Caractérisation du rythme à partir de l'analyse du signal audio

Ugo Marchand

THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Ugo MARCHAND

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Caractérisation du rythme à partir de l'analyse du signal audio

Directeur de thèse

Geoffroy PEETERS

soutenue le 28 novembre 2016

devant le jury composé de :

M. Xavier SERRA	Rapporteur
M. Emmanuel VINCENT	Rapporteur
M. Carlos AGON	Examinateur
Mme Isabelle BLOCH	Examinatrice
M. Bertrand DAVID	Examinateur
M. Andre HOLZAPFEL	Examinateur

Remerciements

Je remercie tout d'abord Geoffroy PEETERS de m'avoir encadré pendant ces trois années, pour sa pédagogie, son excellence scientifique et pour avoir su partager son goût pour la recherche.

Je remercie ensuite l'IRCAM et plus particulièrement l'équipe analyse synthèse pour m'avoir accueilli dans leur locaux.

Merci aussi aux projets Bee Music et ABCDj qui ont financé mes travaux de recherche.

Je ne peux pas citer tous mes collègues de l'IRCAM, encore là ou déjà partis, avec qui j'ai partagé mes trois années de doctorat. Merci en particulier à Enrico, David, Alicœ qui ont été des co-bureaux fantastiques.

Je remercie les Gros qui ont été d'un soutien moral sans faille et pour leur aimable contribution à la détente et à la procrastination.

Merci à mes parents pour leur soutien pendant toutes ces longues années d'études (c'est enfin fini, promis j'arrête là !).

Merci enfin à Eve pour tout, et en particulier pour son soutien (surtout à la fin) et ses très nombreuses relectures de manuscrit.

Table des matières

Ugo MARCHAND 28 novembre 2016

Caractérisation du rythme à partir de l'analyse du signal audio

Résumé : Cette thèse s'inscrit dans le cadre de l'analyse automatique de la musique. La finalité de œ champ de recherche est d'extraire des informations de la musique, ou autrement dit, de faire comprendre ce qu'est la musique à un ordinateur. Les applications sont nombreuses: fabriquer des systèmes de recommandation musicale, transcrire une partition à partir du signal ou générer automatiquement de la musique. Nous nous intéressons dans ce manuscrit à l'analyse automatique du rythme. Notre objectif est de proposer de nouvelles descriptions du rythme qui s'inspirent d'études perceptives et neurologiques.

La représentation du rythme d'un signal musical audio est un problème complexe. Il ne s'agit pas simplement de détecter la position des attaques et la durée des notes comme sur une partition mais plus généralement de modéliser l'interaction temporelle entre les différents instruments présents et collaborant à l'établissement d'un rythme de manière compacte, discriminante et invariante. Nous cherchons à obtenir des représentations invariantes à certains paramètres (tels la position dans le temps, les variations faibles de tempo ou d'instrumentation) mais à l'inverse sensibles à d'autres (comme le motif rythmique, les paramètres fins d'interprétation ou le swing). Nous étudions les trois aspects fondamentaux pour la description du rythme: le tempo les déviations et les motifs rythmiques.

Avant de proposer nos différentes méthodes d'analyse du rythme, nous décrivons les différents corpus que nous allons utiliser. Nous présentons en particulier deux corpus que nous avons spécialement créés pour nos travaux d'analyse des motifs rythmiques et du swing.

Dans un premier chapitre, nous étudions la prédiction de l'accord entre auditeurs sur la perœption du tempo. En effet, la perœption du tempo peut ne pas être partagée pour un même morceau, les auditeurs pouvant se focaliser sur différents niveaux métriques. Cette perception dépend bien évidemment de l'auditeur mais nous faisons l'hypothèse que certaines caractéristiques du signal audio peuvent faciliter ou non la perception partagée du tempo. Pour étudier cela, nous avons utilisé plusieurs descripteurs (onset, similarité à court-terme, balance spectrale et changement d'harmonicité) pour modéliser les différentes caractéristiques du signal audio. Nous avons ensuite proposé quatre modèles de prédiction de l'accord/désaccord entre auditeurs sur base de œs descripteurs. Notre meilleure prédiction 75% repose sur un modèle utilisant la cohérence des quatre tempos estimés individuellement à l'aide des quatre descripteurs. Ce modèle valide notre hypothèse de départ: si l'information de tempo est partagée entre les descripteurs, les utilisateurs auront tendance à être d'accords sur la perception du tempo.

Dans un second chapitre, nous avons étudions les déviations systématiques de position des évènements dans la musique jazz: le swing. Nous proposons plusieurs méthodes d'estimation de ces déviations systématiques, toutes basées sur l'auto-corrélation de la fonction d'onset du signal. Cette fonction d'auto-corrélation est très utile pour l'estimation du swing (ou plus généralement de la métrique) car elle possède un pic pour chaque niveau métrique important. La première méthode, notée ACF, compare l'auto-corrélation d'un morceau à une base de prototypes des couples swing/tempo. La seconde, notée LLACF, fait de même mais en utilisant l'auto-corrélation en échelle logarithmique de décalages. La dernière, notée PIC, modélise les pics de la fonction d'auto-corrélation correspondants

aux différents niveaux métriques. Nous comparons ces trois méthodes pour une tâche de classification de morceaux en *Swing*/*NoSwing* (selon que le morceau possède ou non du swing) et pour une tâche d'estimation du ratio de swing. Nous montrons en particulier que notre méthode PIC donne de meilleurs résultats que l'état de l'art (LLACF [Dittmar et al., 2015]) en détection de swing et en estimation de son ratio.

Dans un troisième chapitre, nous étudions la représentation des motifs rythmiques. Une telle représentation se doit de satisfaire deux contraintes: être invariante aux décalages temporels et être invariante aux changements de tempo. Pour cela, nous nous appuyons sur l'invariance aux décalages temporels fournie par l'auto-corrélation ou le module de la DFT et la quasi-invariance aux changements de tempo fournie par la transformée d'échelle. Nous proposons le Modulation Scale Spectrum (MSS) comme l'application de la transformée d'échelle sur différentes bandes de fréquence. Nous montrons que notre Modulation Scale Spectrum (MSS) donne de meilleurs résultats que l'état de l'art [Holzapfel et al., 2011] sur les deux corpus de référence BALLROOM et CRETE. Ceci démontre que la prise en compte de la localisation fréquentielle des évènements est importante pour la description des motifs rythmiques. Le MSS modélise cependant les différentes bandes de fréquences de manière indépendante. Nous proposons ensuite le 2DMSS et le MASSS qui modélisent les inter-relations entre les bandes de fréquence par utilisation d'une transformée de Fourier 2D suivie d'une transformée d'échelle 2D (2DMSS) et par la fusion tardive du MSS et de coefficients de corrélations croisées entre les bandes de fréquences inspirés d'expériences perceptives (MASSS). Nous montrons que notre MASSS fourni les meilleurs résultats à l'heure actuelle et ce quel que soit le corpus. Ceci démontre que la prise en compte non seulement de la localisation fréquentielle des évènements mais également de leurs inter-relations est importante pour la description des motifs rythmiques.

Mots clés : MIR, rythme, tempo perceptif, ratio de swing, motifs rythmiques, perception

Acronymes

DAT Théorie de l'Attention Dynamique ou « Dynamic Attending Theory ». 7, 10–12

DTW Déformation Temporelle Dynamique ou « Dynamic Time Warping ». 24, 25

GMM Modèle de Mélange de Gaussiennes ou « Gaussian Mixture Model ». 55, 56, 118

GTTM Théorie Générative de la Musique Tonale ou « Generative Theory of Tonal Music ». 7, 10, 12, 13, 15

KNN algorithme des k plus proches voisins ou « K-Nearest Neighbors algorithm ». 94, 96–100, 117

MFCC « Mel-frequency cepstral coefficients ». 16, 24, 25, 50, 105

MIR « Music Information Retrieval ». 1, 94, 113

MIREX « Music Information Retrieval Evaluation eXchange ». 21

MSS Modulation Scale Spectrum. x, 2, 3, 91, 95, 96, 98, 101, 102, 106, 107, 109, 110, 113

PCA Analyse en Composantes Principales ou « Principal Component Analysis ». 54

SVM Machines à Vecteurs de Support ou « Support Vector Machine ». 51, 56, 78, 94, 96–98, 100, 117, 118

Chapitre 1

Introduction

Cette thèse s'inscrit dans le cadre de l'analyse automatique de la musique ou « Music Information Retrieval » (MIR). La finalité de ce champ de recherche est d'extraire des informations de la musique, ou autrement dit, de faire comprendre ce qu'est la musique à un ordinateur. Les applications sont nombreuses : fabriquer des systèmes de recommandation musicale, transcrire une partition à partir du signal ou générer automatiquement de la musique. Nous nous intéressons dans ce manuscrit à l'analyse automatique du rythme.

