence, c'est-à-dire centrés sur la même catégorie. Si cette hypothèse est plausible pour les instruments à clavier, elle est complètement mise à mal pour les instruments de

1. voir Bowdle et Gentner (1997) pour des expériences de psychologie humaine sur la question.

hal-00849736, version 1 - 17 Sep 2013

l'orchestre symphonique tels que le violon ou le hautbois : une très légère modification du jeu du musicien pouvant conduire à de drastiques décalages des partiels, surtout dans les hautes fréquences, il nous semble illusoire d'espérer classifier des sons musicaux en se basant sur le module de leur transformée de Fourier. Pour pallier ce problème, nous proposerons, à la section 4.2, de remplacer cette représentation sonore par un banc de filtres dit *à facteur de qualité constant*, c'est-à-dire dont la bande passante est proportionnelle à la fréquence centrale de chaque filtre ; d'où, on le verra, une propriété de stabilité aux petites déformations temporelles. Ce choix permet, tout en gardant le même cadre statistique, d'améliorer considérablement la performance de la structuration.

3.1.5 *Familles exponentielles régulières*

A travers l'équation 3.1.6, on revisite le cadre des statistiques traditionnelles en l'abordant comme un problème géométrique. Cette démarche, détaillée par Amari et Nagaoka (2000), ne se limite pas à la comparaison d'histogrammes : pour toute famille exponentielle de paramètre naturel $\boldsymbol{\theta}$ et de log-normalisatrice F, on peut construire, sur le modèle de l'équation 3.1.5, deux divergences de Bregman d_F et d_G, avec $G = F^*$. En exploitant le diagramme de dualité de Legendre et la relation 3.1.2, il est aisé de montrer que ces deux divergences sont égales, à condition de permuter l'ordre des arguments. De plus, ces divergences de Bregman sur les paramètres reviennent à calculer une divergence KL entre les densités de probabilité elles-mêmes.

$$\forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathcal{S},\ d_F\left(\boldsymbol{\theta} \| \boldsymbol{\theta}'\right) = d_G\left(\boldsymbol{\eta}' \| \boldsymbol{\eta}\right) = d_{\mathrm{KL}}\left(\mathbb{P}_{\lambda'} \| \mathbb{P}_{\lambda}\right) \tag{3.1.11}$$

À l'inverse, on peut chercher à caractériser les divergences de Bregman régulières, soit celles auxquelles on peut associer une famille exponentielle. Banerjee et al. (2005) ont montré que pour toute divergence de Bregman d_F, une condition nécessaire et suffisante de régularité est que la normalisatrice $f = \exp F$ soit exponentiellement convexe [2] et que l'espace de paramètres naturels $\mathcal{M}$ soit ouvert. On bénéficie par conséquent d'une bijection entre familles exponentielles régulières et divergences de Bregman.

2. On définit, pour $n \in \mathbb{N}^*$, l'ensemble $\mathcal{K}_n = \left\{\left(\boldsymbol{\eta}^1 \dots \boldsymbol{\eta}^n\right) \in \mathcal{M}^n | \forall i, j \in [\![1; n]\!],\ \boldsymbol{\eta}^i + \boldsymbol{\eta}^j \in \mathcal{M}\right\}$. Une fonction f de $\mathcal{M}$ dans $\mathbb{R}_+^*$ est dite *exponentiellement convexe* si et seulement si le noyau $K_f : (\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{K}_2 \longmapsto f\left(\boldsymbol{\eta} + \boldsymbol{\eta}'\right)$ vérifie, pour tout $\mathbf{z} \in \mathbb{C}^n$ et pour tout n-uplet $\left(\boldsymbol{\eta}^1 \dots \boldsymbol{\eta}^n\right) \in \mathcal{K}_n$, l'inégalité $\mathbf{z}^\top K_f\left(\boldsymbol{\eta}^i, \boldsymbol{\eta}^j\right) \overline{\mathbf{z}} \geq 0$ pour tous i, $j \in [\![1; n]\!]$.

hal-00849736, version 1 - 17 Sep 2013

3.1.6 *Application à l'ostinato de Ravel*

Au cours de l'expérience de segmentation sur l'ostinato (voir sous-section 2.4.6), nous avons, pour chaque segment $\Xi = \left(\tilde{\lambda}^1 \dots \tilde{\lambda}^n\right)$ de longueur n, gardé en mémoire la statistique exhaustive

$$\widetilde{\eta} = \mathbb{T}(\Xi) = \frac{1}{n}\sum_{t=1}^{n} \eta^t, \tag{3.1.12}$$

notée $\tilde{\eta}_0$ au chapitre précédent. En raison de l'équation précédente, ce vecteur est appelé *centre de masse* du modèle.

La figure 3.1.1 est une matrice de dissimilarité construite avec les divergences de Bregman — c'est-à-dire, dans le cas particulier d'un flux d'histogrammes, une divergence KL — entre les centres de masses respectifs de chaque modèle. Les cellules composées des états 1 à 7 ressortent sous forme de sous-diagonale, tandis que la similarité entre états 4 et 7 forme un point sombre isolé.

En guise de comparaison, nous avons représenté, à la figure 3.1.2, la même matrice de dissimilarité KL, mais sur les observations elles-mêmes. Bien que les cellules structurelles y ressortent aussi, les variations transitoires des observations en compromettent la lisibilité. Par ailleurs, il faut remarquer que la matrice des modèles occupe 400 octets en mémoire environ, tandis que la matrice des observations occupe 4 mégaoctets. En effet, on rappelle que le pas d'analyse pour une observation est de 5 ms, tandis que la durée d'un évènement est de l'ordre de 500 ms. L'utilisation d'une segmentation appropriée montre que l'information contenue dans une SDM traditionnelle peut être compressée d'un facteur 10^4 sans en perdre les qualités.

hal-00849736, version 1 - 17 Sep 2013

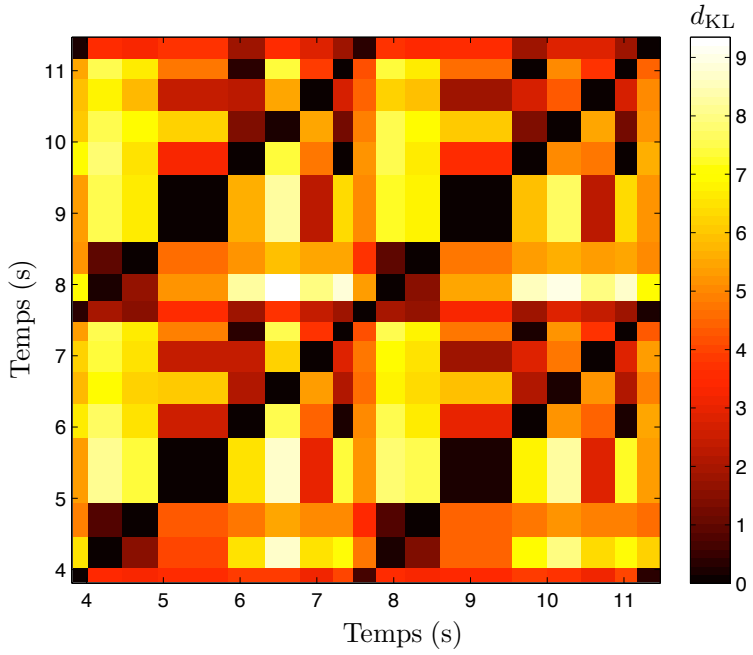


FIGURE 3.1.1: Matrice de dissimilarité KL entre centres de masses respectifs des modèles de l'ostinato.

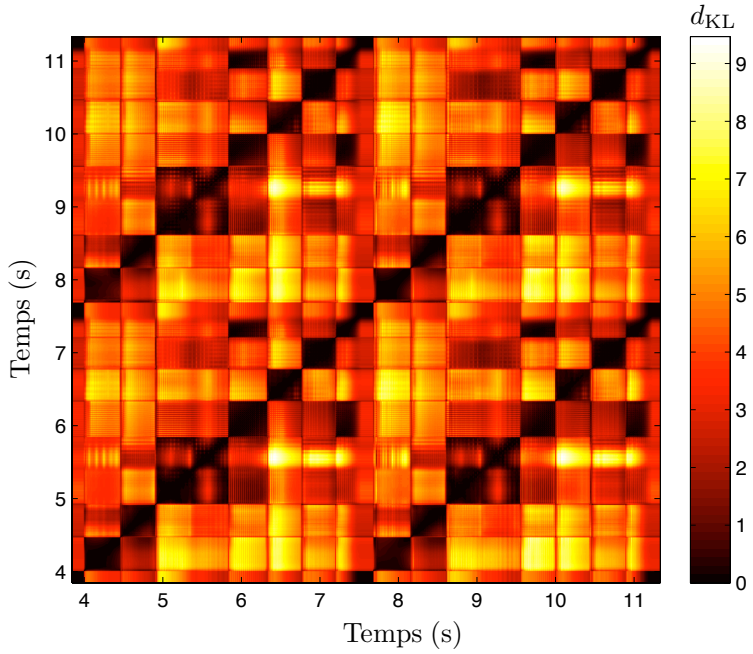


FIGURE 3.1.2: Matrice de dissimilarité KL entre les observations de l'ostinato.

hal-00849736, version 1 - 17 Sep 2013

3.1.7 *Boules de Bregman*

Afin d'améliorer la métrique entre centres de masses, on cherche à représenter chaque modèle comme un objet géométrique à part entière. Celui qui s'est imposé à nous, à la fois par sa simplicité et par sa popularité est la *boule de Bregman*, qui généralise la boule euclidienne à toutes sortes de géométrie, y compris la géométrie KL.

Dans l'espace $\mathcal{M}$ des paramètres des moments, une *boule de Bregman à droite* est la donnée d'un *centroïde* $\mathbf{c} \in \mathcal{M}$ et d'un rayon $R \geq 0$. Cette notion généralise la notion de boule euclidienne aux divergences de Bregman.

$$\mathcal{B} = \{\boldsymbol{\eta} \in \mathcal{M} \mid d_G(\boldsymbol{\eta} \| \mathbf{c}) \leq R\} \tag{3.1.13}$$

Or, Banerjee et al. (2005) ont démontré que le centre de masse minimise la divergence à droite moyenne avec tous les éléments du modèle, et ce, indépendamment de la divergence de Bregman choisie. Puisqu'il a déjà été calculé lors de la segmentation, on choisit le centre de masse, au moins dans un premier temps, comme centroïde pour chaque boule de Bregman.

Le rayon de la boule est alors défini comme la moyenne arithmétique des divergences entre le centroïde et les points du modèle. C'est à partir de ces « boules informationnelles » (Cont et al., 2010) que l'on va pouvoir mettre en œuvre des algorithmes de comparaison de modèles. Nous reportons notre lecteur à Nielsen et Nock (2009) pour un aperçu des différents choix de centroïdes possibles.

3.2 TEST D'INCLUSION

3.2.1 *Problème primal*

Soient $\mathcal{B}_0 = (\widetilde{\boldsymbol{\eta}_0}, R_0)$ et $\mathcal{B}_1 = (\widetilde{\boldsymbol{\eta}_1}, R_1)$ deux boules de Bregman appartenant à l'espace $\mathcal{M}$. On cherche à savoir si $\mathcal{B}_0$ est incluse dans $\mathcal{B}_1$. L'objectif de cette section est de résoudre le problème d'optimisation sous contrainte suivant :

$$\begin{aligned} \hat{\boldsymbol{\eta}} &= \underset{\boldsymbol{\eta} \in \mathcal{M}}{\operatorname{argmax}}\, d_G(\boldsymbol{\eta} \| \widetilde{\boldsymbol{\eta}_1}) \\ &\text{tel que } d_G(\boldsymbol{\eta} \| \widetilde{\boldsymbol{\eta}_0}) \leq R_0 \end{aligned} \tag{3.2.1}$$

En effet, on a alors équivalence entre la propriété géométrique $\mathcal{B}_0 \subset \mathcal{B}_1$ et l'inégalité $d_G(\hat{\boldsymbol{\eta}} \| \widetilde{\boldsymbol{\eta}_1}) \leq R_1$. Bien qu'il s'agisse d'un problème de maximisation non concave, il est possible d'en effectuer une relaxation de contrainte. Le lagrangien associé vaut :

$$\nu : (\boldsymbol{\eta}, \alpha) \in \mathcal{M} \times \mathbb{R}_+ \longmapsto d_G(\boldsymbol{\eta} \| \widetilde{\boldsymbol{\eta}_1}) + \alpha \times (R_0 - d_G(\boldsymbol{\eta} \| \widetilde{\boldsymbol{\eta}_0}))\,. \tag{3.2.2}$$

Ainsi, le problème primal 3.2.1 s'écrit alternativement :

$$(\hat{\boldsymbol{\eta}}, \hat{\alpha}) = \arg \max_{\boldsymbol{\eta} \in \mathcal{M}} \min_{\alpha \geq 0} \nu(\boldsymbol{\eta}, \alpha) \tag{3.2.3}$$

hal-00849736, version 1 - 17 Sep 2013

3.2.2 *Condition nécessaire d'optimalité*

Le problème dual procède d'une permutation des opérateurs max et min dans la définition précédente. A l'issue de cette permutation, on va tâcher d'exprimer $\overline{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}\in\mathcal{M}} \nu(\boldsymbol{\eta}, \alpha)$ pour tout $\alpha \geq 0$. Dans la suite, on ne rappellera pas la dépendance de $\overline{\boldsymbol{\eta}}$ en α.

$$(\hat{\boldsymbol{\eta}}, \hat{\alpha}) = \arg\min_{\alpha\geq 0}\max_{\boldsymbol{\eta}\in\mathcal{M}} \nu(\boldsymbol{\eta}, \alpha) = \arg\min_{\alpha\geq 0} \nu(\overline{\boldsymbol{\eta}}, \alpha) \tag{3.2.4}$$

En tout optimum local $(\overline{\boldsymbol{\eta}}, \alpha)$ de ν, il est nécessaire que le gradient de ν selon $\boldsymbol{\eta}$ soit nul. Or on a :

$$\forall \boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathcal{M},\ \boldsymbol{\nabla}_{\boldsymbol{\eta}} d_G(\boldsymbol{\eta} \| \boldsymbol{\eta}') = \boldsymbol{\nabla} G(\boldsymbol{\eta}) - \boldsymbol{\nabla} G(\boldsymbol{\eta}') = \boldsymbol{\theta} - \boldsymbol{\theta}'. \tag{3.2.5}$$

Dès lors, en posant successivement $\overline{\boldsymbol{\theta}} = \boldsymbol{\nabla} G(\overline{\boldsymbol{\eta}})$, $\widetilde{\boldsymbol{\theta}_0} = \boldsymbol{\nabla} G(\widetilde{\boldsymbol{\eta}_0})$ et $\widetilde{\boldsymbol{\theta}_1} = \boldsymbol{\nabla} G(\widetilde{\boldsymbol{\eta}_1})$, la condition nécessaire $\boldsymbol{\nabla}_{\boldsymbol{\eta}} \nu(\overline{\boldsymbol{\eta}}, \alpha) = 0$ se traduit par l'équation

$$\overline{\boldsymbol{\theta}} = \frac{1}{1-\alpha}\left(\widetilde{\boldsymbol{\theta}_1} - \alpha\widetilde{\boldsymbol{\theta}_0}\right) \tag{3.2.6}$$

, à supposer que $\alpha \neq 1$.

L'égalité ci-dessus définit implicitement $\overline{\boldsymbol{\eta}}$ à partir du réel positif α ; c'est sur ce dernier que l'on se concentre à présent. Puisque l'on cherche à maximiser ν selon $\boldsymbol{\eta}$, une autre condition nécessaire est le fait que ses dérivées partielles secondes forment une matrice carrée définie négative en $(\overline{\boldsymbol{\eta}}, \alpha)$. On obtient la propriété :

$$\boldsymbol{\nabla}^2 \nu(\overline{\boldsymbol{\eta}}, \alpha) = (1-\alpha)\,\boldsymbol{\nabla}^2 G(\overline{\boldsymbol{\eta}}) < 0 \tag{3.2.7}$$

3.2.3 *Géodésique*

Par stricte convexité de G sur $\mathcal{M}$, la matrice $\boldsymbol{\nabla}^2 G$, appelée *hessienne* de G, est définie positive en tout point de $\mathcal{M}$. Dès lors, la condition $\boldsymbol{\nabla}^2 \nu(\overline{\boldsymbol{\eta}}, \alpha) < 0$ équivaut à l'inégalité réelle stricte $\alpha > 1$. On pose

$$\rho = \frac{\alpha}{\alpha - 1} \tag{3.2.8}$$

, ce qui justifie l'équation 3.2.6 et lui donne la forme

$$\overline{\boldsymbol{\theta}} = \rho\widetilde{\boldsymbol{\theta}_1} + (1-\rho)\,\widetilde{\boldsymbol{\theta}_0}. \tag{3.2.9}$$

Le vecteur $\overline{\boldsymbol{\theta}}$ appartient à la demi-droite ouverte $\mathcal{D} \subset \mathcal{N}$ d'origine $\widetilde{\boldsymbol{\theta}_1}$ et engendrée par le vecteur $\left(\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}\right)$. Dans l'espace $\mathcal{M}$, $\overline{\boldsymbol{\eta}}$ est sur la demi-*géodésique* correspondante

$$\mathcal{D}^* = \left\{\boldsymbol{\eta}_\rho = \boldsymbol{\nabla} F(\rho\boldsymbol{\nabla} G(\widetilde{\boldsymbol{\eta}_1}) + (1-\rho)\,\boldsymbol{\nabla} G(\widetilde{\boldsymbol{\eta}_0})) \mid \rho > 1\right\}. \tag{3.2.10}$$

Une géodésique généralise la notion de ligne droite à des espaces non euclidiens, en traçant le plus court chemin entre deux points sur une variété différentielle donnée.

Par dualité faible, on a l'inégalité suivante :

$$\forall \rho > 1,\ \nu(\hat{\boldsymbol{\eta}}, \hat{\alpha}) \leq \nu(\overline{\boldsymbol{\eta}}, \alpha) \tag{3.2.11}$$

hal-00849736, version 1 - 17 Sep 2013

3.2.4 *Résolution*

A présent, on cherche un point $\check{\boldsymbol{\eta}}$ de $\mathcal{D}^*$ minimisant le réel $\check{\alpha}$ associé. La dérivée partielle de $d_G\left(\boldsymbol{\eta}\|\widetilde{\boldsymbol{\eta}_1}\right)$ par rapport à ρ s'exprime grâce à la règle de dérivation en chaîne, en passant par les variables $\boldsymbol{\eta}$ et $\boldsymbol{\theta}$.

$$\frac{\partial d_G}{\partial \rho}\left(\boldsymbol{\eta}\|\widetilde{\boldsymbol{\eta}_0}\right) = \boldsymbol{\nabla}_{\boldsymbol{\eta}} d_G\left(\boldsymbol{\eta}\|\widetilde{\boldsymbol{\eta}_0}\right)^\top \boldsymbol{\nabla}^2 F\left(\boldsymbol{\theta}\right) \frac{\partial \boldsymbol{\theta}}{\partial \rho} \tag{3.2.12}$$

D'où, en adaptant à $\widetilde{\boldsymbol{\eta}_1}$ les équations 3.2.5 et 3.2.9 :

$$\frac{\partial d_G}{\partial \rho}\left(\check{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right) = \left(\check{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}_1}\right)^\top \boldsymbol{\nabla}^2 F\left(\check{\boldsymbol{\theta}}\right)\left(\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}\right), \text{ soit} \tag{3.2.13}$$

$$\frac{\partial d_G}{\partial \rho}\left(\check{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right) = (\rho - 1) \times \left(\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}\right)^\top \boldsymbol{\nabla}^2 F\left(\check{\boldsymbol{\theta}}\right)\left(\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}\right). \tag{3.2.14}$$

On cherche un minorant uniforme sur $\mathcal{B}_0$ de cette dérivée partielle. Puisque $\mathcal{B}_0^* = \{\boldsymbol{\theta} \in \mathcal{N} \mid \boldsymbol{\nabla} F\left(\boldsymbol{\theta}\right) \in \mathcal{B}_0\}$ est un compact de $\mathcal{N}$, le réel

$$c = \min_{\boldsymbol{\theta} \in \mathcal{B}_0^*} \left((m-1) \times \min \boldsymbol{\nabla}^2 F\left(\boldsymbol{\theta}\right)\right) \tag{3.2.15}$$

existe et est strictement positif. On aboutit à l'inégalité suivante :

$$\frac{\partial d_G}{\partial \rho}\left(\overline{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right) \geq (\rho - 1) \times c \times \left(\min |\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}|\right)^2 \tag{3.2.16}$$

La fonction $\rho \longmapsto d_G\left(\overline{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right)$ est continue, strictement convexe et strictement croissante tant que $d_G\left(\overline{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right) \leq R_0$. Par continuité selon ρ, et en vertu du théorème des valeurs intermédiaires, il existe un réel $\check{\rho} > 1$ tel que $d_G\left(\check{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right) = R_0$. Autrement dit, $\check{\boldsymbol{\eta}}$ est à l'intersection de la « surface » de la boule $\mathcal{B}_0$ et de la demi-géodésique $\mathcal{D}^*$. Puisque le vecteur $\check{\boldsymbol{\eta}}$ est la solution optimale du problème dual 3.2.4, et qu'il respecte la contrainte du problème primal $d_G\left(\check{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right) \leq R_0$, on a en fait $\check{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}$, ce qui achève la résolution de ce dernier.

3.2.5 *Algorithme*

En résumé, on a ramené le test d'inclusion 3.2.3, problème non convexe et de dimension élevée, à la recherche de l'antécédent de R_0 par la fonction $\rho \longmapsto d_G\left(\overline{\boldsymbol{\eta}}\|\widetilde{\boldsymbol{\eta}_0}\right)$, définie sur l'intervalle ouvert $]1; +\infty[$. Grâce à sa vitesse de convergence quadratique, la méthode de Newton-Raphson, ou méthode des tangentes, fournit une très bonne approximation de $\hat{\rho}$ en peu d'itérations. En pratique, on pose arbitrairement $\rho\left(0\right) = 2$ et on applique la mise à jour récursive suivante une dizaine de fois.

$$\rho\left(l+1\right) = \rho\left(l\right) - \frac{d_F\left(\widetilde{\boldsymbol{\theta}_0}\|\boldsymbol{\theta}_{\rho(l)}\right) - R_0}{\left(\rho\left(l\right) - 1\right) \times \left(\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}\right)^\top \boldsymbol{\nabla}^2 F\left(\boldsymbol{\theta}_{\rho(l)}\right)\left(\widetilde{\boldsymbol{\theta}_1} - \widetilde{\boldsymbol{\theta}_0}\right)} \tag{3.2.17}$$

hal-00849736, version 1 - 17 Sep 2013

Dans l'expression précédente, nous avons remplacé le terme $d_G\left(\boldsymbol{\eta}_{\rho(l)} \| \widetilde{\boldsymbol{\eta}_0}\right)$ au numérateur par la divergence $d_F\left(\widetilde{\boldsymbol{\theta}_0} \| \boldsymbol{\theta}_{\rho(l)}\right)$, en vertu de la propriété 3.1.11. Ainsi, une fois évaluées les constantes $\widetilde{\boldsymbol{\theta}_0}$ et $\widetilde{\boldsymbol{\theta}_1}$, cet algorithme procède essentiellement par des mises à jour barycentriques et des produits matriciels, des tâches que des langages comme MATLAB ou Python traitent avec beaucoup d'efficacité. A chaque itération, les deux seules opérations qui prennent un temps d'exécution non négligeable concernent le calcul de $d_F\left(\widetilde{\boldsymbol{\theta}_0} \| \boldsymbol{\theta}_{\rho(l)}\right)$ et de $\boldsymbol{\nabla^2} F\left(\boldsymbol{\theta}_{\rho(l)}\right)$. Afin de s'épargner des répétitions inutiles et d'accélérer encore l'algorithme, nous suggérons de garder la valeur constante $F\left(\widetilde{\boldsymbol{\theta}_0}\right)$ en mémoire, et de réutiliser le gradient $\boldsymbol{\nabla} F\left(\boldsymbol{\theta}_{\rho(l)}\right)$, intervenu dans la divergence $d_F\left(\widetilde{\boldsymbol{\theta}_0} \| \boldsymbol{\theta}_{\rho(l)}\right)$, pour le calcul de la matrice hessienne $\boldsymbol{\nabla^2} F\left(\boldsymbol{\theta}_{\rho(l)}\right)$. Par stricte convexité de F, $\boldsymbol{\nabla^2} F$ est définie positive, ce qui permet de définir un système de coordonnées curvilignes sur la famille exponentielle $\mathscr{P}$, appelé *métrique d'information de Fisher*. Initiée par Rao (1945), cette idée a non seulement conduit à l'expression d'une borne inférieure sur la variance d'un estimateur sans biais (dite *borne de Cramér-Rao*), mais est aujourd'hui considérée comme l'acte de naissance de la géométrie de l'information Nielsen (2013). Or, par la formule 3.2.17, nous avons montré que la *matrice d'information de Fisher* $\boldsymbol{\nabla^2} F$ est au cœur de la méthode de Newton-Raphson, quand cette dernière s'applique à une marche géodésique entre deux points.