Le rythme est une composante essentielle de la musique et possède un lien assez particulier avec notre espèce et notre corps. En effet, nous nous différencions des autres espèces entre autre par notre capacité à percevoir le rythme. Le rythme est aussi intimement lié au mouvement : tout le monde est capable de frapper des mains sur les temps ou de danser sur un morceau de musique. Le rythme est enfin une composante assez universelle de la musique, contrairement à la tonalité par exemple. Certaines musiques ne sont même composées que de rythmes, comme de nombreuses musiques traditionnelles africaines ou la batucada brésilienne.

L'objectif de cette thèse est d'étudier l'estimation automatique des paramètres relatifs au rythme à partir de l'analyse d'un signal musical. Si l'attention de la communauté scientifique s'est portée pour l'instant essentiellement sur l'estimation automatique du tempo, de la métrique, de la position des battements et des premiers temps, la description du rythme au sens général est restée peu étudiée. Cette recherche donc dans un premier temps à définir les paramètres les plus pertinents pour décrire un rythme et dans un deuxième temps à étudier les meilleures représentations du contenu audio permettant l'estimation automatique de ces paramètres.

La représentation du rythme d'un signal musical audio est un problème complexe. Il ne s'agit pas simplement de détecter la position des attaques et la durée des notes comme sur une partition mais plus généralement de modéliser l'interaction temporelle entre les différents instruments présents et collaborant à l'établissement d'un rythme de manière compacte, discriminante et invariante. Nous cherchons à obtenir des représentations invariantes à certains paramètres (tels la position dans le temps, les variations faibles de tempo ou d'instrumentation) mais à l'inverse sensibles à d'autres (comme le motif rythmique, les paramètres fins d'interprétation ou le swing).

Tout au long de ce manuscrit, nous allons montrer l'importance des études perceptives pour l'analyse automatique du rythme. Notre objectif est de proposer de nouvelles descriptions du rythme qui s'inspirent d'études perceptives et neurologiques.

Nous allons étudier trois aspects fondamentaux pour la description du rythme : le tempo, les déviations et les motifs rythmiques. Chacun de ces aspects

fera l'objet d'un chapitre à part.

Résumé des chapitres

Chapitre 2 : État de l'art

Nous définissons tous les concepts associés au rythme. Nous insistons particulièrement sur les études perceptives et neurologiques liée à la compréhension de la perception du rythme afin de voir comment elles peuvent profiter à l'analyse automatique. Nous proposons enfin un état de l'art sur l'analyse automatique du rythme.

Chapitre 3 : Corpus d'évaluation

Avant de proposer nos différentes méthodes d'analyse du rythme, nous décrivons les différents corpus que nous allons utiliser. Nous présentons en particulier deux corpus que nous avons spécialement créés pour nos travaux d'analyse des motifs rythmiques et du swing.

Chapitre 4 : Tempo perceptif

Nous y décrivons le problème de l'ambiguïté de tempo. Nous proposons ensuite plusieurs systèmes de prédiction de l'accord entre auditeurs sur la perception du tempo. Notre meilleure prédiction 75% repose sur un modèle utilisant la cohérence de quatre tempos estimés individuellement à l'aide des quatre descripteurs modélisant différentes caractéristiques du signal audio. Ce modèle valide notre hypothèse de départ : si l'information de tempo est partagée entre les descripteurs, les utilisateurs auront tendance à être d'accord sur la perception du tempo.

Chapitre 5 : Estimation des déviations systématiques

Nous présentons trois méthodes temporelles d'estimation d'un type de déviations systématiques de position des évènements dans la musique jazz : le swing. Les trois méthodes reposent sur la fonction d'auto-corrélation qui permet de mettre en évidence les différents niveaux métriques d'un signal audio. Nous montrons que notre méthode PIC donne de meilleurs résultats que l'état de l'art sur les deux tâches de détection du swing et d'estimation de son ratio.

Chapitre 6 : Estimation des motifs rythmiques

Dans ce chapitre, nous étudions la représentation des motifs rythmiques. Une telle représentation se doit de satisfaire deux contraintes : être invariante aux décalages temporels et être invariante aux changements de tempo. Pour cela, nous nous appuyons sur l'invariance aux décalages temporels fournie par l'auto-corrélation ou le module de la DFT et la quasi-invariance aux changements de tempo fournie par la transformée d'échelle. Nous proposons le Modulation Scale Spectrum (MSS) comme l'application de la transformée d'échelle sur différentes bandes de fréquence. Nous montrons que notre MSS donne de meilleurs résultats que l'état de l'art sur les deux corpus de référence. Ceci démontre que la prise en compte de la localisation fréquentielle des évènements est importante pour la description des motifs rythmiques.

Le MSS modélise cependant les différentes bandes de fréquences de manière indépendante. Nous proposons ensuite le 2DMSS et le MASSS qui permettent de modéliser les inter-relations entre les bandes de fréquence. Le 2DMMS utilise une transformée de Fourier 2D suivie d'une transformée d'échelle 2D et le MASSS la fusion tardive du MSS et de coefficients de corrélations croisées entre les bandes de fréquences inspirés d'expériences perceptives. Nous montrons que notre MASSS fourni les meilleurs résultats à l'heure actuelle et ce quel que soit le corpus. Ceci démontre que la prise en compte non seulement de la localisation fréquentielle des évènements mais également de leurs inter-relations est importante pour la description des motifs rythmiques.

Chapitre 2

État de l'art

Contenu

2.1 Le rythme

2.1.1 Approche historique et générale

Historiquement, une des plus anciennes définitions du rythme est due à [Platon, -350] qui le définit par « l'ordre dans le mouvement ». Cette définition met déjà l'accent sur les deux points fondamentaux nécessaires à la définition du rythme, à savoir le temps qui passe (mouvement), et la présence de régularités (ordre). Il est à noter que « l'ordre » ainsi défini ne se limite pas à une séquence isochrone, mais peut être plus complexe en matière de durées.

D'un point de vue général, nous avons retenu deux façons complémentaires de définir le rythme qui se complètent.

Dans la première, le rythme est une répétition d'évènements identiques au cours du temps (c'est le cas du rythme des saisons ou du rythme cardiaque). Étymologiquement, rythme provient du grec ῥυθμός (« rhytmos », « any regular recurring motion ») [Henry George et al., 1999] : récurrence (ou motif) revenant de façon régulière dans le temps. Cette définition peut s'appliquer à n'importe quel phénomène ayant une fréquence ou une périodicité temporelle, pouvant aller de la microseconde au million d'années (le tic-tac d'une horloge, l'alternance des saisons, la révolution des objets stellaires...). Dans cette définition, le rythme est la répétition d'évènements identiques au cours du temps. Dans le cadre de la musique, cette définition du rythme s'appliquerait à la structure musicale : à l'échelle d'un morceau, on observe la répétition de parties quasi-identiques.

La seconde définition voit le rythme comme un enchaînement d'évènements non-isochrones mais néanmoins basés sur une régularité (horloge, métronome), c'est le cas du rythme musical. Cette définition s'applique spécifiquement aux arts (musique, danse, poésie, ...). Le rythme y est défini comme l'instant temporel des évènements (sons musicaux et silences, pas de danses, métrique d'un poème). Il faut bien noter que ces instants sont loin d'être choisis au hasard, mais sont souvent basés sur une grille temporelle régulière (le tempo dans le cadre de la musique). L'échelle des tempos possibles est beaucoup plus restreinte que dans la définition précédente : les tempos sont souvent de l'ordre de quelques Hertz (ou 120 bpm comme ou pourra le voir plus tard). Cette échelle est calquée sur la biologie du corps humain, et sur ce que qu'il est physiquement possible de percevoir. En effet, une hypothèse commune expliquant pourquoi notre préférence va vers les tempos aux alentour de 120 bpm est que ce tempo est similaire aux rythmes du corps humain (rythme cardiaque, de la respiration, de la marche ou des ondes cérébrales). C'est cette seconde définition, qui se rapproche le plus du rythme musical et qui va nous intéresser par la suite.