3.2.6 *Test d'inclusion*

Pour tout couple de segments (σ, σ'), auxquels on a préalablement associé des boules de Bregman $\mathcal{B}_\sigma$ et $\mathcal{B}_{\sigma'}$, on définit le rapport d'inclusion entre $\mathcal{B}_\sigma$ et $\mathcal{B}_{\sigma'}$ comme le réel positif

$$\operatorname{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) = \frac{d_F\left(\widetilde{\boldsymbol{\theta}}_{\sigma'} \| \hat{\boldsymbol{\theta}}\right)}{R_{\sigma'}}, \tag{3.2.18}$$

où $\hat{\boldsymbol{\theta}} = \hat{\rho}\widetilde{\boldsymbol{\theta}_1} + (1-\hat{\rho})\,\widetilde{\boldsymbol{\theta}_0}$ (voir équation 3.2.9) et $\hat{\rho}$ est obtenu en appliquant la méthode de Newton-Raphson comme appliqué précédemment, avec $\widetilde{\boldsymbol{\theta}_0} = \widetilde{\boldsymbol{\theta}}_\sigma$ et $\widetilde{\boldsymbol{\theta}_1} = \widetilde{\boldsymbol{\theta}}_{\sigma'}$. Le test d'inclusion devient :

$$\operatorname{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) \underset{\mathcal{B}_\sigma \not\subset \mathcal{B}_{\sigma'}}{\overset{\mathcal{B}_\sigma \subset \mathcal{B}_{\sigma'}}{\lessgtr}} 1 \tag{3.2.19}$$

3.2.7 *Application à l'ostinato de Ravel*

Un rapport d'inclusion est une quantité positive non bornée. Afin de le représenter sur une échelle de couleur, on a plutôt intérêt à le ramener, de façon monotone sur l'intervalle $[0; 1[$. Sur la figure 3.2.1, on a représenté la quantité

hal-00849736, version 1 - 17 Sep 2013

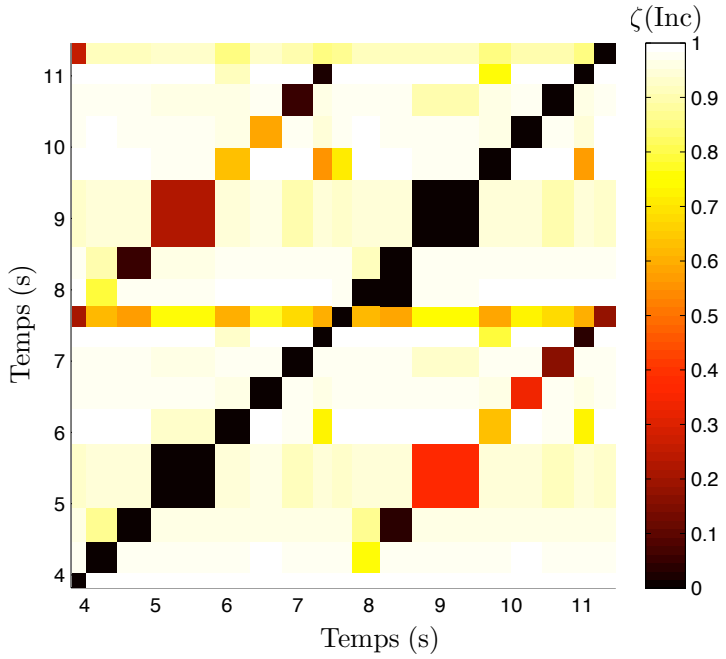


FIGURE 3.2.1: Matrice des rapports d'inclusion de l'ostinato.

$$\zeta\left(\mathrm{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right)\right) = \frac{\mathrm{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right)}{1 + \mathrm{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right)} \tag{3.2.20}$$

pour chaque paire de modèles (σ, σ'). La fonction ζ est monotone, vaut 0 en 0 et 1 vers $+\infty$. Dans cette représentation, c'est autour de 0,5, c'est-à-dire $\mathrm{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) = 1$, que l'échelle de couleur est la plus sensible. On choisit de poser $\zeta = 0$ sur la diagonale principale, en accord avec les matrices d'auto-distance vues précédemment. Ces choix de visualisation n'ont aucune incidence sur la qualité de la reconnaissance, puisque l'on ne tient pas compte de l'auto-similarité de la diagonale principale, afin de ne pas biaiser les résultats.

On constate que, si $\zeta\left(\mathrm{Inc}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right)\right)$ varie de 0 à 0,7 quand les évènements σ et σ' sont similaires, il est presque toujours supérieur à 0,9 quand ils ne le sont pas. Le rapport d'inclusion est donc un bon moyen d'écarter un grand nombre de paires de modèles, et d'ainsi éviter des détections erronnées.

3.3 TEST D'INTERSECTION

3.3.1 *Problème primal*

Soient $\mathcal{B}_0 = (\widetilde{\boldsymbol{\eta}_0}, R_0)$ et $\mathcal{B}_1 = (\widetilde{\boldsymbol{\eta}_1}, R_1)$ deux boules de Bregman appartenant à l'espace $\mathcal{M}$. On cherche à savoir si $\mathcal{B}_0$ et $\mathcal{B}_1$ sont disjointes ou pas. De façon intéressante, il est possible de suivre un raisonnement analogue à celui de la section précédente.

hal-00849736, version 1 - 17 Sep 2013

Soit $\widehat{\boldsymbol{\eta}_1}$ le projeté de $\widetilde{\boldsymbol{\eta}_0}$ sur $\mathcal{B}_1$. Cette fois, il s'agit de *maximiser* le lagrangien $\nu_1 = \nu$ (voir équation 3.2.2) selon le multiplicateur α, puis de le *minimiser* selon $\boldsymbol{\eta}$. La compacité de $\mathcal{B}_1$ nous assure de l'existence d'un optimum $(\widehat{\boldsymbol{\eta}_1}, \widehat{\alpha_1})$.

$$(\widehat{\boldsymbol{\eta}_1}, \widehat{\alpha_1}) = \arg\max_{\alpha_1 \geq 0} \min_{\boldsymbol{\eta} \in \mathcal{M}} \nu_1 (\boldsymbol{\eta}, \alpha_1) \tag{3.3.1}$$

3.3.2 *Condition nécessaire d'optimalité*

Un raisonnement semblable au début de la section précédente conduit à poser

$$\rho_1 = \frac{\alpha_1}{\alpha_1 + 1} \tag{3.3.2}$$

afin de donner à la condition nécessaire $\boldsymbol{\nabla} \nu_1 (\widehat{\boldsymbol{\eta}_1}, \widehat{\alpha_1}) = 0$ la forme :

$$\widehat{\boldsymbol{\theta}_1} = \widehat{\rho_1} \widetilde{\boldsymbol{\theta}_1} + (1 - \widehat{\rho_1}) \widetilde{\boldsymbol{\theta}_0} \tag{3.3.3}$$

3.3.3 *Géodésique*

Cette fois, $\widehat{\boldsymbol{\eta}_1}$ est sur la géodésique $\mathcal{G}^*$ reliant $\widetilde{\boldsymbol{\eta}_0}$ à $\widetilde{\boldsymbol{\eta}_1}$, le sous-ensemble de $\mathcal{M}$ dual de la corde $\mathcal{G}$ reliant $\widetilde{\boldsymbol{\theta}_0}$ à $\widetilde{\boldsymbol{\theta}_1}$.

$$\mathcal{G}^* = \left\{ \boldsymbol{\eta}_\rho = \boldsymbol{\nabla} F (\rho \boldsymbol{\nabla} G (\widetilde{\boldsymbol{\eta}_0}) + (1 - \rho) \boldsymbol{\nabla} G (\widetilde{\boldsymbol{\eta}_0})) \mid \rho \in [0;\, 1] \right\} \tag{3.3.4}$$

De même, en intervertissant le rôle des boules $\mathcal{B}_0$ et $\mathcal{B}_1$, on définit le lagrangien du problème symétrique :

$$\nu_0 : (\boldsymbol{\eta}, \alpha_0) \in \mathcal{M} \times \mathbb{R}_+ \longmapsto d_G (\boldsymbol{\eta} \| \widetilde{\boldsymbol{\eta}_0}) + \alpha_0 \times (R_1 - d_G (\boldsymbol{\eta} \| \widetilde{\boldsymbol{\eta}_1})) \tag{3.3.5}$$

D'où, en posant $\rho_0 = \dfrac{\alpha_0}{\alpha_0 + 1}$, et en écrivant $\boldsymbol{\nabla} \nu_0 (\widehat{\boldsymbol{\eta}_0}, \widehat{\alpha_0}) = 0$:

$$\widehat{\boldsymbol{\theta}_0} = \widehat{\rho_0} \widetilde{\boldsymbol{\theta}_0} + (1 - \widehat{\rho_0}) \widetilde{\boldsymbol{\theta}_1} \in \mathcal{G} \tag{3.3.6}$$

Par conséquent, $\widehat{\boldsymbol{\eta}_0}$ et $\widehat{\boldsymbol{\eta}_1}$ font tous deux parties de la géodésique $\mathcal{G}^*$. En fait, puisque $\widehat{\boldsymbol{\eta}_1} \in \mathcal{G}^*$, on peut définir alternativement $\widehat{\boldsymbol{\eta}_0}$ comme le projeté de $\widehat{\boldsymbol{\eta}_1}$ sur $\mathcal{B}_0$, et $\widehat{\boldsymbol{\eta}_1}$ comme le projeté de $\widehat{\boldsymbol{\eta}_0}$ sur $\mathcal{B}_1$.

3.3.4 *Théorème de Pythagore généralisé*

Supposons temporairement que $\mathcal{B}_0 \cap \mathcal{B}_1 \neq \emptyset$. On cherche à prouver qu'alors $\mathcal{B}_0 \cap \mathcal{B}_1 \cap \mathcal{G}^* \neq \emptyset$. Pour ce faire, il est suffisant de montrer qu'en particulier $\widehat{\boldsymbol{\eta}_0} \in \mathcal{B}_0 \cap \mathcal{B}_1$ ou $\widehat{\boldsymbol{\eta}_1} \in \mathcal{B}_0 \cap \mathcal{B}_1$. Par l'absurde, supposons qu'aucune de ces deux propriétés ne soit vérifiée. Soit

$$\mathbb{1}_{\mathcal{B}_1} : \boldsymbol{\eta} \in \mathcal{M} \longmapsto \begin{cases} 1 & \text{si } \boldsymbol{\eta} \in \mathcal{B}_1 \\ 0 & \text{sinon} \end{cases} \tag{3.3.7}$$

hal-00849736, version 1 - 17 Sep 2013

la fonction indicatrice de l'ensemble $\mathcal{B}_1$. Le théorème de dualité forte de Fenchel[3] donne :

$$\min_{\boldsymbol{\eta}\in\mathcal{M}} \left[d_G\left(\boldsymbol{\eta}\|\widehat{\boldsymbol{\eta}_1}\right) + \mathbb{1}_{\mathcal{B}_1}\left(\boldsymbol{\eta}\right)\right] = \max_{\boldsymbol{\theta}\in\mathcal{N}} \left[-d_G^*\left(\boldsymbol{\theta}\|\widehat{\boldsymbol{\theta}_1}\right) - \mathbb{1}_{\mathcal{B}_1}^*\left(-\boldsymbol{\theta}\right)\right] \quad (3.3.8)$$

La fonction conjuguée de l'indicatrice vaut, par définition :

$$\mathbb{1}_{\mathcal{B}_1}^* : \boldsymbol{\theta_1} \in \mathcal{N} \longmapsto \max_{\boldsymbol{\eta_1}\in\mathcal{B}_1} \left(-\boldsymbol{\theta}^\top \boldsymbol{\eta_1}\right) \quad (3.3.9)$$

Quant à la conjuguée de $\boldsymbol{\eta} \longmapsto d_G\left(\boldsymbol{\eta}\|\widehat{\boldsymbol{\eta}_1}\right)$, c'est Rockafellar (1970, section 12.3) qui nous la donne :

$$d_G^*\left(\cdot\|\widehat{\boldsymbol{\theta}_1}\right) : \boldsymbol{\theta} \in \mathcal{N} \longmapsto F\left(\boldsymbol{\theta} + \widehat{\boldsymbol{\theta}_1}\right) + G\left(\widehat{\boldsymbol{\eta}_1}\right) - \widehat{\boldsymbol{\theta}_1}^\top \widehat{\boldsymbol{\eta}_1} \quad (3.3.10)$$

Tandis que $\widehat{\boldsymbol{\eta}_0}$ minimise le problème primal — membre de gauche de l'égalité 3.3.8 — c'est $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}_0} - \widehat{\boldsymbol{\theta}_1}$ qui maximise le problème dual[4]. On a par conséquent :

$$\begin{aligned}
d_G\left(\widehat{\boldsymbol{\eta}_0}\|\widehat{\boldsymbol{\eta}_1}\right) &= -d_G^*\left(\widehat{\boldsymbol{\theta}}\|\widehat{\boldsymbol{\theta}_1}\right) - \max_{\boldsymbol{\eta_1}\in\mathcal{B}_1}\left(-\boldsymbol{\theta}^\top\boldsymbol{\eta_1}\right) \\
&= -\sup_{\boldsymbol{\eta}\in\mathcal{M}}\left(\left(\widehat{\boldsymbol{\theta}_0} - \widehat{\boldsymbol{\theta}_1}\right)^\top \boldsymbol{\eta} - \left(G\left(\boldsymbol{\eta}\right) - G\left(\widehat{\boldsymbol{\eta}_1}\right) - \widehat{\boldsymbol{\theta}_1}\left(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}_1}\right)\right)\right) \\
&\quad - \max_{\boldsymbol{\eta_1}\in\mathcal{B}_1}\left(\widehat{\boldsymbol{\theta}_1} - \widehat{\boldsymbol{\theta}_0}\right)^\top \boldsymbol{\eta_1} \\
&= -\sup_{\boldsymbol{\eta}\in\mathcal{M}}\left(\widehat{\boldsymbol{\theta}_0}^\top\boldsymbol{\eta} - G\left(\boldsymbol{\eta}\right)\right) - G\left(\widehat{\boldsymbol{\eta}_1}\right) + \widehat{\boldsymbol{\theta}_1}^\top\boldsymbol{\eta_1} \\
&\quad - \max_{\boldsymbol{\eta_1}\in\mathcal{B}_1}\left(\widehat{\boldsymbol{\theta}_1} - \widehat{\boldsymbol{\theta}_0}\right)^\top \boldsymbol{\eta_1} \\
&= G\left(\widehat{\boldsymbol{\eta}_0}\right) - \widehat{\boldsymbol{\eta}_0}^\top\widehat{\boldsymbol{\theta}_0} - G\left(\widehat{\boldsymbol{\eta}_1}\right) - \widehat{\boldsymbol{\theta}_1}^\top\widehat{\boldsymbol{\eta}_0} + \widehat{\boldsymbol{\theta}_1}\widehat{\boldsymbol{\eta}_0} + \widehat{\boldsymbol{\theta}_1}^\top\widehat{\boldsymbol{\eta}_1} \\
&\quad - \max_{\boldsymbol{\eta_1}\in\mathcal{B}_1}\left(\widehat{\boldsymbol{\theta}_1} - \widehat{\boldsymbol{\theta}_0}\right)^\top \boldsymbol{\eta_1} \\
&= d_G\left(\widehat{\boldsymbol{\eta}_0}\|\widehat{\boldsymbol{\eta}_1}\right) - \max_{\boldsymbol{\eta_1}\in\mathcal{B}_1}\left(\widehat{\boldsymbol{\theta}_1} - \widehat{\boldsymbol{\theta}_0}\right)^\top\left(\boldsymbol{\eta_1} - \widehat{\boldsymbol{\eta}_0}\right) \qquad (3.3.11)
\end{aligned}$$

On conclut avec :

$$\max_{\boldsymbol{\eta_1}\in\mathcal{B}_1}\left(\widehat{\boldsymbol{\theta}_1} - \widehat{\boldsymbol{\theta}_0}\right)^\top\left(\boldsymbol{\eta_1} - \widehat{\boldsymbol{\eta}_0}\right) = 0 \quad (3.3.12)$$

Soit $\boldsymbol{\eta}_a$ un élément arbitraire de $\mathcal{B}_0 \cap \mathcal{B}_1$. On a :

$$\begin{aligned}
&d_G\left(\boldsymbol{\eta_a}\|\widehat{\boldsymbol{\eta}_1}\right) - d_G\left(\boldsymbol{\eta_a}\|\widehat{\boldsymbol{\eta}_0}\right) - d_G\left(\widehat{\boldsymbol{\eta}_0}\|\widehat{\boldsymbol{\eta}_1}\right) \\
&= \widehat{\boldsymbol{\theta}_1}^\top\left(\widehat{\boldsymbol{\eta}_1} - \boldsymbol{\eta_a}\right) + \widehat{\boldsymbol{\theta}_0}^\top\left(\boldsymbol{\eta_a} - \widehat{\boldsymbol{\eta}_0}\right) + \widehat{\boldsymbol{\theta}_1}^\top\left(\widehat{\boldsymbol{\eta}_0} - \widehat{\boldsymbol{\eta}_1}\right) \\
&= \left(\widehat{\boldsymbol{\theta}_0} - \widehat{\boldsymbol{\theta}_1}\right)^\top\left(\boldsymbol{\eta_a} - \widehat{\boldsymbol{\eta}_0}\right) \qquad (3.3.13)
\end{aligned}$$

3. voir à ce sujet Rockafellar (1970, section 31).
4. voir à ce sujet Rockafellar (1970, section 23 et 31).

hal-00849736, version 1 - 17 Sep 2013

Puisque $\boldsymbol{\eta_a} \in \mathcal{B}_1$, le lemme 3.3.12 nous permet d'écrire l'inégalité suivante :

$$\left(\widehat{\boldsymbol{\theta}_0} - \widehat{\boldsymbol{\theta}_1}\right)^\top (\boldsymbol{\eta_a} - \widehat{\boldsymbol{\eta}_0}) \geq \min_{\boldsymbol{\eta_1} \in \mathcal{B}_1} \left(\widehat{\boldsymbol{\theta}_0} - \widehat{\boldsymbol{\theta}_1}\right)^\top (\boldsymbol{\eta_1} - \widehat{\boldsymbol{\eta}_0}) = 0 \tag{3.3.14}$$

Soit, en remontant les calculs effectués en 3.3.13 :

$$d_G\left(\boldsymbol{\eta_a} \| \widehat{\boldsymbol{\eta}_1}\right) \geq d_G\left(\boldsymbol{\eta_a} \| \widehat{\boldsymbol{\eta}_0}\right) + d_G\left(\widehat{\boldsymbol{\eta}_0} \| \widehat{\boldsymbol{\eta}_1}\right) \tag{3.3.15}$$

A première vue, la relation ci-dessus a la forme d'une inégalité triangulaire, une propriété vérifiée par toutes les distances (mais pas nécessairement les divergences) d'un espace métrique. Pourtant, il s'agit en fait de l'inégalité inverse, et seulement valable parce que $\widehat{\boldsymbol{\eta}_1}$ est le projeté de $\widehat{\boldsymbol{\eta}_0}$ sur le convexe $\mathcal{B}_1$. Si $\mathcal{B}_1$ était affine, il est aisé de démontrer que 3.3.15 deviendrait une égalité. Avec une divergence de Bregman euclidienne (voir sous-section 3.1.3), on retrouve... le théorème de Pythagore ! C'est ce qui vaut au résultat présenté ici le nom de *théorème de Pythagore généralisé*.

3.3.5 *Résolution*

Comme on l'a évoqué plus haut, $\widehat{\boldsymbol{\eta}_0}$ peut être vu comme le projeté de $\widehat{\boldsymbol{\eta}_1}$ (et plus seulement $\widetilde{\boldsymbol{\eta}_1}$) sur le convexe $\mathcal{B}_0$. Dès lors, on peut écrire le théorème de Pythagore généralisé une seconde fois, en invoquant le fait que $\boldsymbol{\eta_a}$ appartient à $\mathcal{B}_0$. Toutes les équations 3.3.7 à 3.3.14 peuvent être réitérées en permutant les indices « 0 » et « 1 ».

$$d_G\left(\boldsymbol{\eta_a} \| \widehat{\boldsymbol{\eta}_1}\right) + d_G\left(\widehat{\boldsymbol{\eta}_1} \| \widehat{\boldsymbol{\eta}_0}\right) \leq d_G\left(\boldsymbol{\eta_a} \| \widehat{\boldsymbol{\eta}_0}\right) \tag{3.3.16}$$

En soustrayant 3.3.15 à l'inégalité ci-dessus, on obtient

$$d_G\left(\widehat{\boldsymbol{\eta}_1} \| \widehat{\boldsymbol{\eta}_0}\right) \leq -d_G\left(\widehat{\boldsymbol{\eta}_0} \| \widehat{\boldsymbol{\eta}_1}\right) \leq 0\,; \tag{3.3.17}$$

d'où, par positivité de la divergence de Bregman $d_G\left(\widehat{\boldsymbol{\eta}_0} \| \widehat{\boldsymbol{\eta}_1}\right)$,

$$d_G\left(\widehat{\boldsymbol{\eta}_1} \| \widehat{\boldsymbol{\eta}_0}\right) = d_G\left(\widehat{\boldsymbol{\eta}_0} \| \widehat{\boldsymbol{\eta}_1}\right) = 0. \tag{3.3.18}$$

On a donc $\widehat{\boldsymbol{\eta}_0} = \widehat{\boldsymbol{\eta}_1}$; soit, puisque par hypothèse $\widehat{\boldsymbol{\eta}_0} \in \mathcal{B}_0$ et $\widehat{\boldsymbol{\eta}_1} \in \mathcal{B}_1$, $\widehat{\boldsymbol{\eta}_0} \in \mathcal{B}_0 \cap \mathcal{B}_1$. Or, on avait supposé $\widehat{\boldsymbol{\eta}_0} \notin \mathcal{B}_0 \cap \mathcal{B}_1$ au début de cette section, ce qui est absurde.

3.3.6 *Test d'intersection*

En associant le raisonnement par l'absurde détaillé plus haut et la contraposée de la propriété triviale $\mathcal{B}_0 \cap \mathcal{B}_1 = \emptyset \Rightarrow \widehat{\boldsymbol{\eta}_1} \notin \mathcal{B}_0$, on conclut avec l'équivalence suivante :

$$\mathcal{B}_0 \cap \mathcal{B}_1 \neq \emptyset \iff \widehat{\boldsymbol{\eta}_1} \in \mathcal{B}_0 \tag{3.3.19}$$

hal-00849736, version 1 - 17 Sep 2013

Ce résultat, initialement dû à Cayton (2008), nous donne un critère simple d'intersection entre deux boules de Bregman $\mathcal{B}_\sigma$ et $\mathcal{B}_{\sigma'}$. Il suffit d'estimer le projeté $\widehat{\boldsymbol{\eta}_{\sigma'}}$ de $\widetilde{\boldsymbol{\eta}_\sigma}$ sur $\mathcal{B}_{\sigma'}$, et de calculer le rapport d'intersection suivant :

$$\text{Int}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) = \frac{d_G\left(\widehat{\boldsymbol{\eta}_{\sigma'}} \| \widetilde{\boldsymbol{\eta}_\sigma}\right)}{R_\sigma} = \frac{d_F\left(\widetilde{\boldsymbol{\theta}_\sigma} \| \widehat{\boldsymbol{\theta}_{\sigma'}}\right)}{R_\sigma} \tag{3.3.20}$$

3.3.7 *Algorithme*

Afin d'évaluer $\widehat{\boldsymbol{\theta}}_{\sigma'}$, on procède par dichotomie sur $\rho \in [0; 1]$. Une fois posé arbitrairement $\rho(0) = 0.5$, on calcule, à chaque itération $l \geq 0$, $d_F\left(\widetilde{\boldsymbol{\theta}_{\sigma'}} \| \boldsymbol{\theta}_{\rho(l)}\right)$. Si cette divergence est plus grande que $R_{\sigma'}$, on incrémente ρ de 2^{-l}. Dans le cas contraire, on décrémente ρ de 2^{-l}. La recherche dichotomique ayant une vitesse de convergence exponentielle, il suffit de se restreindre à une dizaine d'itérations pour obtenir une approximation satisfaisante de la valeur de $\widehat{\boldsymbol{\theta}}_{\sigma'}$. Le test d'intersection se ramène à :

$$\text{Int}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) \underset{\mathcal{B}_\sigma \cap \mathcal{B}_{\sigma'} = \emptyset}{\overset{\mathcal{B}_\sigma \cap \mathcal{B}_{\sigma'} \neq \emptyset}{\lessgtr}} 1 \tag{3.3.21}$$

3.3.8 *Application à l'ostinato de Ravel*

De même qu'à la sous-section 3.2.7, on a représenté la matrice de rapports d'intersection de l'ostinato de Ravel, en faisant passer ceux-ci par la fonction

$$\zeta : z \longmapsto \frac{z}{1+z}. \tag{3.3.22}$$

Quand deux évènements σ et σ' sont similaires, la quantité $\zeta\left(\text{Int}\left(\sigma \| \sigma'\right)\right)$ s'étend de 0 à 0,4. Par ailleurs, si l'on fait exception des évènements des silence, celle-ci est toujours supérieure à 0,6 quand σ et σ' sont différents. De façon générale, il nous semble que le rapport d'intersection est un critère plus efficace que le rapport d'inclusion.

3.4 CONSTRUCTION D'UNE MÉTRIQUE MIXTE

Traditionnellement, la dissimilarité entre deux segments sonores σ et σ'est calculée comme une divergence (le plus souvent KL) entre $\widetilde{\boldsymbol{\eta}_\sigma}$ et $\widetilde{\boldsymbol{\eta}_{\sigma'}}$, leurs centroïdes à droite respectifs. Nous avons vu, à la fin de la section 3.1, que ce choix n'est légitime que quand le spectre de Fourier est quasi-stationnaire à l'échelle du segment, hypothèse non vérifiée en pratique. Nous souhaitons ici exploiter les résultats des deux sections précédentes afin de construire une métrique de dissimilarité enrichie, qui prenne en compte la variabilité temporelle

hal-00849736, version 1 - 17 Sep 2013

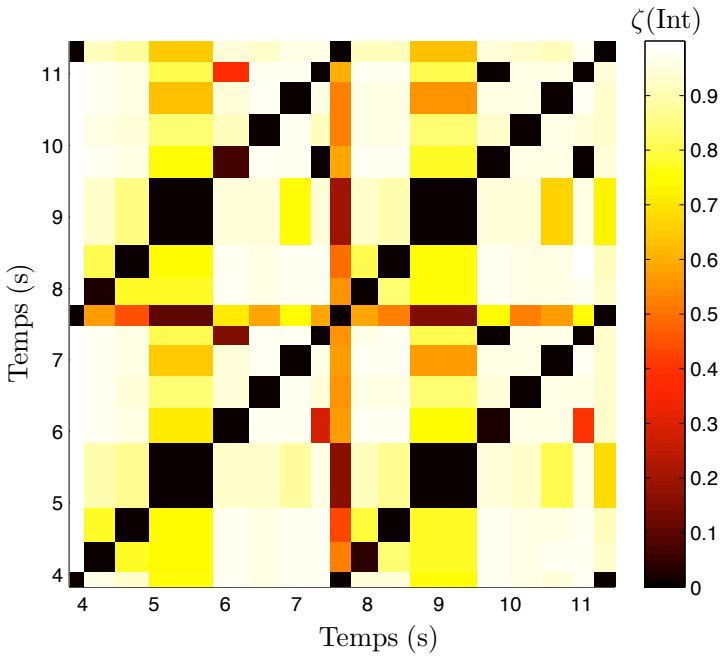


FIGURE 3.3.1: Matrice des rapports d'intersection de l'ostinato.

des signaux. Nous proposons donc, une fois définies des boules de Bregman $\mathcal{B}_\sigma$ et $\mathcal{B}_{\sigma'}$, de calculer le produit

$$D\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) = d_G\left(\widetilde{\boldsymbol{\eta}_\sigma} \| \widetilde{\boldsymbol{\eta}_{\sigma'}}\right) \times \xi_{\text{Inc}}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right) \times \xi_{\text{Int}}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right), \quad (3.4.1)$$

où $\xi_{\text{Inc}}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right)$ et $\xi_{\text{Int}}\left(\mathcal{B}_\sigma \| \mathcal{B}_{\sigma'}\right)$ sont des coefficients de pondération, dépendant des rapports d'inclusion et d'intersection entre les boules $\mathcal{B}_\sigma$ et $\mathcal{B}_{\sigma'}$ (voir équations 3.2.18 et 3.3.20). L'heuristique suivante, qui juxtapose un quart d'ellipse, une plage constante et une branche exponentielle, nous a semblé intéressante.

$$\xi : z \in \mathbb{R}_+ \longmapsto \begin{cases} \sqrt{1 - \left(1 - \dfrac{z}{z_0}\right)^2} & \text{si} \quad 0 \leq z < z_0 \\ 1 & \text{si} \quad z_0 \leq z < z_1 \\ \exp\left((z - z_1)^2\right) & \text{si} \quad z_1 \leq z \end{cases} \quad (3.4.2)$$

Les paramètres z_0 et z_1 sont des seuils constants. Intuitivement, z_0 est censé minorer l'ensemble des rapports géométriques entre segments non similaires, tandis que z_1 majore l'ensemble des rapports géométriques entre segments similaires. Entre ces deux valeurs, les rapports géométriques sont incertains : afin de ne pas risquer de déteriorer la métrique, ξ est alors égal à l'identité.