2.1.2 Divergences sur la définition de rythme.

Le rythme a fait l'objet de très nombreuses études ces dernières décennies. Nous donnons un aperçu de la complexité de définir ce qu'est le rythme dans cette partie. Pour [Cooper, 1963], étudier le rythme c'est étudier la musique dans son ensemble (« to study rhythm is to study all of music », p1). Pour [Randel, 1945] (p639), trouver une définition du rythme acceptable, même pour une minorité de personnes, semble une tâche impossible (« It would be a hopeless task to search for a definition of rhythm which would prove acceptable event to a small minority of musicians and writers on music »). Ce constat est repris par [Fraisse, 1974].

Définitions Les définitions du rythme sont très nombreuses. Elles peuvent être assez générales, comme pour [Honing et al., 2014], où le rythme est la façon dont les évènements sont ordonnés dans le temps (« The way events are ordered in time ») ou pour [Fraisse, 1974] (p107) « La perception du rythme est faite tout à la fois de la perception de structures et de leur répétition. ». Elles peuvent être plus spécifiquement basées sur l'organisation des évènements musicaux. Pour [Cooper, 1963], ressentir le rythme est le fait de grouper des sons séparés en motifs structurés (« to experience rhythm is to group separate sounds into structured patterns », p1). Elles peuvent découler de phénomènes perœptifs : pour [Clarke, 1999] et [Fraisse, 1974], le rythme se limite à des durées courtes. Clarke fait la différence entre les évènements court-termes (« rhythm ») et long-termes (« form »). Fraisse explique (p79) que « Plus l'intervalle entre les sons augmente, plus le nombre d'éléments que l'on peut saisir en une suite est petit », « Réciproquement, pour perœvoir plus d'éléments, il faut réduire la durée des intervalles qui les séparent ». Le rythme peut-être aussi vu d'un point de vue sensori-moteur : il peut être définit comme le côté dansant de la musique, œ qui nous donne l'envie de bouger. Plusieurs travaux confirment cette définition empirique et montrent que certaines zones motriœs du cerveau s'activent lors de la production et la perception de rythme [J. L. Chen et al., 2008 ; Grahn et al., 2007 ; Schubotz et al., 2000]. Nous n'avons cité que les définitions les plus courantes, de nombreuses autres définitions sont présentées dans les travaux de [Waadeland, 2000] (p16) ou de [Gouyon, 2005] (p8).

Grandes théories du rythme Il existe plusieurs grandes théories du rythme.

Les premières présentent la musique comme stratifiée ([Cooper, 1963] parle d'architecture du rythme (« architectonic ») et [Yeston, 1976] de stratification du rythme). Cela signifie que le rythme s'établit sur plusieurs niveaux, ou sur plusieurs échelles temporelles, et que les propriétés du rythme sont les mêmes quelque soit l'échelle. Cooper utilise des groupements basés sur des mètres poétiques (iamb, anapest, ...) pour définir des règles rythmiques. Cette théorie du rythme est donc l'art de grouper ensemble des évènements sonores.

La Théorie Générative de la Musique Tonale ou « Generative Theory of Tonal Music » (GTTM), proposée par [Lerdahl et al., 1985], permet d'expliquer comment un auditeur se construit une compréhension de la musique. Les auteurs cherchent à décrire un ensemble infini (la musique) par un ensemble fini de règles. C'est ce qu'on appelle une théorie générative. Le langage, par exemple, en est une : un certain nombre de règles (grammaire) nous permet de générer un nombre infini de phrases que nous n'avons jamais entendues avant. Les auteurs proposent donc une grammaire capable de représenter œ qu'est la musique pour un humain. La GTTM se concentre sur quatre systèmes hiérarchiques créant nos intuitions musicales : « grouping structure », « metrical structure », « time-span reduction » et « prolongationnal reduction ». Ils énoncent une série de règles permettant d'établir ces systèmes. Cette théorie a été à l'origine de très nombreux travaux tant dans le domaine de la théorie musicale que dans le domaine de la perception.

[Jones et al., 1989] proposent une théorie de l'attention et de la perception d'évènements temporels appelée Théorie de l'Attention Dynamique ou « Dynamic Attending Theory » (DAT). Elle montre que le contexte de perception d'un évènement temporel est crucial. Chaque évènement va être anticipé par l'auditeur en fonction du passé. La DAT peut être résumée simplement par : le cerveau tend à fournir un minimum d'effort. Notre attention est plus forte lorsque l'on

sait qu'un évènement sonore va arriver (par exemple, au moment de la pulsation, ou sur un point fort d'un motif rythmique), et elle baisse le reste du temps. Le cerveau cherche donc à minimiser sa durée d'attention. Nous allons être moins précis lorsqu'un évènement arrive à un moment incohérent avec les précédents. Il est à noter qu'il n'y a pas de notion d'horloge régulière inférée par l'auditeur. Cette théorie s'en abstrait car les motifs rythmiques ne sont pas toujours simples, évidents ou fixés rigidement.

Dans les neuro-sciences, la question de la perception du rythme par le cerveau est toujours en suspens. Pour l'instant, il existe deux théories principales. De très nombreuses publications sont disponibles à ces sujets, nous n'en citons que quelques-unes pour chacune des deux théories. La première est la théorie du codage d'intervalle, qui se base sur l'existence d'accumulateurs temporels neuronaux ou « chronomètres ». Dans cette théorie, le codage du temps est explicite, et la précision diminue avec la durée (selon la loi de Weber $\Delta T \sim kT$). Récemment, [Merchant et al., 2011] ont montré l'existence de plusieurs types de neurones sensibles au rythme : neurones du temps relatif, neurones du temps absolus et neurones accumulateurs qui sont ces fameux neurones chronomètres. La deuxième théorie est la théorie de l'entraînement des ondes cérébrales (ou « neural resonance theory »). Ici, le temps est codé de façon périodique, et la précision temporelle est très haute et se dégrade peu. Le cœur de cette théorie est que les rythmes et les oscillations neuronales sont couplés entre eux. Les derniers résultats de cette théorie sont présentés par [Large et al., 2009]. On peut citer, par exemple, [Fujioka et al., 2012] qui montrent que même en l'absence totale de mouvement, certaines oscillations de neurones (aux alentours de 20Hz) se synchronisent avec le rythme.

Et pour aller plus loin, [Patel, 2014] pose la question majeure du rapport entre le rythme et le langage, sachant que le langage est un point différenciant les humains des autres espèces, et qu'il est supposé que langage et perception temporelle sont étroitement liés. La perception du rythme est-elle nécessaire à la mise en place du langage ou le langage précède-t-il la perception du temps ? [Winkler et al., 2009] montrent que les nouveaux-nés ont une perception partielle du rythme innée : leur cerveau se synchronise avec les débuts de mesure. Le langage ne se développe que vers 2-3 ans et la capacité de se synchroniser sur un rythme vers 5 ans. De plus, nous n'expliquons toujours pas pourquoi certaines espèces sont capables de se synchroniser sur des temps (l'exemple du perroquet dansant sur Mickael Jackson a fait le tour du monde).

[Stober et al., 2014a,b] montrent qu'il est possible d'identifier des rythmes perçus à partir de l'EEG d'un auditeur et de réseaux de neurones convolutionels. Il va donc sûrement être possible prochainement de comprendre le fonctionnement du cerveau en ce qui concerne la perception temporelle.

2.1.3 Choix d'une définition

Nous venons de voir que la définition du rythme est loin de faire consensus. Nous allons choisir une définition du rythme et la conserver tout au long de ce manuscrit. Notre définition est présentée dans la figure 2.1. Le rythme est défini comme l'un des quatre aspects fondamentaux de la musique. La musique est composée de quatre dimensions : la mélodie (ce qui est joué), l'intensité (comment), le timbre (qui joue) et le rythme (quand). La mélodie est la succession de hauteurs de l'ensemble des notes jouées. Le timbre représente la couleur de

l'extrait musical, c'est-à-dire son instrumentation. L'intensité (ou la nuance) représente la force à laquelle les notes sont jouées. Le rythme représente donc, pour nous, toute l'information temporelle d'une pièce musicale.