À titre d'exemple, nous avons représenté, sur la figure 3.4.1, la fonction ξ que nous avons utilisée pour pondérer les rapports d'intersection. On a dans ce cas : $z_0 = 0{,}4$ et $z_1 = 0{,}9$. Ces seuils ont étés fixés

hal-00849736, version 1 - 17 Sep 2013

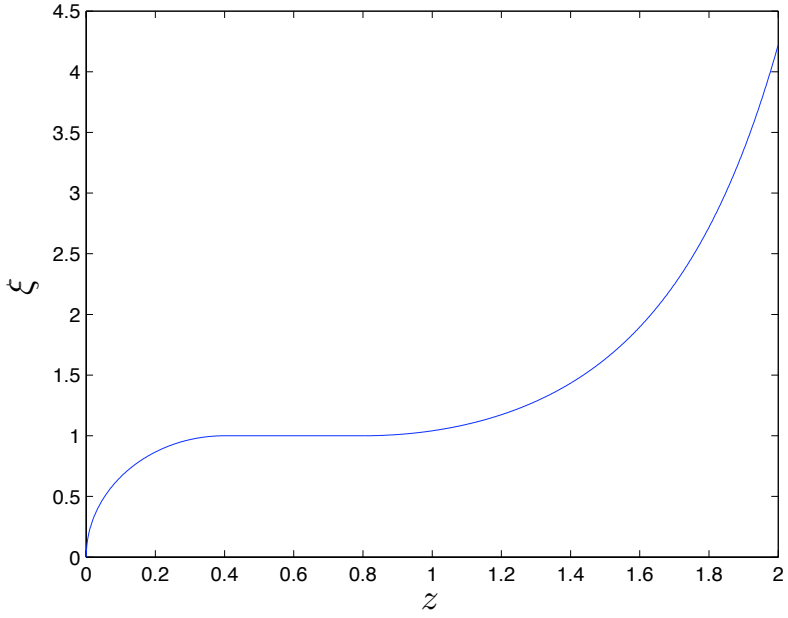


FIGURE 3.4.1: Exemple de fonction de pondération ξ.

manuellement, par un processus d'essais et d'erreurs. Dans le cadre d'une évaluation à grande échelle, il serait préférable de les optimiser automatiquement.

Les figures de ce chapitre permettent de visualiser que la métrique mixte permet de mieux distinguer les cellules structurelles répétées dans un extrait de musique, par rapport à un simple calcul de divergence KL sur les centroïdes de chaque modèle. Au chapitre suivant, nous quantifions l'apport de la métrique mixte par rapport à d'autres approches, lors de tests de difficulté réaliste.

hal-00849736, version 1 - 17 Sep 2013

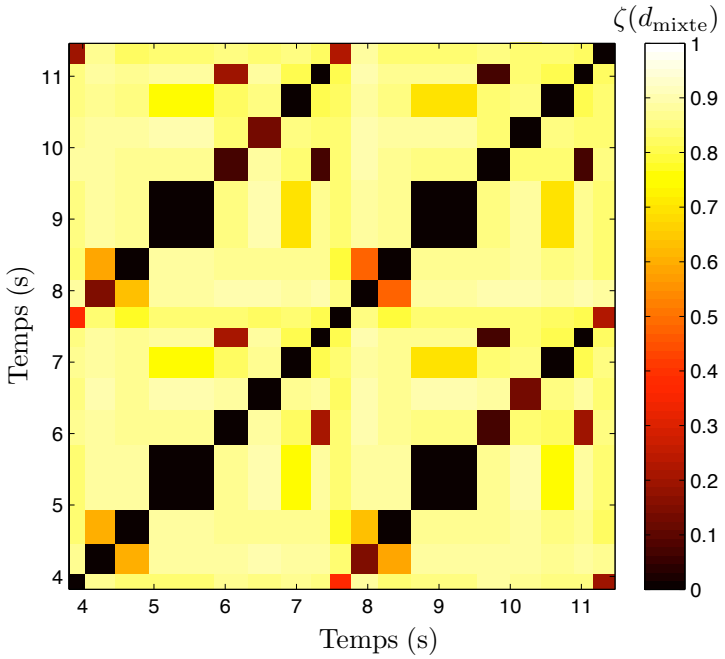


FIGURE 3.4.2: Matrice de dissimilarité mixte de l'ostinato.

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

4 APPLICATIONS AUX SIGNAUX DE MUSIQUE

Dans ce chapitre, on présente deux expériences de découverte de structures appliquées aux signaux de musique. La première porte sur un extrait de piano comprenant l'ostinato étudié précédemment, tandis que la seconde est un thème de jazz pour quatre soufflants. Si l'extrait de piano est convenablement représenté par un spectre de Fourier, l'extrait de jazz nécessite une observation plus stable aux déformations temporelles, appelée *spectre de scattering*. Dans les deux cas, l'apport de la géométrie de l'information augmente considérablement les performances du système.

L'évaluation d'un système de reconnaissance n'est pas une tâche aisée. Le plus délicat est souvent de s'accorder sur une référence jugée parfaite, appelée *ground truth* en anglais. Dans la première expérience, cette référence est automatiquement issue d'une transcription symbolique des évènements livrée avec la base de données. En revanche, dans la seconde expérience, nous avons produit la référence « à l'oreille », sans passer par une transcription mélodique.

4.1 STRUCTURATION D'UN EXTRAIT DE PIANO

4.1.1 *Méthode d'évaluation*

Outre ses nombreux fichiers sonores, la base RWC est accompagnée d'une annotation MIDI, dans laquelle tous les évènements sont répertoriées en temps absolu, comme des activations de hauteurs. Afin de lire automatiquement cette annotation, on a utilisé la boîte à outils conçue par Schutte (2006). Un point de changement est exprimé par un message `NOTE ON` ou `NOTE OFF`. On adopte une tolérance de 50 ms pour agréger des points de changement non exactement simultanés. Afin de ne pas biaiser l'évaluation, on n'inclut pas la diagonale principale ni les similarités entre zones de silence.

On a représenté la matrice de dissimilarité vraie à la figure 4.1.1. On reconnaît, dans le quart inférieur gauche, l'ostinato étudié au chapitre précédent.

Pour une mesure de dissimilarité d et un certain seuil de décision $\delta \in]0; 1[$, le test

$$d\left(\sigma \| \sigma'\right) \underset{H_1}{\overset{H_0}{\lessgtr}} \delta \tag{4.1.1}$$

détermine si les segments σ et σ' sont similaires (hypothèse nulle H_0) ou pas (hypothèse alternative H_1). On aboutit à une matrice binaire de dissimilarité binaire, dont les coefficients valent 0 si H_0 est vérifiée

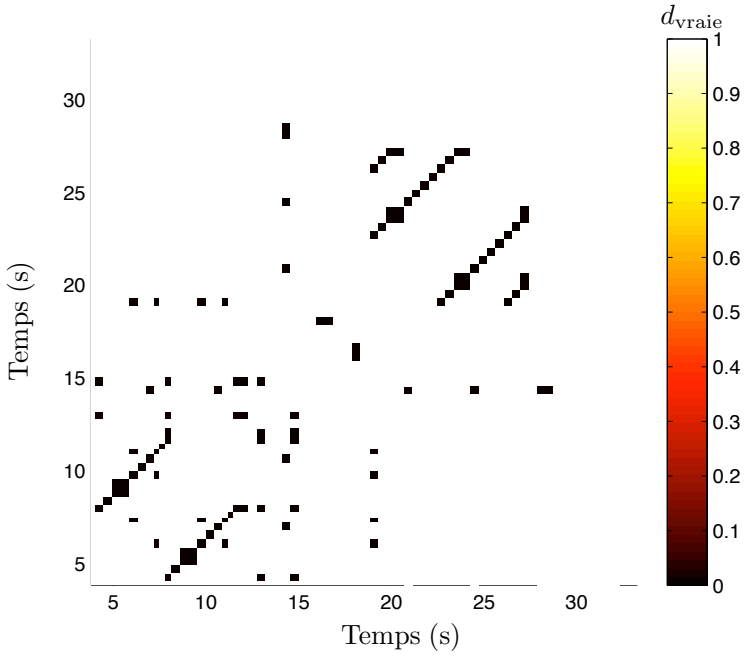


FIGURE 4.1.1: Matrice de dissimilarité vraie pour l'extrait de piano.

et 1 si H_1 est vérifiée. Dès lors, cette matrice binaire peut être évaluée par sa *précision*

$$\mathsf{P}(\delta) = \frac{\text{nombre de similarités correctement retrouvées}}{\text{nombre total de similarité retrouvées}} \tag{4.1.2}$$

et son *rappel*

$$\mathsf{R}(\delta) = \frac{\text{nombre de similarités correctement retrouvées}}{\text{nombre total de vraies similarités}}, \tag{4.1.3}$$

deux nombres compris entre 0 et 1. Par construction, $\mathsf{R}(\delta)$ est une fonction décroissante sur l'intervalle ouvert $]0; 1[$.

Une fois ces deux quantités calculées, il est fréquent de mesurer la performance globale d'un algorithme de reconnaissance par sa *F-mesure*, c'est-à-dire la moyenne harmonique de la précision et du rappel. La F-mesure est aussi un nombre compris entre 0 et 1.

$$\mathsf{F}(\delta) = 2 \times \frac{\mathsf{P}(\delta) \times \mathsf{R}(\delta)}{\mathsf{P}(\delta) + \mathsf{R}(\delta)} \tag{4.1.4}$$

Remarquons qu'un système parfait aurait, pour tout seuil $\delta \in]0; 1[$, une précision de 1 et un rappel de 1, d'où une F-mesure de 1. En pratique, on recherche, par pas de 0,001, le seuil δ maximisant F. La précision et le rappel dits optimaux sont alors les valeurs de $\mathsf{P}(\delta)$ et $\mathsf{R}(\delta)$ correspondantes.

hal-00849736, version 1 - 17 Sep 2013

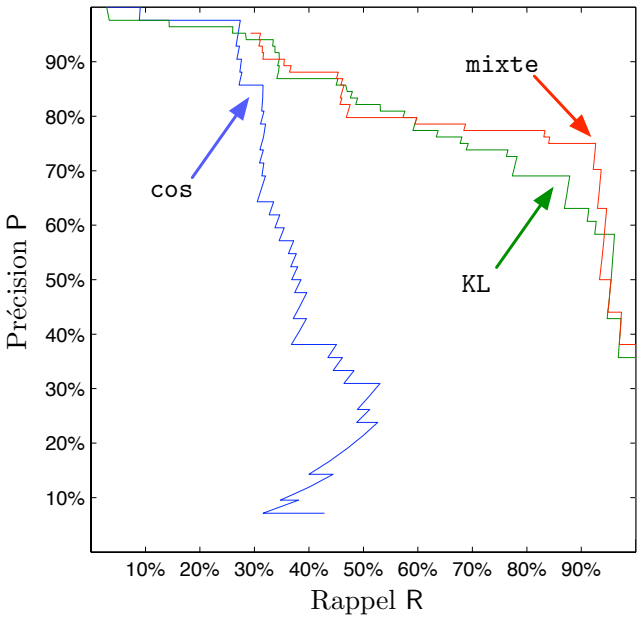


FIGURE 4.1.2: Courbes précision/rappel pour différentes structurations de l'extrait de piano. Les flèches pointent les coordonnées (P, R) maximisant la F-mesure F.

	cos	KL	mixte
P	85,7%	69,0%	75,0%
R	31,6%	87,8%	92,7%
F	46,1%	77,3%	82,9%

TABLE 4.1.1: Résultats de découverte de structures pour l'extrait de piano.

4.1.2 *Résultats*

On a réalisé le protocole précédent pour trois mesures de dissimilarité :

- `cos`, relatif à $\cos(\lambda, \lambda') = 0.5 - \lambda^\top \lambda'$, le cosinus de l'angle que font λ et λ' sur la portion de sphère $\mathcal{S}$; sous le nom de *cosine distance*, elle est souvent utilisée par la communauté MIR, et notamment par Foote (2000) ;
- `KL`, la divergence de Kullback-Leibler entre centroïdes, introduite à la section 3.1 ; et
- `mixte`, la métrique mixte introduite à la section 3.4.

La figure 4.1.2 résume les résultats de l'expérience. Les valeurs de P et R optimisant F sont rassemblées dans le tableau 4.1.1. On constate que la métrique mixte parvient à atteindre des performances légèrement supérieures par rapport aux deux autres mesures de dissimilarité.

hal-00849736, version 1 - 17 Sep 2013

4.1.3 *Extension à des flux d'histogrammes arbitraires*

De toute évidence, la notion de métrique mixte n'est pas limitée au spectre de Fourier. Le flux d'histogrammes $\boldsymbol{h}^t$ introduit au début du chapitre 2 peut être défini avec n'importe quelle observation de dimension $m > 1$ sur le signal, pourvu qu'une sorte de loi de conservation de l'énergie y soit respectée. Par conséquent, il est immédiat d'adapter le cadre géométrique développé au chapitre 3 à de nombreux descripteurs utilisés par la communauté MIR, tels que les bancs de filtres auditifs ou les chromagrammes.

Pour illustrer cette généralisation, nous présentons, à la section 4.3, une seconde expérience de découverte de structures. Celle-ci ayant pour objet d'étude un signal de difficulté réaliste, il est nécessaire de le représenter par une observation plus robuste aux petites déformations temporelles que le spectre de Fourier. Nous introduisons brièvement le spectre de scattering à la section suivante, en motivant son emploi ; ensuite, nous expliquons comment adapter le procédé de découverte de structures décrit précédemment à ce nouvel outil.

4.2 VERS LE SPECTRE DE SCATTERING

On se place dans l'espace de Hilbert $L^2(\mathbb{R})$ des fonctions de carré intégrable. Le temps s et la fréquence ω sont des grandeurs réelles.

4.2.1 *Invariance et stabilité*

Soit y un signal continu de carré intégrable. Etant donné $c \in \mathbb{Z}$, on définit $y_c(s) = y(s-c)$, issu de y par une translation temporelle de c échantillons. On a une relation de déphasage entre les transformées de Fourier de ces deux signaux :

$$\forall \omega \in \mathbb{R},\ \widehat{y_c}(\omega) = \mathrm{e}^{-\mathrm{i}c\omega}\hat{y}(\omega). \tag{4.2.1}$$

Par conséquent, on dit que le module de la transformée de Fourier est un opérateur *invariant par translation* :

$$\forall \omega \in \mathbb{R},\ |\widehat{y_c}(\omega)| = |\hat{y}(\omega)|. \tag{4.2.2}$$

On appelle $D^1(\mathbb{R})$ l'espace des fonctions dérivables sur $\mathbb{R}$, et

$$D_1^1 = \left\{ \tau \in D^1(\mathbb{R}) \mid \forall s \in \mathbb{R},\ |\tau'(s)| < 1 \right\} \tag{4.2.3}$$

l'ensemble des *petites déformations* sur $\mathbb{R}$. Une représentation $\Phi(y)$ est dite *stable* si et seulement si elle est relativement peu altérée par une petite déformation. Cette propriété peut s'exprimer comme une condition de régularité lipschitzienne, i.e.

$$\exists C > 0,\ \forall \tau \in D_1^1,\ \|\Phi(y) - \Phi(y_\tau)\| \leq C \times \sup|\tau'| \times \|x\|, \tag{4.2.4}$$

hal-00849736, version 1 - 17 Sep 2013

où $y_\tau(s) = y(s - \tau(s))$ est le signal déformé. Malgré son intérêt évident, une représentation fondée sur le module de la transformée de Fourier est gravement *instable*. A titre d'exemple, considérons un signal harmonique de fréquence fondamentale ξ et lentement modulé en amplitude par une fonction g — un modèle simpliste de note de musique comportant un trémolo :

$$y(s) = g(s) \times \sum_{k=1}^{K} a_k \cos(k\xi s) \tag{4.2.5}$$

La transformée de Fourier de y est une somme de partiels équidistants. Puisque les variations de g sont lentes, l'essentiel de l'énergie de sa transformée de Fourier $\hat{g}$ est comprise dans un intervalle fréquentiel restreint autour de 0, dont la largeur Δ vaut tout au plus 20 Hz.

$$\hat{y}(\omega) = \sum_{k=1}^{K} \frac{a_k}{2} \times (\hat{g}(\omega - k\xi) + \hat{g}(\omega + k\xi)) \tag{4.2.6}$$

Considérons une dilatation temporelle uniforme $\tau(s) = \epsilon s$, avec $0 < \epsilon \ll 1$. Après déformation, chaque partiel $\widehat{g}(\omega \pm k\xi)$ est translaté de $k\epsilon\xi$. Même si le partiel principal n'est pas beaucoup altéré relativement à sa largeur (ce qui se traduit par $\epsilon\xi < \Delta$), l'entier k rend les spectres $\widehat{y}$ et $\widehat{y_\tau}$ disjoints en hautes fréquences. La distance euclidienne entre $|\widehat{y}|$ et $|\widehat{y_\tau}|$ est alors élevée alors que le produit $\sup|\tau'| \times \|x\|$ est faible, ce qui est en contradiction avec la condition 4.2.4. C'est pour cette raison que l'on souhaite remplacer la transformée de Fourier à court terme, qui s'apparente à un banc de filtres équidistants en fréquence, par un banc de filtres inspiré de l'échelle logarithmique de Mel. Nous nous limitons ici à donner les bases du formalisme des ondelettes ; pour une exploration beaucoup plus approfondie, nous renvoyons le lecteur à l'ouvrage classique de Mallat (2009).

4.2.2 *Transformée en ondelettes*

L'idée centrale de la représentation en ondelettes est de construire un dictionnaire de filtres passe-bande ψ_λ issus d'une « mère » $\psi \in L^2(\mathbb{R})$ par homothétie et translation. Par convention, $\widehat{\psi}$ a pour fréquence centrale 1.

$$\forall \lambda > 0,\ \psi_\lambda(s) = \lambda\psi(\lambda s) \tag{4.2.7}$$

L'ondelette mère considérée ici est dite de Gabor : son module est une fonction gaussiene.

$$\psi(s) = \exp\left(\mathrm{i}s - \frac{Q}{2}s^2\right). \tag{4.2.8}$$

L'entier $Q > 0$, appelé *facteur de qualité*, est égal au nombre de filtres par octave. La valeur $Q = 8$ est très répandue en traitement de la

hal-00849736, version 1 - 17 Sep 2013

parole ; en ce qui nous concerne, nous choisissons intuitivement $Q = 12$ afin de correspondre aux intervalles de la gamme tempérée. Les résultats obtenus sont assez satisfaisants, mais en aucun cas nous ne garantissons que cette valeur soit optimale. Les ondelettes que l'on va construire auront pour fréquences centrales respectives $\lambda = 2^{j/Q}$, où j est un entier relatif.

Le support de $\widehat{\psi_\lambda}$ est centré sur λ et est large d'environ λ/Q , de même que le support de ψ_λ est centré sur 0 et large d'environ $2\pi Q/\lambda$. Or, on a limité la durée d'une observation à T : on restreint donc la définition 4.2.7 à $\lambda \geq 2\pi Q/T$. Les basses fréquences, correspondant à l'intervalle $[0; 2\pi Q/T]$, sont alors couvertes avec $(Q-1)$ filtres équidistants de bande passante $2\pi/T$. On note Λ l'ensemble de toutes les fréquences centrales du banc de filtres obtenu ; celui-ci est « mixte », dans le sens où les fréquences $\lambda \in \Lambda$ croissent d'abord linéairement puis logarithmiquement. Λ est infini en théorie mais, en pratique, il n'est évidemment plus nécessaire de le calculer au-delà de $F_s/2$, c'est-à-dire les confins de l'audition humaine.

La *transformée en ondelettes* Wy rassemble tous les signaux résultant d'une convolution de y par les ondelettes ψ_λ ; ainsi que le signal $y \star \phi$, où ϕ est un filtre passe-bas gaussien de fréquence de coupure T^{-1} :

$$Wy = (y \star \phi,\, y \star \psi_\lambda)_{\lambda \in \Lambda}\,. \tag{4.2.9}$$

4.2.3 *Mel-spectrogramme*

Le banc de filtres d'Andén et Mallat est conçu de telle sorte que l'intégralité de l'axe fréquentiel soit couverte de façon quasi-uniforme : la quantité

$$A(\omega) = \left|\widehat{\phi}(\omega)\right|^2 + \frac{1}{2}\sum_{\lambda \in \Lambda}\left(\left|\widehat{\psi_\lambda}(\omega)\right|^2 + \left|\widehat{\psi_\lambda}(-\omega)\right|^2\right) \tag{4.2.10}$$

vérifie la condition d'admissibilité

$$1 - \alpha \leq A(\omega) \leq 1 \tag{4.2.11}$$

pour toute fréquence $\omega \in \mathbb{R}$, et avec $\alpha = 0{,}02$. La formule de Plancherel (voir Mallat, 2009), qui généralise l'égalité de Parseval à $L^2(\mathbb{R})$, donne

$$(1-\alpha)\,\|y\|^2 \leq \|Wy\|^2 \leq \|y\|^2 \tag{4.2.12}$$

où $\|y\|^2 = \int |x(s)|^2\, \mathrm{d}s$ est la norme L^2 du signal y, et où

$$\|Wy\|^2 = \int |y \star \phi(s)|^2\, \mathrm{d}s + \sum_{\lambda \in \Lambda}\int |y \star \psi_\lambda(t)|^2\, \mathrm{d}t. \tag{4.2.13}$$

Soit Λ_1 un tel banc de filtres. On note

$$\widetilde{W_1}y = (y \star \phi,\, |y \star \psi_{\lambda_1}|)_{\lambda_1 \in \Lambda_1} \tag{4.2.14}$$

hal-00849736, version 1 - 17 Sep 2013

le module de la transformée en ondelettes de y par Λ_1. Le *mel-spectrogramme* S_1 peut alors être défini comme l'ensemble des signaux

$$S_1 y\,(t, \lambda_1) = |y \star \psi_{\lambda_1}| \star \phi\,(s) \tag{4.2.15}$$

avec $\lambda_1 \in \Lambda_1$.

Par rapport à la transformée en ondelettes Wy, le mel-spectrogramme perd une donnéee importante : la phase du signal $y \star \psi_{\lambda_1}$. Cette phase nous informe sur les modulations d'amplitude de y, c'est-à-dire les variations temporelles lentes de l'énergie (de l'ordre de quelques Hertz) dans les bandes de fréquences des ondelettes. En musique, ces modulations peuvent être dues à des modes de jeu spécifiques tels que le vibrato, le trémolo ou le staccato, mais aussi à des interférences polyphoniques. Bien que la fréquence de ces modulations soit en-dessous des limites d'audition de la cochlée, leur rôle dans les processus neuronaux de l'audition est capitale, comme Atlas et Shamma (2003) l'ont démontré.

Pour des fenêtres d'observation de $T = 23\,\text{ms}$, correspondant à une fréquence de coupure de 43 Hz pour ϕ, les modulations d'amplitude peuvent être négligées. En effet, le mel-spectrograme donne des descripteurs localement invariants aux déformations petites devant T. En revanche, $T = 370\,\text{ms}$ correspond à une fréquence de coupure de 2,7 Hz, soit moins que la fréquence typique d'un vibrato. Andén et Mallat (2011) estiment que si 97% de l'énergie de y est captée par le mel-spectrogramme pour $T = 23\,\text{ms}$, cette proportion chute à 70% pour $T = 370\,\text{ms}$. De toute évidence, l'énergie perdue à cause du filtrage ϕ est aussi une *information* perdue, qui risque de nous faire défaut pour comparer des évènements musicaux non stationnaires, et longs de plusieurs dixièmes de seconde.

4.2.4 *Spectre de scattering*

Afin de récupérer cette information, Andén et Mallat (2011) calculent une seconde série de transformées en ondelettes. Pour chacun des signaux $|y \star \psi_{\lambda_1}|$, on calcule

$$\widetilde{W}_2\,|y \star \psi_{\lambda_1}| = (|y \star \psi_{\lambda_1}| \star \phi,\, ||y \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)_{\lambda_2 \in \Lambda_2} \tag{4.2.16}$$

sur le modèle de l'équation 4.2.14. Λ_2 est un banc de filtres à facteur de qualité constant $Q_2 = 1$.

$$S_2 x\,(s, \lambda_1, \lambda_2) = ||y \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi\,(s) \tag{4.2.17}$$

S_2 peut être appréhendé comme un mel-spectrogramme à l'ordre 2 : avant d'échantillonner $|y \star \psi_{\lambda_1}|$ sur une cadence $T = 370\,\text{ms}$, on quantifie l'éparpillement (en anglais *scattering*) de son énergie à basses

hal-00849736, version 1 - 17 Sep 2013

fréquences sur le banc de filtres Λ_2. Ces transformées en ondelettes peuvent être itérées un nombre arbitraire de fois, donnant ainsi lieu à une représentation dite *profonde*. Cependant, étant donnée la durée caractéristique T de notre problème, il n'est pas nécessaire d'effectuer un troisième filtrage : sur les 30% d'énergie manquant au mel-spectrogramme initial, l'ordre 2 parvient à en récupérer 24. Nous renvoyons notre lecteur à Andén et Mallat (2013) pour une présentation plus complète de la transformée de scattering.