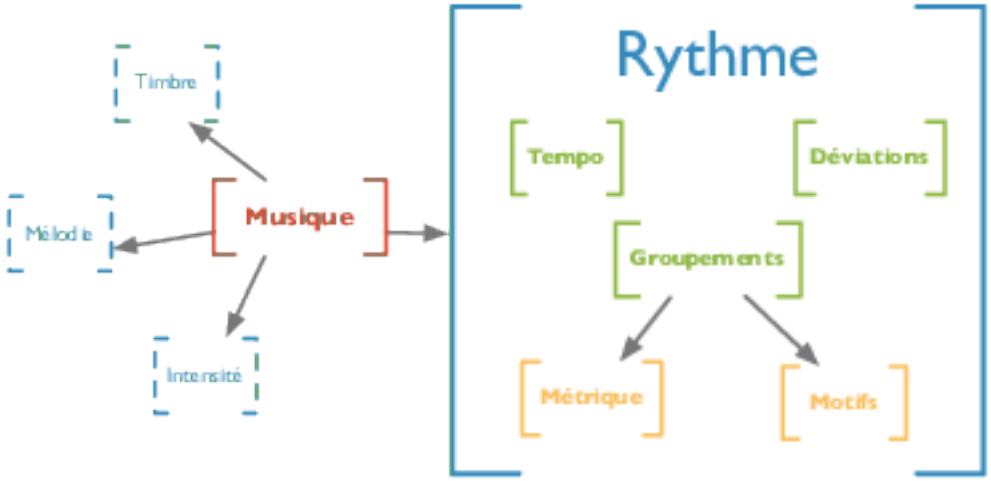

FIGURE 2.1 – *Le rythme dans la musique.*

Nous découpons le rythme similairement à [Gouyon, 2005] et à [Honing, 2001] en tempo, groupements (métrique et motifs) et déviations. Ces concepts vont être définis en détails dans la partie 2.2. Ce choix est motivé par la multiplicité des définitions du mot rythme. En effet, le mot rythme peut-être utilisé, selon les contextes, pour désigner tous les aspects temporels d'une pièce. Il peut désigner le tempo (quand on dit « cette musique a du rythme », on se réfère souvent à l'allure de la pièce), les motifs rythmiques ([Cooper, 1963] nomme ses motifs rythmiques « rhythm ») ou certaines déviations systématique (comme le swing dans la musique jazz qui fait le « rythme » d'une pièce). C'est pourquoi nous avons choisi la définition la plus générale possible.

Le rythme est donc l'ensemble des informations et des relations temporelles de la musique, il inclut tempo, métrique, motifs et déviations. Nous allons définir ces notions dans la partie suivante.

2.2 La description du temps dans la musique.

Dans cette partie, nous définissons tous les concepts associés au rythme, en insistant sur les aspects perceptifs. Nous allons donc étudier les accents, le tempo, les groupements et les déviations.

2.2.1 Accents

À la base de tout rythme, il y a des accents, c'est-à-dire des notes qui sortent du contexte de leur série de notes (« accent is a stimulus (in a serie of stimuli) which is marked for consciousness in some way. », [Cooper, 1963] p8). L'origine des accents est multiple et n'est toujours pas entièrement comprise. Pour Cooper, elle est psychologique, et peut provenir de plusieurs facteurs, comme la durée, l'intensité, le contour mélodique, la régularité, la hauteur, le timbre... La note accentuée doit cependant rester similaire à la série de son à laquelle elle se rapporte, sinon elle devient un son isolé. L'harmonie peut elle aussi jouer un rôle

[Lerdahl et al., 1985]. Cooper définit aussi le « stress », qui est une intensification dynamique d'un évènement, qu'il soit un accent ou pas (« dynamic intensification of a beat, whether accented or unaccented »). Dans la GTTM [Lerdahl et al., 1985] citent trois types d'accents :

— Le premier, est équivalent au « stress » de Cooper et Meyer. Il correspond à n'importe quel événement sur la surface musicale qui donne de l'emphase au flux musical. Cela correspond aux attaques, aux accents locaux (comme un sforzando), aux changements soudains de dynamique, de timbre, au passage de notes très hautes à très basses, aux notes longues, aux changements harmoniques. C'est œ type d'accent qui va servir d'indice à l'auditeur pour qu'il extrapole une métrique.
— Le second correspond aux accents métriques, c'est-à-dire tout temps qui est plus fort dans son contexte métrique.
— Enfin, les accents structurels, qui sont les accents causés par les points de gravité harmoniques ou mélodiques dans une phrases musicale.

Phénomène d'accentuation subjective Dans une série d'évènements isochrones et identiques, les auditeurs peuvent quand même percevoir des accents. C'est le phénomène de d'accentuation subjective (« subjective rhythmization »).

[Povel et al., 1981] étudient les accents émergeant dans une séquence de stimuli identiques en tout point (fréquence, contenu spectral, intensité, durée), le seul paramètre variant est la durée du silence inter-stimuli. Une séquence répétée de deux stimuli identiques dont la durée entre les stimuli alterne va être perçue comme un ensemble de groupes de deux notes. Une des deux notes sera perçue comme ayant un accent clair, bien que rien ne la différencie de l'autre. L'accentuation subjective se fera soit sur la première note, soit sur la seconde, en fonction de la durée relative des silences. Une autre expérience montre que le phénomène est loin d'être anodin étant donné qu'il faut augmenter l'intensité des notes non perçues comme accentuées de 4dB pour compenser l'accentuation subjective. [Parncutt, 1994] a demandé à des auditeurs de battre le tempo de différentes séquences de rythme. Il obtient des résultats variables sur les vitesses et les emplacements des accents, même quand les sons sont physiquement tous identiques et isochrones, montrant que la perœption des accents est un phénomène subjectif.

Certains modèles perceptifs ont essayé d'expliquer ce phénomène d'accentuation subjective On peut citer le modèle « d'horloge interne » d'[Essens et al., 1985; Povel et al., 1985]. Dans plusieurs expériences, ceux-ci montrent que la perception et la reproduction de rythme est meilleure et plus précise quand les rythmes sont composés de multiples entiers d'une pulsation de base appelée horloge interne. Cette horloge interne est inférée par l'auditeur à partir des accents d'un extrait musical, et elle est d'autant plus forte que les accents sont périodiques. [Jones et al., 1989], dans sa DAT, explique que des auditeurs à qui on présente une séquence d'évènements sonores extraient certaines régularités dès le premier évènement. Ils vont ensuite anticiper les futures propriétés de la série à partir de celles-ci. Cette anticipation va donc cibler l'attention sur certains évènements, qui seront donc perçus comme accentués même s'ils ne le sont pas.

[Brochard et al., 2003] essayent d'apporter des preuves physiologiques du phénomène d'accentuation subjective. Pour en trouver une manifestation physique, les auteurs mesurent les potentiels évoqués (ou Event-Related Potential ERP) du cerveau d'un auditeur lorsqu'on lui présente une séquence de sons isochrones et identiques. Les mesures des ERP suggèrent qu'une structure métrique

binaire est perçue par défaut. Ils indiquent aussi que le phénomène est partiellement affecté par l'expertise musicale. [Drake et al., 2000a] puis [Potter et al., 2009] confirment l'étude précédente, et montrent que notre perception du rythme est altérée par les écoutes précédentes (confirmant là aussi la DAT de [Jones et al., 1989]), et par notre expérience personnelle (à savoir si l'auditeur est musicien ou non-musicien [Drake et al., 2000b]).