4.2.5 *Schéma de structuration multiéchelles*

L'idée de tenir compte des modulations d'amplitude à des fins de découverte de structures en musique est due à (Peeters, 2004), qui proposait déjà d'utiliser des « descripteurs dynamiques » *(dynamic features*). Nous souhaitons poursuivre cet effort en montrant que l'évolution temporelle d'une représentation à court terme, telle que l'énergie dans une bande de fréquence, peut être appréhendée comme une donnée intrinsèque au modèle.

Concevoir un dispositif de découverte de structures, c'est faire face à un dilemme sur la durée T des fenêtres d'observation : si T est trop faible, la moyenne glissante ϕ détruit une part précieuse de l'information musicale ; mais si T est trop grand, les points de rupture ne sont plus détectés avec assez d'acuité, et les modèles se chevauchent. Grâce à l'analyse multiéchelles, il est possible de satisfaire ces deux exigences. L'idée est d'utiliser une représentation à court terme (typiquement $T_{\text{obs}} = 23{,}3\,\text{ms}$) afin de segmenter le signal avec des observations au premier ordre $S_{1,\text{obs}}y\,(\lambda_1, s) = |y \star \psi_{\lambda_1}| \star \phi_{\text{obs}}\,(s)$; mais avant de filtrer par ϕ_{obs}, on stocke temporairement en mémoire les signaux $|y \star \psi_{\lambda_1}|\,(s)$. Quand un point de changement est détecté, on s'arrange pour que les signaux $|y \star \psi_{\lambda_1}|\,(s)$ du segment aient une durée multiple de $T_{\text{seg}} = 370\,\text{ms}$, éventuellement au prix d'une symétrisation ou d'une périodisation. Pour obtenir la représentation à long terme associée, il ne reste plus qu'à calculer $S_{1,\text{seg}}y\,(\lambda_1, s) = |y \star \psi_{\lambda_1}| \star \phi_{\text{seg}}\,(s)$ et $S_2y\,(\lambda_1, \lambda_2, s) = ||y \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{\text{seg}}\,(s)$. Bien sûr, afin de ne pas surcharger la mémoire, on se permettra de sous-échantillonner chacun de ces signaux à une fréquence double de leur bande passante. Ce schéma d'analyse imbrique plus encore les tâches de segmentation et de modélisation, et permet d'éviter de répéter deux fois les mêmes calculs.

La métrique mixte du spectre de scattering comprend alors, d'une part, la divergence KL entre centroïdes respectifs de $\{S_{1,\text{seg}}y\,(s)\,; S_2y\,(s)\}$, et d'autre part, les rapports d'inclusion et d'intersection entre boules d'observations à court terme $S_{1,\text{obs}}y$.

hal-00849736, version 1 - 17 Sep 2013

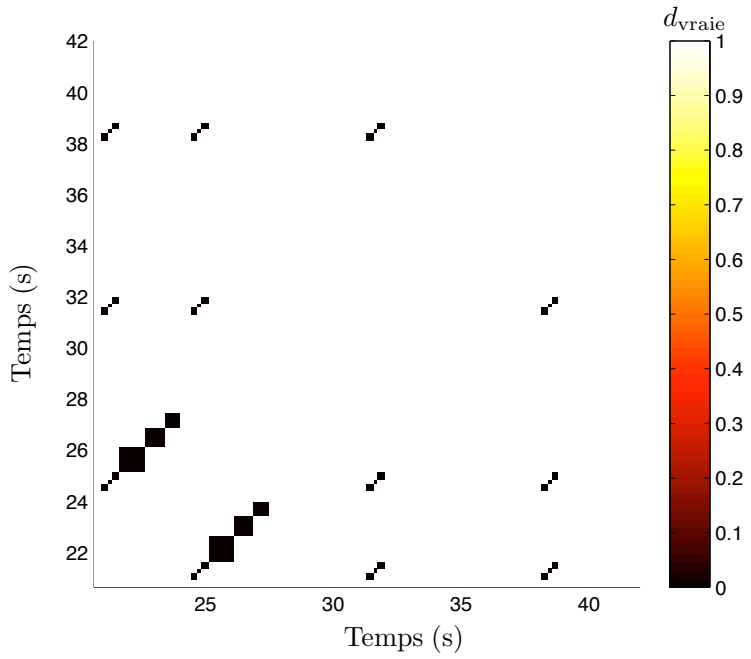


FIGURE 4.3.1: Matrice de dissimilarité vraie pour l'extrait de jazz.

4.3 STRUCTURATION D'UN EXTRAIT DE JAZZ

4.3.1 *Procédé*

Dans l'optique de tester les apports de ce travail dans un contexte aussi large que possible, on a appliqué les algorithmes présentés à un autre extrait de la base RWC, issu du répertoire jazz. Il s'agit de « Lounge Away », un thème de jazz interprété par un choral de quatre instruments à vents, et accompagné par une section rythmique de quatre musiciens. Ce thème suit une forme de blues en douze mesures ; les deux premières phrases sont strictement identiques, et l'amorce initiale de trois notes est retrouvée deux fois de plus dans le reste du thème. L'enregistrement audio porte le nom suivant : RWC-MDB-J-2001 n°28. Nous avons représenté la *ground truth*, transcrite à la main, sur la figure 4.3.1. Encore une fois, on ne tient pas compte de la diagonale principale, ni des similarités entre zones de silence.

4.3.2 *Résultats*

Au cours de cette expérience, nous avons confronté cinq méthodes :

- `Foote` : distance « cosinus » entre coefficients MFCC (voir Foote (2000))
- `Fourier-KL` : divergence KL entre histogrammes de spectre de Fourier,

hal-00849736, version 1 - 17 Sep 2013

	Foote	Fourier-KL	Fourier-mixte	Scat-KL	Scat-mixte
P	47,7%	61,9%	75,8%	83,3%	94,0%
R	23,3%	72,2%	83,3%	74,5%	83,8%
F	31,3%	66,7%	79,4%	78,7%	88,6%

TABLE 4.3.1: Résultats de découverte de structures pour l'extrait de jazz.

- `Fourier-mixte` : métrique mixte entre histogrammes de spectre de Fourier (voir section 3.4),
- `Scat-KL` : divergence KL entre histogrammes de scattering, et
- `Scat-mixte` : métrique mixte entre histogrammes de spectre de scattering (voir sous-section 4.2.5).

Les résultats sont rassemblés dans le tableau 4.3.1. Qu'il s'agisse du spectre de Fourier ou du spectre de scattering, l'utilisation de notre métrique mixte améliore les performances de structuration, tant en termes de précision que de rappel. Il est intéressant de constater que `Fourier-mixte` parvient à atteindre des performances comparables à `Scat-KL` alors que le spectre de Fourier est un descripteur à très court terme. On peut donc dire, en quelque sorte, que le traitement géométrique en aval du calcul de centroïdes compense l'absence de coefficients de second ordre. Profitant de ces deux atouts, la métrique `Scat-mixte` surclasse amplement les autres mesures de dissimilarité.

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

5 CONCLUSION

5.1 RÉCAPITULATIF

Notre stage au sein de l'équipe-projet MuTant était consacré à la découverte automatique de structures musicales en temps réel dans des fichiers sonores, en se fondant sur le cadre de la géométrie de l'information. Dans ce rapport, après avoir présenté l'état de l'art en découverte de structures au chapitre 1, nous avons progressivement défini, au chapitre 2, les outils fondamentaux en géométrie de l'information pour étudier un flux d'observations multidimensionnelles. Nous avons commencé par appréhender chacune des observations comme le paramètre d'une variable aléatoire catégorique ; ensuite, nous avons inclus cette modélisation particulière dans le cadre, plus général, des familles exponentielles de probabilité. Dans un troisième temps, nous avons introduit la notion de vraisemblance d'une réalisation aléatoire, avant de l'étendre sur l'espace des paramètres des moments. Par les propriétés du maximum de vraisemblance d'une famille exponentielle, nous avons construit une statistique exhaustive valable pour tout ensemble d'observations, comme une simple moyenne arithmétique. Puisque la découverte de stuctures doit débuter, selon notre procédé, par une segmentation du flux sonore en évènements musicaux, on a abordé le problème de la détection séquentielle de rupture, en revisitant les rapports de vraisemblance généralisés sous une forme géométrique. Nous avons explicité un algorithme de segmentation reposant sur une règle de décision non bayésienne. Le chapitre 2 s'est terminé par une expérience simple de segmentation en évènements pour un ostinato de piano.

Le chapitre 3 était consacré à la comparaison d'évènements musicaux à partir de leur modélisation géométrique sous-jacente. Notre contribution principale était de rassembler les tâches de segmentation et de calcul de similarité dans le formalisme unique de la géométrie de l'information. Nous avons montré comment les paramètres d'une famille exponentielle imposent une mesure de dissimilarité qui leur est propre : la divergence de Bregman. Après en avoir commenté les propriétés, nous avons exploité cette divergence afin de retrouver des similarités entre évènements musicaux. Pour chaque segment, une fois modélisé le nuage des observations qui le composent comme une boule de Bregman, nous avons proposé d'étudier les relations d'intersection et d'inclusion entre ces boules. Dans ce but, nous avons mis à profit deux algorithmes récents de géométrie computationnelle. Une de nos contributions majeures a été d'introduire des métriques conti-

nues, appelées *rapports d'intersection* et *d'inclusion*, fondées sur les résultats de ces algorithmes ; ainsi que des heuristiques pour quantifier leur importance décisionnelle. Enfin, un apport majeur de ce travail de recherche est d'avoir construit une métrique dite *mixte*, qui prend en compte les trois critères géométriques précédents. L'usage de chacun d'entre eux est illustré par un exemple simple.

Le chapitre 4, enfin, s'était donné pour objectif d'appliquer les contributions théoriques de ce rapport à des extraits musicaux réalistes. Dès lors, nous avons entrepris deux expériences de découverte de structures, faisant appel à des représentations différentes sur le signal. Pour chacune d'entre elles, nous avons proposé une heuristique de métrique mixte, qui se révèle plus performante que des mesures de dissimilarités classiquement présentes dans la littérature scientifique, telle que la « distance cosinus » ou la divergence de Kullback-Leibler. La contribution principale de ce chapitre est de proposer un schéma séquentiel unifié pour utiliser des représentations dites profondes, telles que le spectre de scattering, à des fins de découverte de structures en temps réel.

5.2 BILAN ET PERSPECTIVES

La dernière décennie a vu l'apparition des premiers dispositifs de découverte automatique de structures dans les flux audio. S'épanouissant progressivement en tant que telle, cette problématique a suscité de nombreuses innovations en traitement du signal et en traitement du langage naturel appliqué à la musique. Nous avons montré que ces deux domaines scientifiques opèrent sur des échelles naturellement disjointes — voir figure 1.2.1 pour rappel. Pourtant, les méthodes actuelles en découverte de structures se contentent le plus souvent, pour passer de l'une à l'autre, d'effectuer un calcul de moyenne. Au cours de ce stage, notre travail a notamment consisté à mettre en évidence la richesse de l'information musicale à l'échelle du dixième de seconde : ce faisant, nous avons étudié le comportement de descripteurs sonores *autour* de leurs moyennes respectives. À ce titre, ce stage s'achève par une réussite sur le plan expérimental.

Le dispositif présenté ici se concentre sur la recherche de similarités exactes entre évènements musicaux. Or, de toute évidence, l'oreille du musicien est capable de déceler des similarités entre cellules structurelles, mêmes quand celles-ci sont exprimées selon des tonalités, des orchestrations ou des modes de jeux différents. Il faut donc reconnaître que notre approche ne couvre qu'une partie de la véritable découverte de structures en musique. La construction d'une représentation du signal musical qui soit invariante à ces trois modifications reste une question ouverte. Pour envisager de la résoudre, nous pensons qu'il est nécessaire de créer un système de reconnaissance qui opère sur de nombreuses échelles temporelles, tout en reliant celles-ci

hal-00849736, version 1 - 17 Sep 2013

en un procédé computationnel unifié ; sur le schéma, en somme, de la cognition musicale humaine.

S'agissant de la recherche de similarités exactes, nous sommes conscients que notre machine d'écoute artificielle demeure à l'état de prototype, et est loin d'être prête pour affronter une évaluation à grande échelle. Tout d'abord, si l'on désire poursuivre la comparaison de boules informationnelles, il est évident que les conséquences du choix d'un certain type de centroïde et d'un certain type de boule doivent être explicitées. De plus, si l'on continue à construire des métriques mixtes, il est nécessaire d'exhiber des fonctions de pondération ξ uniformément robustes pour la recherche des similarités. Même si l'on progressait considérablement dans le choix de métriques appropriées pour la comparaison de modèles, le passage à une implémentation commerciale ne serait pas immédiat. L'article de Peeters et al. (2012) décrit bien les nombreux défis à relever pour faire éclore, dans le contexte de la découverte de structures en musique, une innovation scientifique vers le milieu industriel.