2.2.2 Tempo, pulsation

Avant de parler de tempo, nous définissons brièvement ce que sont le tatum et le tactus. Le **tatum** est l'évènement le plus court de la structure métrique. Selon [Bilmes, 1993], c'est la division temporelle qui coïncide le plus fréquemment avec tous les évènements sonores (« time division that most highly coincide with all note onset »). Il est présent dans la littérature sous une multitude de noms différents : « tick » [Gouyon et al., 2002], « attack-point » [Schloss, 1985], « tatum » [Bilmes, 1993], « basic time unit » [Parncutt, 1994], « atomic beat » [Hofmann-Engl, 2002]. D'un point de vue musical, le tatum est souvent le niveau de la croche ou de la double croche. Il est à noter que dans un morceau très syncopé, le tatum peut ne pas être explicite et être seulement induit.

Le **tactus** est le nom donné à la pulsation ou au temps (« beat » en anglais). Il est défini par [Cooper, 1963] comme un stimulus parmi une série de stimuli identiques, réguliers et récurrents (« one of a series of regularly recurring, precisely equivalent stimuli »). Le tactus est le niveau métrique qui définit le tempo. Une définition commune du tempo est la vitesse à laquelle des auditeurs vont taper dans leur mains en écoutant la musique. C'est en effet du côté de la perception qu'il faut aller chercher la définition du tempo. [Scheirer, 2000] en propose deux qui conviennent : la fréquence d'une bande-son composée de clics et ajustée pour qu'elle soit perçue comme ayant la même vitesse que le stimulus (« the frequency of a click-track adjusted to have the same perceived speed as the stimulus ») ou le tempo d'un son pour une personne est juste ce qu'il pense être le bon tempo (« the tempo of a sound to a listener is just whatever the listener thinks it is »).

Le **tempo** symbolise la vitesse d'un extrait musical. Il est apparu sur les partitions au XVIII^{e} sous forme d'indication relative, et ce n'est qu'au XIX^{e} siècle, avec l'apparition du métronome, que le compositeur a pu indiquer précisément le tempo qu'il désire pour son morceau. Il est mesuré en battements par minute (bpm). Une source de confusion de la définition de tempo est qu'il est possible que le niveau métrique inscrit sur la partition ne soit pas celui sur lequel les gens vont taper dans leur main.

Ambiguïté du tempo De plus, plusieurs auditeurs peuvent entendre des tempos différents pour un même morceau. Il existe relativement peu d'articles traitant de l'ambiguïté de la perception d'un tempo, donc sur l'existence d'un tempo partagé par tout le monde ou non. Les premières études sur ce sujet ont été réalisées par [McKinney et al., 2004, 2006; Moelants et al., 2004] qui ont proposé un modèle de résonance pour expliquer l'apparition d'un tempo préférentiel. En effet, lors de plusieurs expériences perceptives où il était demandé aux sujets de taper les battements d'un extrait musical, il est apparu un tempo préférentiel, souvent centré autour de 120 bpm. Ce modèle est validé par des tests de perceptions dans leur laboratoire. Ils proposent aussi une explication basée sur les accents rythmiques lorsque les résultats s'éloignent de leur modèle de résonance. [Moelants et al., 2004] vont plus loin et font l'hypothèse que la perception

du tempo est partagée (c'est-à-dire que tous les sujets battent le même tempo) si l'extrait musical contient un niveau métrique proche de ce tempo de résonance (120 bpm). Par contre, si l'extrait contient un niveau métrique ayant deux pics de part et d'autre du tempo préférentiel, la perception du tempo risque d'être ambiguë (c'est-à-dire que les sujets ne sont pas d'accord entre eux). L'hypothèse d'un tempo de résonance est confirmée par les travaux de [Drake et al., 2001] qui montrent qu'il existe une zone temporelle de traitement optimal pour l'humain aux alentours de 600 ms (ce qui correspond à un tempo de 100 bpm).

Une des rares autres études traitant du désaccord sur la perception du tempo entre utilisateurs a été réalisée par [Zapata et al., 2012]. Ils montrent que l'on peut prédire la confiance à apporter aux résultats des algorithmes d'estimation automatique de tempo, grâce à une mesure de l'accord entre annotateurs (MMA « mean mutual agreemeent »). Cette mesure (MMA) permet aussi la sélection de l'annotateur de tempo le plus fiable pour un extrait musical donné. Les résultats de cette expérience montrent que les corpus classiques sur lesquels travaillent les algorithmes d'estimation de tempo sont souvent biaisés et fournissent trop d'exemples faciles. D'après leur article précédent [Holzapfel et al., 2012], ce biais est à l'origine du plafonnement des algorithmes d'estimation automatique du tempo.

En ce qui concerne l'estimation automatique du tempo perceptif, peu de méthodes existent. Les algorithmes actuels souffrent des fameuses erreurs d'octave (l'algorithme estime le double ou la moitié du tempo), l'objectif principal de la recherche dans ce domaine est donc de réduire ces erreurs d'octaves. Nous verrons ces méthodes dans le chapitre 4 dédié à l'estimation du tempo perceptif.

2.2.3 Groupements

Métrique et motifs rythmiques La GTTM de [Lerdahl et al., 1985] différencie deux types de groupements des évènements sonores : les motifs rythmiques « grouping structure » et la métrique « metric structure ». La **métrique** y est définie comme un motif régulier de temps forts et faibles. Pour [Cooper, 1963], ce sont les évènements sonores accentués qui permettent d'inférer la métrique. Des accents conflictuels ou irréguliers vont donner une notion floue ou ambiguë de la métrique, alors que des accents en phase vont donner une métrique ferme. Si une métrique est établie, il faudra des éléments contradictoires important pour en changer (ce résultat s'accorde avec la DAT proposée par [Jones et al., 1989]). Les accents métriques sont donc purement subjectifs. Pour [Honing et al., 2014], la métrique représente les multiples niveaux de régularités d'un rythme musical, qui ensemble, créent un motif hiérarchique prédominant (« multiple levels of regularity in a musical rhythm, which together create a hierarchical pattern of saliency »). [Lerdahl et al., 1985] font bien la distinction entre la métrique qui est induite par les évènements accentués et les motifs rythmiques. Les **motifs rythmiques**, (qui sont appelés rythme dans [Cooper, 1963]) sont la façon dont sont groupés les évènements non-accentués avec les évènements accentués. Pour la GTTM, un auditeur organise naturellement ce qu'il entend en phrases, mesures, périodes, ... que l'on appelle génériquement groupes. Le musicien va chercher à respirer entre ces groupes plutôt que pendant.

Niveaux hiérarchiques Pour [Cooper, 1963], [Yeston, 1976] ainsi que [Clarke, 1999], la musique est stratifiée. Cela signifie que le rythme s'établit sur plusieurs niveaux (ou sur plusieurs échelles temporelles). Cooper parle d'architecture du

rythme (« architectonic ») et Yeston de stratification du rythme. Cette stratification du rythme a pour conséquence que tous les niveaux ou toutes les échelles temporelles doivent s'analyser de la même façon. Une règle théorique définissant comment les groupements rythmiques se font doit être valable quel que soit le niveau d'analyse auquel on se place. En général, il y a 5 ou 6 niveaux hiérarchiques (ou métriques) dans une pièce, la métrique annotée (barre de mesure)[1] étant souvent le niveau intermédiaire. Tout le monde n'entend pas pour autant tous les niveaux métriques de façon équivalente. Un auditeur a tendance à se concentrer sur un ou deux niveaux principaux qui sont de l'ordre de la vitesse de battement du pied.

Règles théoriques de groupement [Cooper, 1963] et [Lerdahl et al., 1985] proposent des règles permettant de regrouper les évènements sonores entre eux afin de former la métrique et les motifs rythmiques. Dans la GTTM, les règles principales de formation de la structure métrique sont : chaque accent présent dans un niveau métrique doit l'être aussi dans tous les niveaux inférieurs, chaque niveau métrique doit être constitué d'accents équidistants et les divisions sont de deux ou trois.

En ce qui concerne les règles de groupement en motifs rythmiques, elle sont présentées sous la forme de règle de bonne formation de rythmes. En pratique, elles ne sont pas suffisantes pour faire les choix nécessaires à la formation les rythmes à partir des évènements sonores. Pour Cooper, créer les rythmes est un exercice abstrait et non concret, il n'y a pas de règles précises et rapides pour trouver quel est le rythme dans un extrait particulier (« rythmic grouping is a mental fact, not a physical one. There are no hard and fast rules for calculating what in any particular instance the grouping is »). Cet exercice fait appel à un certain nombre de connaissances, à de l'intuition, à de l'expérience et à de la sensibilité. C'est donc tout un art.

De façon générale, la similarité entre évènements forme la cohésion (donc le rythme) et la répétition créé la séparation entre groupes. Deux sons très similaires ne se sont différenciés que par un aspect perceptif (durée, mélodie, instrumentation, ornementations...) Il serait possible de faire correspondre tous les niveaux perceptifs afin de créer des rythmes non-ambigus. Cependant, cela supprimerait les niveaux supérieurs d'organisation (phrases, structure...).

Importance des ratios entiers Dans la musique occidentale, la partition est quantifiée. Les durées des évènements sonores sont des multiples ou des divisions entières de la pulsation de base. L'objectif de la quantification rythmique automatique est de faire correspondre une liste d'évènements sonores à une partition dans la notation occidentale (mesure, rythmes, ...) où les durées choisies (1, 1/4, 1/6, ...) le sont par rapport à une signature temporelle de mesure (4/4, 6/8, ...) et un tempo.

[Essens et al., 1985] travaillent aussi sur la métrique. À partir de plusieurs expériences de production ou de reproduction de motifs rythmiques, ils montrent que, spontanément, les sujets vont produire des intervalles inter-groupes en moyenne deux fois plus longs que les intervalles intra-groupes. Ce résultat est

1. La barre de mesure peut être un bon indice de la métrique dans la musique occidentale. Il faut cependant faire attention car elle est parfois utilisée de façon un peu arbitraire, et certains compositeurs comptaient sur l'interprète pour donner la sensation de la bonne métrique et n'indiquait donc pas celle-ci. Ce sont même parfois les éditeurs de musique qui ont ajouté plus ou moins arbitrairement les barres de mesure à la musique qu'ils imprimaient.

indépendant du motif rythmique à reproduire et du tempo. Ils montrent aussi qu'il n'y a pas l'air d'avoir de distinction entre les intervalles avec ratio entier (2 :1, 3 :1, 4 :1) et les intervalles avec ratio non-entier (1.5 :1, 2.5 :1, 3.5 :1) L'intervalle le mieux reproduit reste le 2 :1. Le ratio 1.5 :1 est sur-évalué, et les ratios supérieurs à 2 :1 sont sous-évalués. Enfin, les motifs rythmiques qui peuvent être interprétés par la métrique sont plus facilement reproduits que ceux qui ne le peuvent pas. Le cas du ratio 2 :1 fait exception, dans ce cas-là, tous les motifs sont bien reproduits même si ceux-ci ne sont pas interprétables par la métrique. Il y a donc une place toute particulière du ratio 2 :1 dans la perception et la production de rythme.

Cette observation perceptive est confirmée par des expériences neurologiques de [Abecasis et al., 2005] sur l'accentuation subjective. L'analyse des ERP du cerveau d'auditeurs pendant l'écoute de motifs rythmiques (long-court ou long-court-court) confirment que le groupage est binaire par défaut, et non ternaire.

[Patel et al., 2005] cherchent à répondre entre autre à la question : comment la synchronisation avec une structure métrique diffère de la synchronisation avec une périodicité ? Ils réalisent une expérience où les sujets doivent battre la pulsation sur des rythmes ayant une métrique plus ou moins présente. Certains rythmes sont très métriques, c'est-à-dire qu'ils ont un évènement sur tous les temps ayant une importance métrique comme la pulsation. D'autres le sont beaucoup moins. Toutes les séquences sont basées sur des rythmes tirés des expériences précédentes de [Essens et al., 1985]. Ils montrent que la précision de frappe est la même si les gens se synchronisent sur un métronome ou sur un rythme très métrique, ce qui est contradictoire avec les résultats de [Large et al., 1999] qui montraient une meilleure synchronisation grâce à des rythmes par rapport à un métronome seul. Large et Jones montrent aussi que la précision de la frappe est moins bonne sur les rythmes non-métriques, validant par là les expériences d'Essens.

Deux autres expériences de [Essens, 1986] montrent que les motifs temporels sont représentés hiérarchiquement, et que les hauts niveaux sont liés aux bas niveaux par des intervalles entiers égaux à 2 ou 3. Pour avoir une reproduction précise des motifs rythmiques, les expériences montrent que les unités d'horloge doivent être divisées en parties qui forment des ratios entiers. La reproduction de motifs temporels dont les ratios ne sont pas entiers est moins précise, de même que celle dont les motifs qui ont des ratios supérieurs à 5. La reproduction d'intervalles est précise pour les ratios 2, 3 et 4. Les divisions en 2 et en 4 n'interfèrent pas entre elles, suggérant qu'elles sont liées. Cette étude soutient l'idée d'un modèle d'une représentation interne des motifs rythmiques dans lesquels l'horloge perçue est subdivisée en deux ou trois parties, elles-mêmes divisées en deux ou trois. Ce modèle est confirmé par des observation neurologiques par [Sakai et al., 1999] qui ont montré que la représentation neuronale d'un rythme (fMRI) dépend de si le rythme est métrique ou non-métrique. Pour les ratios 1 :2 :4 et 1 :2 :3, une partie de l'hémisphère gauche est principalement activée, alors que pour les ratios 1 :2.5 :3.5, c'est une partie de l'hémisphère droit qui est principalement activée. Cela conforte les expériences psychologiques d'Essens, montrant qu'il y a une différence entre les motifs rythmiques basés sur une métrique à ratio entier, et sur une métrique à ratio non-entiers.

[Collier et al., 1995] ont de plus mis en évidence le phénomène de régularisation des rythmes. Quand il a été demandé à des personnes de jouer des rythmes

ayant des rapports d'intervalles temporels non-entiers, œs personnes ont eu tendance à régulariser les ratios non-entiers : par exemple le ratio 2.5 :1 est devenu 2 :1, le ratio 3 :5 est devenu 3 :1. [Large et al., 2002] et [Madison et al., 2002] confirment ces résultats, en montrant que les auditeurs ont tendance à percevoir des temps non-réguliers comme réguliers. Large et al. nomment œ phénomène régularisation subjective (« subjective rhythmisation »).

2.2.4 Déviations

Sur une partition où la musique est quantifié, les déviations temporelles expressives des notes ne sont pas indiquées quantitativement mais sous forme d'indication (rubato, ral., ...). La GTTM, qui est une représentation abstraite de la partition, possède le défaut de ne pas prendre en compte les déviations expressives ni systématiques. Une déviation temporelle est définie comme l'écart entre la position temporelle théorique d'une note selon la quantification rythmique et sa position réellement jouée.

Mathématiquement, il est toujours possible de représenter une déviation par un changement local de tempo, et inversement [Honing, 2001]. Musicalement, cependant, modifier le tempo d'une pièce, ou utiliser des déviations temporelles pour accentuer l'expressivité sont bien différents. En général, l'interprétation la plus probable est : le tempo est constant et le morceau possède des déviations temporelles locales. Prenons l'exemple d'une musique jazz ayant du swing : il est plus probable que tout le monde partage le même tempo constant, et que certains musiciens jouent hors des temps, plutôt que d'avoir un tempo variable pour chaque musicien qui jouerait alors sur ses propres temps. Dans la pratique, la courbe de tempo n'est jamais constante, comme l'ont fait remarqué [Dixon et al., 2006] (« the regular pulse, which is the basis of rhythmic notation in common music notation, is anything but regular when the timing of performed notes is measured. »). Ils ont aussi montrés que les participants ont tendance à lisser la courbe de tempo, et donc que les instants des temps perçus sont un compromis entre pulsation parfaitement régulière et déviations nulles.

Il existe deux types de déviations temporelles dans la musique : les déviations expressives, qui peuvent être utilisées pour créer des accents locaux, ou pour instaurer une certaine émotions dans une pièce de façon globale, et les déviations systématiques, comme le swing dans la musique jazz. [Palmer, 1997] a montré que les déviations sont loin d'être aléatoires et [Honing et al., 2008] que les musiciens jazz professionnels ont un contrôle extrêmement précis sur le facteur de swing qu'ils veulent atteindre.