Pour autant, l'engouement actuel pour la classification de données en haute dimension suscite un effort de recherche inédit en géométrie de l'information computationnelle. Rassemblée autour de conférences telles que *Geometric Science of Information*, la communauté scientifique du domaine produit des avancées théoriques surprenantes, et est aujourd'hui capable de les expérimenter sur de très grandes bases de données. De plus, la géométrie de l'information réunit des chercheurs venus d'horizons divers mais visant à des résultats souvent très analogues. Ainsi, de même que ce travail s'est abreuvé d'algorithmes pour la fouille de données sémantiques (Cayton, 2008, 2009), on peut espérer qu'il inspirera d'autres champs d'application que le traitement du signal musical.

Nous débuterons, en septembre prochain, un doctorat portant sur l'« analyse géométrique pour la classification de sons », sous la direction de Stéphane Mallat. Une des pistes de développement possibles serait de combiner spectre de scattering et réseaux de neurones profonds, afin d'incorporer des relations structurelles sur des échelles de plus en plus longues. Il ne fait aucun doute qu'à cette occasion, les concepts d'apprentissage statistique et de géométrie différentielle acquis au cours de ce stage seront de précieux alliés.

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

hal-00849736, version 1 - 17 Sep 2013

BIBLIOGRAPHIE

S.-i. Amari, H. Nagaoka. *Methods of Information Geometry. Translations of Mathematical Monographs*, S. Kobayashi, M. Takesaki, éditeurs. American Mathematical Society (AMS), Providence, RI, États-Unis, 2000. (Cité aux pages 6 et 33.)

J. Andén, S. Mallat. Multiscale scattering for audio classification. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Miami, FL, États-Unis, octobre 2011. (Cité aux pages 12 et 55.)

J. Andén, S. Mallat. Deep Scattering Spectrum. *CoRR*, abs/1304.6763, 2013. (Cité à la page 56.)

R. André-Obrecht. *A new statistical approach for the automatic segmentation of continuous speech signals.* Publication interne INRIA n°287, Rennes, France, février 1986. (Cité à la page 13.)

X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals. Speaker Diarization : A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)* 20(1), p. 356–370, 2012. (Cité à la page 8.)

L. Atlas, S. Shamma. Joint acoustic and modulation frequency. *EURASIP J. Appl. Signal Process.*, vol. 7, p. 668–675, 2003. (Cité à la page 55.)

A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6, p. 1705–1749, 2005. (Cité aux pages 33 et 36.)

G. Berry, P. Couronné, G. Gonthier. *Systèmes réactifs et programmation synchrone.* Rapport de recherche INRIA n°524, Sophia-Antipolis, France, mai 1986. (Cité à la page 6.)

B. F. Bowdle, D. Gentner. Information and Asymmetry in Comparisons. *Cognitive Psychology*, 34, p. 244–286, 1997. (Cité à la page 32.)

A. von Brandt. Detecting and estimating parameter jumps using ladder algorithms and likelihood ratio tests. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 8, p. 1017–1020, Boston, MA, États-Unis, avril 1983. (Cité à la page 21.)

L. M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3), p. 200–217. (Cité à la page 30.)

L. Cayton. Fast nearest neighbor retrieval for Bregman divergences. In *25th Int. Conf. on Machine Learning (ICML)*, pages 112–119, Helsinki, Finlande, juillet 2008. (Cité aux pages 44 et 61.)

L. Cayton. Efficient Bregman range search. *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pages 243–251. NIPS Foundation, La Jolla, CA, États-Unis, 2009. (Cité aux pages 11, 29 et 61.)

N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*. *Translations of Mathematicals Monographs*, vol. 23, L. J. Leifman, éditeur. American Mathematical Society (AMS), Providence, RI, États-Unis, 1982. (Cité à la page 6.)

A. Cont. ANTESCOFO : Anticipatory Synchronization and Control of Interactive Parameters in Computer Music. In *Proc. Int. Computer Music Conf. (ICMC)*, Belfast, Irlande, août 2008. (Cité à la page 4.)

A. Cont. *Modeling Musical Anticipation.* Thèse de doctorat, U. San Diego / U. Pierre et Marie Curie, octobre 2008. (Cité à la page 7.)

A. Cont. Synchronisme musical et musiques mixtes : du temps écrit au temps produit. *Circuit : musiques contemporaines* 2(1), p. 9–24. Presses de l'Université de Montréal, QC, Canada, mai 2012. (Cité à la page 5.)

A. Cont, S. Dubnov, G. Assayag. On the Information Geometry of Audio Streams with Applications to Similarity Computing. *IEEE Trans. Audio, Speech and Language Processing (TASLP)*, 19(4), p. 837–846, 2010. (Cité aux pages 5 et 36.)

R. B. Dannenberg, M. Goto. Music structure analysis from acoustic signals. In *Handbook of Signal Processing in Acoustics*, p. 305–331, Springer, New York, États-Unis, 2009.

G. Darmois. Sur les lois de probabilités à estimation exhaustive. In *Comptes Rendus des Séances Hebdomadaires de l'Académie des Sciences*, vol. 200, p. 1265–1266. Gauthier-Villars, Paris, France, 1935. (Cité à la page 7.)

I. Deliège. *Understanding musical structure and form : papers in honour of Irène Deliège*. *Musical Scientiae Special Issue*, ESCOM, 2010. (Cité à la page 2.)

A. Dessein. *Computational Methods of Information Geometry with Real-Time Applications in Audio Signal Processing*. Thèse de doctorat, U. Pierre et Marie Curie, décembre 2012. (Cité aux pages 11, 13 et 32.)

A. Dessein, A. Cont. An Information-Geometric Approach to Real-Time Audio Segmentation. *IEEE Signal Processing Letters*, 20(4), p. 331–334, 2013. (Cité à la page 24.)

hal-00849736, version 1 - 17 Sep 2013

J. S. Downie. Music information retrieval (Chapter 7). *Annual Review of Information Science and Technology* 37, p. 295–340. (Cité à la page 3.)

J.-P. Eckmann, S. Oliffson Kamphorst, D. Ruelle. Recurrence Plots of Dynamical Systems. *Europhys. Lett.*, 4(9), p. 973–977, novembre 1987. (Cité à la page 9.)

S. Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, 11(3), p. 793–803, septembre 1983. (Cité à la page 6.)

R. A. Fisher. Two new properties of mathematical likelihood. *Proc. Royal Society of London : Series A, Containing Papers of a Mathematical and Physical Character* 144(852), p. 285–307, mars 1934. (Cité à la page 7.)

J. T. Foote. Visualizing music and audio using self-similarity. In *7th ACM Int. Conf. on Multimedia*, volume 1, pages 77–80, Orlando, FL, États-Unis, octobre-novembre 1999. (Cité à la page 9.)

J. T. Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, vol. 1, p. 452–455, New York, NY, États-Unis, juillet-août 2000. (Cité aux pages 10, 51 et 57.)

M. A. Girschick et H. Rubin. A Bayes approach to a quality control model. *The Annals of Mathematical Statistics*, 23(1), p. 114–125, mars 1952. (Cité à la page 13.)

H. Gish, M.-H. Siu, R. Rohlicek. Segregation of speakers for speech recognition and speaker dentification. In *1991 Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, p. 873–876, may 1991. (Cité à la page 8.)

M. Goto, H. Hashiguchi, T. Nishimura, R. Oka. RWC Music Database : Popular, Classical, and Jazz Music Databases. *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, p. 287–288, octobre 2002. (Cité à la page 25.)

F. Kaiser, G. Peeters. Multiple hypotheses at multiple scales for audio novelty computation within music. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, mai 2013. (Cité à la page 11.)

B. O. Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3), p. 399–409, mai 1936. (Cité à la page 7.)

G. Lorden. Procedures for reacting to a change in distributions. *The Annals of Mathematical Statistics*, 42(6), p. 1897–1908, décembre 1971. (Cité aux pages 13 et 21.)

hal-00849736, version 1 - 17 Sep 2013

J. F. Lynch Jr., J. G. Josenhans, R. E. Crochiere. Speech/silence segmentation for real-time coding via rule based adaptive endpoint detection. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 12, p. 1348–1351, Dallas, TX, États-Unis, avril 1987. (Cité à la page 32.)

S. Mallat. *A wavelet tour of signal processing*, 3rd edition. Academic Press, Waltham, MA, États-Unis, 2009. (Cité aux pages 53 et 54.)

P. Manoury. Considérations (toujours actuelles) sur l'état de la musique en temps réel. *L'étincelle* 3, novembre 2007. (Cité à la page 5.)

F. Nielsen. Cramér-Rao Lower Bound and Information Geometry. *Infinity II : On the work of Indian mathematicians*, R. Bhatia et C.S. Rajan, éditeurs, Hindustan Book Agency, 2013. (Cité à la page 39.)

F. Nielsen, R. Nock. Sided and symmetrized Bregman centroids. *IEEE Trans. Information Theory*, 55(6), p. 2882–2904. (Cité à la page 36.)

R. Nock, F. Nielsen. Fitting the Smallest Enclosing Bregman Ball. In *16th Eur. Conf. on Machine Learning (ECML)*, p. 649–656, Porto, Portugal, octobre 2005. (Cité à la page 11.)

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2), p.100–115, juin 1954. (Cité à la page 13.)

J. Paulus, M. Müller, A. Klapuri. Audio-based music structure analysis. In *11th Int. Soc. for Music Information Retrieval Conference* (ISMIR), pages 625–636, Utrecht, Pays-Bas, aout 2010. (Cité à la page 11.)

G. Peeters, A. Laburthe, X. Rodet. Toward automatic music audio summary generation from signal analysis. In *2nd Int. Soc. for Music Information Retrieval (ISMIR)*, p. 94-100, octobre 2002. (Cité à la page 4.)

G. Peeters. Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation : "Sequence" and "State" Approach. In *Proceedings of the Int. Symp. on Computer Music Modeling and Retrieval (CMMR)*, Montpellier, France, p. 143–166, mai 2003. (Cité aux pages 9 et 56.)

G. Peeters. Music Structure Discovery : measuring the "state-ness" of times. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Miami, FL, États-Unis, octobre 2011. (Cité à la page 10.)

G. Peeters, F. Cornu, D. Tardieu, C. Charbuillet, J. J. Burred, M. Ramona, M. Vian, V. Botherel, J.-B. Rault, J.-P. Cabanal. A multimedia search and navigation prototype, including music and video-clips. In *Proc. Int. Soc. for Music Information Retrieval*, Porto, Portugal, October 2012. (Cité à la page 61.)

hal-00849736, version 1 - 17 Sep 2013

G. Peeters. Indexation automatique de contenus audio musicaux. Mémoire d'habilitation à diriger des recherches, U. Pierre et Marie Curie, avril 2013. (Cité à la page 10.)

E. J. G. Pitman. Sufficient statistics and intrinsic accuracy. *Mathematical Proc. Cambridge Philosophical Soc.*, 32(4), p. 567–579, décembre 1936. (Cité à la page 7.)

A. S. Polunchenko, A. G. Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and Computing in Applied Probability*, 14(3), p. 649–684, septembre 2012. (Cité à la page 13.)

V. H. Poor, O. Hadjiliadis. *Quickest Detection*. Cambridge University Press, New York, NY, États-Unis, 2009. (Cité à la page 27.)

S. Quackenbush, A. Lindsay. Overview of MPEG-7 audio. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), p. 725–729, 2001. (Cité à la page 4.)

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Mathematical Soc.*, 37(3), p. 81–91, 1945. (Cité aux pages 6 et 39.)

R. T. Rockafellar. *Convex Analysis*. *Princeton Mathematical Series*, vol. 21. P. A. Griffiths, M. Morse, E. M. Stein, éditeurs. Princeton University Press, Princeton, NJ, États-Unis, 1970. (Cité aux pages 30 et 42.)

J. Serrà, M. Müller, P. Grosche, J.L. Arcos. Unsupervised Detection of Music Boundaries by Time Series Structure Features. In *Proc. Int. Conf. of the Association for the Advancement of Artificial Intelligence (AAAI)*, p. 1613–1619, juillet 2012. (Cité à la page 11.)

E. Scheirer, M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *IEEE Transactions on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1331–1334, 1997. (Cité à la page 8.)

K. Schutte. MIDI MATLAB Toolbox. Voir `www.kenschutte.com/midi`. Version 12, 24 janvier 2006. (Cité à la page 49.)

G. Tzanetakis, P. Cook. Multifeature audio segmentation for audio browsing and annotation. In *Proc. of 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 103–106, New Platz, NY, États-Unis, octobre 1999. (Cité à la page 8.)

B. Vercoe. Synthetic Listeners and Synthetic Performers. In *Int. Symp. "Computer World '90"*, Kobe, Japon, novembre 1990. (Cité à la page 2.)

hal-00849736, version 1 - 17 Sep 2013

H. Vinet. The Representation Levels of Music Information. In *Int. Symp. on Computer Music Modeling and Retrieval (CMMR)*, p. 193–209, Montpellier, France, mai 2003. (Cité à la page 3.)

hal-00849736, version 1 - 17 Sep 2013