2.3 Analyse automatique du rythme

2.3.1 Système général d'analyse automatique de la musique

Les deux points les plus importants d'un système d'apprentissage automatique de la musique sont l'extraction d'information significatives à partir du signal audio (extraction de descripteurs) et la modélisation efficace de la relation entre ces descripteurs et les concepts à estimer (algorithme de classification). La Figure 2.2 résume cela [2].

2. On notera l'existence de méthodes basées sur les réseaux de neurones profonds, qui permettent de réaliser ces deux étapes en une seule passe, les neurones des couches de sorties correspondants au concepts et les neurones des couches intermédiaires représentant les descripteurs.

FIGURE 2.2 – *Système d'analyse automatique de la musique. Il se compose en général d'une étape d'extraction de descripteurs à partir du signal audio, et d'une étape de modélisation des concepts haut-niveaux.*

Nous faisons remarquer que nous nous intéressons seulement aux systèmes basés sur l'analyse du signal audio, par opposition à des systèmes se basant sur une représentation MIDI ou sur une partition comme dans les travaux de [Meudic, 2004].

Descripteurs. Il est intéressant de noter qu'il existe des représentations pour la mélodie (chromas) et le timbre (« Mel-frequency cepstral coefficients » (MFCC)), mais qu'il n'y en a aucune faisant consensus pour une description « universelle » du rythme[3].

Nous ne décrirons pas ici tous les types de descripteurs existant pour l'analyse automatique de la musique. Nous verrons seulement dans la partie 2.3.2 les descripteurs associés à l'analyse du rythme.

Applications. Le cadre de l'analyse automatique du rythme comprend de nombreuses applications comme la classification en rythmes prédominants (ou classification en motifs rythmiques) [Holzapfel et al., 2009 ; Jensen et al., 2009 ; Peeters, 2005], l'estimation de tempo et le « beat-tracking » (localisation des battements) [Elowsson et al., 2015 ; Paulus et al., 2002], l'estimation des déviations systématiques [Dittmar et al., 2004 ; Gouyon et al., 2003a ; Laroche, 2001], l'estimation de la métrique [Quinton et al., 2015], l'estimation de la structure [Peeters et al., 2014a] ou encore la création de résumé audio [Antonopoulos et al., 2007 ; Bartsch et al., 2005].

2.3.2 Cas spécifique du rythme

Nous venons de voir le cadre général d'apprentissage automatique de la musique. Dans le cas spécifique de l'analyse du rythme (Figure 2.3), la première

Nous ne parlerons pas de ces méthodes dans ce manuscrit pour diverses raisons. D'abord d'un point de vue chronologique les premières études ont commencé à faire leur apparition au début de nos travaux, et le sujet était encore un peu nouveau pour le laboratoire. Les temps de calculs sont décourageants, surtout lorsque l'on ne possède pas le matériel adéquat, comme des cartes graphiques dernier cri ou une ferme de serveurs de calculs. Nous croyons de plus que des descripteurs spécifiquement créés pour une tâche peuvent largement égaler des réseaux de neurones profonds. Nous trouvons plus intéressant de travailler sur l'aspect traitement du signal, c'est-à-dire comment l'information à modéliser se traduit dans le signal audio plutôt que sur l'aspect configuration et optimisation de réseaux de neurones. Enfin, les corpus d'évaluation ne possèdent pour l'instant pas assez d'exemples pour pouvoir mettre en œuvre ce type de méthode, qui nécessitent beaucoup de données d'apprentissage.

3. Attention, nous ne prétendons pas que les Chromas et les MFCC soient suffisants pour résoudre tous les problèmes d'estimation automatique de la mélodie et du timbre, nous avançons simplement qu'ils sont très utilisés, qu'ils possèdent de nombreuses déclinaisons possibles, et qu'ils sont majoritairement utilisés pour ce type de problèmes.

étape d'extraction de descripteurs s'appuie sur l'extraction des événements temporels (partie 2.3.4), suivie d'une représentation temporelle cherchant notamment à rendre compte des périodicités et/ou des motifs rythmiques présents (partie 2.3.5). Ces étapes sont souvent précédées d'un pré-traitement facultatif visant à améliorer les performances de la fonction de détection d'onset ou à séparer le signal en bandes de fréquences (partie 2.3.3).

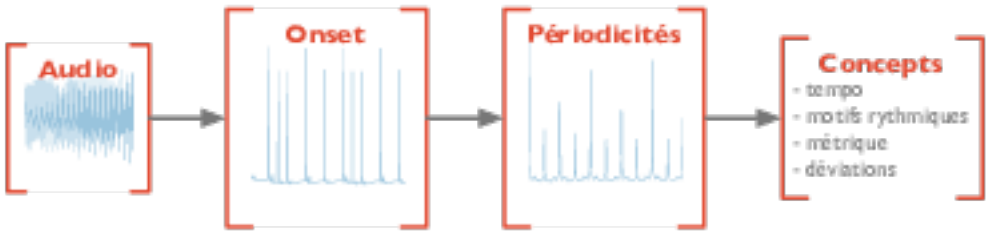

FIGURE 2.3 – *Système d'analyse automatique du rythme. Par rapport à un système d'analyse général, la première étape d'extraction de descripteur est souvent divisée en une étape d'extraction des accents (onset) et d'une étape de recherche de périodicités.*

L'estimation automatique du tempo, des motifs rythmiques ou des déviations feront tous une utilisation différente des fonctions de périodicités. Il est à noter que nous avons cherché à généraliser l'analyse du rythme, et que, même si la majorité des systèmes existants suivent la Figure 2.3, tous ne le font pas strictement.

2.3.3 Pré-traitement facultatif du signal

Une des principales raisons de pré-traiter le signal dans le cadre de l'analyse du rythme est d'améliorer l'extraction de la fonction de détection d'onset. Pour cela on peut séparer le signal en parties harmoniques et percussives ou séparer le signal en multiples bandes de fréquences.

Séparation harmonique/percussive

La séparation du signal en éléments harmoniques et percussifs est une technique pouvant être très utile à l'analyse du rythme. En effet, les évènements percussifs (ce que joue la batterie dans la musique occidentale, ou plus généralement, ce que jouent toutes les percussions) sont souvent des repères pour les autres musiciens en terme de rythme. Une séparation idéale permettrait d'avoir d'un côté tous les évènements harmoniques (très utiles dans le cadre d'un analyse automatique de la mélodie), et de l'autre tous les évènements percussifs. Nous distinguons trois types d'approches pour la séparation en sources harmoniques et percussives d'un signal audio. La première est dite aveugle, et utilise des méthodes comme la factorisation en matrices non-négative (NMF) [Helen et al., 2005], ou l'analyse en sous-espaces indépendants [Uhle et al., 2003]. La deuxième approche cherche à faire correspondre des modèles de sons de percussions (temporels [Zils et al., 2002] et spectraux [Yoshii et al., 2004]), en les cherchant et les adaptant au signal analysé (« match-and-adapt »). La dernière méthode est basée sur la création d'un modèle discriminant entre la partie percussive et la partie harmonique. [Gillet et al., 2005] considère la partie harmonique comme une somme de sinusoïdes et la partie percussive comme du bruit, tandis que les méthodes de [Fitzgerald, 2010; Ono et al., 2008a,b; Rigaud et al.,

2011] sont basées sur le fait que dans le spectrogramme d'un signal musical, les évènements percussifs seront représentés par des droites verticales, et les évènements harmoniques par des droites horizontales. Un simple filtrage médian du spectrogramme sur les fréquences permet donc de détecter les évènements percussifs, et un filtrage médian sur l'axe temporel permet d'extraire la partie harmonique.

Séparation en bandes de fréquences

Un autre pré-traitement courant consiste à séparer le signal en différentes bandes de fréquences. La Figure 2.4 montre l'importance de ce traitement. En effet les deux rythmes présentés sur cette figure sont sémantiquement différents, pourtant, sans représentation fréquentielle, il ne seraient pas différentiables : une détection basique des accents donnerait dans les deux cas, les mêmes évènements aux même moments.

FIGURE 2.4 – *Deux rythmes dont la différence est la répartition fréquentielle des évènements sonores.*

Séparer un signal en bandes de fréquences permet aussi de se rapprocher de la perception humaine, ce qui est en général est assez intéressant. En particulier cela rend l'interprétation des traitements et des résultats plus facile. [Desain et al., 1998] soulignent le manque de collaboration qu'il peut y avoir entre les approches perceptives, neurologiques et computationnelles. [Benetos et al., 2013] soulignent aussi que les algorithmes d'analyse automatique de la musique pourraient grandement bénéficier des résultats des autres champs de recherche. Dans le cadre de la détection d'onset, de nombreux travaux utilisent déjà des résultats psycho-acoustiques [Collins, 2005].

D'un point de vue signal, séparer en bandes de fréquences permet généralement de gagner en résolution. Il est possible d'analyser ce qui se passe sur chaque bande individuellement, mais aussi d'observer les corrélations ou les non-corrélations entre les bandes.

De nombreux auteurs utilisent la séparation en bandes de fréquences pour améliorer la précision de leur algorithmes. [Scheirer, 1998] est le premier à avoir souligné clairement que pour la détection d'onset, il fallait analyser le signal sur différentes bandes de fréquences et combiner œs analyses ensuite. Cette idée a été reprise, entre autres, par [Klapuri, 1999], [Paulus et al., 2002] et [Dixon et al., 2003]. Tous ces auteurs utilisent 6 à 21 bandes de fréquences espacées logarithmiquement.

D'un point de vue perceptif, il existe plusieurs échelles fréquentielles cherchant à reproduire au mieux le fonctionnement de l'appareil auditif humain (cochlée et oreille interne). Les deux principales sont l'échelle de mel et l'échelle de bark. Elle sont aussi basée sur une répartition logarithmique des bandes de

fréquences. L'échelle mel [4] a été introduite par [Stevens et al., 1937]. Une des formules permettant de convertir f Hertz en m mel est :

$$m = 2410 \log_{10}\left(1 + \frac{f}{625}\right)$$

Il s'agit de l'échelle la plus populaire, elle est utilisée dans de très nombreux travaux. L'échelle bark a été introduite par [Zwicker, 1961]. Elle est légèrement moins populaire que l'échelle de Mel. Elle a été définie par des expériences de mesures de la puissance sonore subjective (« loudness »). Elle possède 24 bandes de fréquences. Au-dessus de 500Hz, celles-ci sont espacées plus ou moins logarithmiquement, et en dessous, l'échelle devient de plus en plus linéaire. Elle a été utilisée par exemple dans [Jehan, 2004; Pampalk et al., 2002].

Enfin, [Patterson et al., 1992] ont montré que la réponse impulsionnelle de la fonction gammatone d'ordre 4 est une très bonne approximation des formes des filtres auditifs humains. La fonction gammatone est définie par :

$$g(t) = at^{n-1}\cos(2\pi ft + \phi)\exp^{-2\pi bt}$$

où n est l'ordre du filtre, b sa bande-passante, a son amplitude, f sa fréquence centrale et ϕ sa phase. Nous montrons la réponse fréquentielle de 12 filtres gammatones répartis entre 20 et 4000 Hz sur la Figure 2.5.

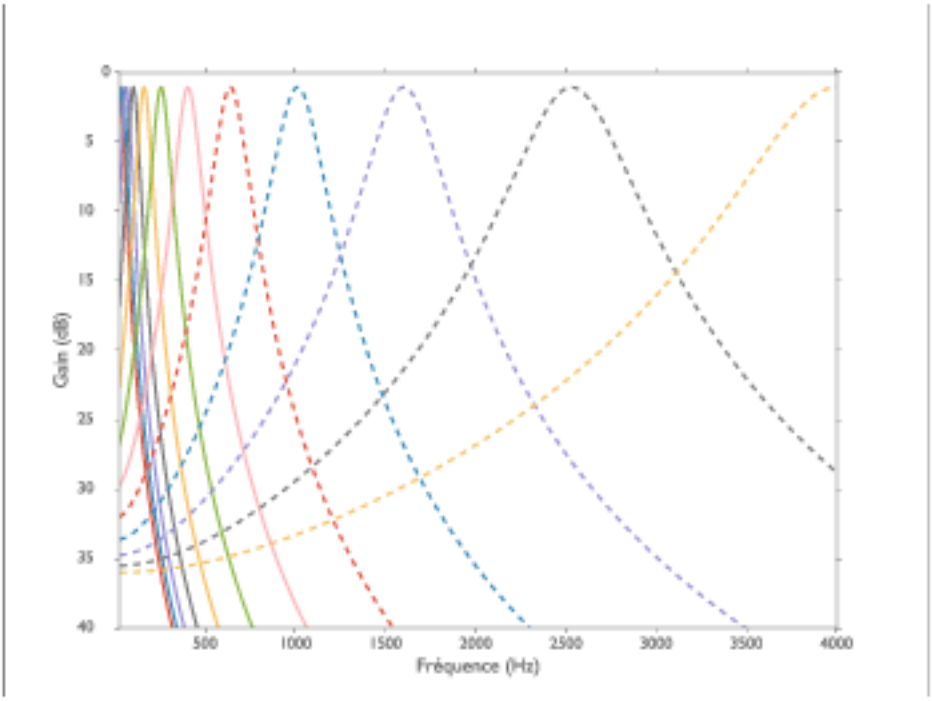

FIGURE 2.5 – *Réponse fréquentielle de 12 filtres gammatone répartis entre 20 et 4000 Hz*

2.3.4 Extraction des évènements temporels du signal

Extraire les onsets d'un signal revient à trouver où sont les accents (partie 2.2.1). De très nombreuses méthodes existent, comme en atteste une littérature

4. Le mot mel vient de mélodie.

très fournie. Les travaux de [Bello et al., 2005] et de [Böck et al., 2012b] donnent un bon aperçu de toutes les possibilités. Nous en résumons les principaux points dans cette partie.

Définition de l'onset. [Klapuri, 1999] le définit comme « le début d'évènements discrets dans un signal acoustique » (« the beginning(s) of discrete event(s) in acoustic signals »). Il peut être perçu comme un changement d'intensité, de hauteur ou de timbre. [Bello et al., 2005] le définissent comme l'instant choisi pour marquer un transient, qui est un court intervalle temporel durant lequel le signal évolue rapidement de façon non triviale. Dans la plupart des cas, l'onset est le début du transient. Un exemple dans le cas d'un signal simple est montré dans la Figure 2.6.

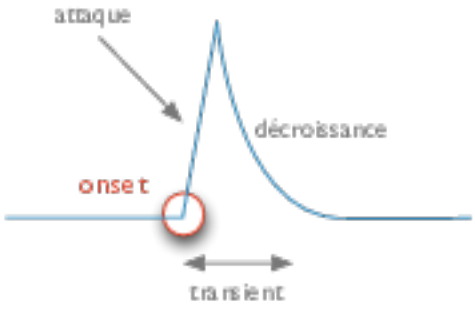

FIGURE 2.6 – *Définition d'un « onset » dans un cas de signal idéal. Figure inspirée de la Fig. 1 de [Bello et al., 2005].*

Algorithmes de détection d'onsets. Dans le cadre d'un signal musical, c'est-à-dire polyphonique, complexe et bruité, il est difficile d'extraire directement les onsets. Il est donc souvent fait appel à une représentation intermédiaire, appelée fonction de détection d'onset[5] (« onset detection function » ou ODF), qui va décrire la quantité de changement qui se produit dans le signal d'un instant à l'autre.

Trois étapes sont donc usuellement réalisées pour extraire les onsets d'un signal musical. La première consiste à pré-traiter le signal. On peut noter l'utilisation de multiples bandes de fréquences [Goto et al., 1996 ; Klapuri, 1999 ; Scheirer, 1998], d'une amplitude exprimée en logarithme [Eyben et al., 2010 ; Klapuri, 1999] ou de la méthode d'adaptative whitening [Stowell et al., 2007].

La seconde étape est le calcul d'ODF (ou l'étape de réduction dans [Bello et al., 2005]). Les méthodes de calcul d'ODF donnant de bons résultats sont par exemple le flux spectral (qui consiste à différencier temporellement l'amplitude du spectrogramme), la déviation de phase pondérée par l'amplitude ([Bello et al., 2004; Dixon, 2006]) ou l'utilisation du domaine complexe ([Duxbury et al., 2003]).

Enfin, les onsets sont extraits par recherche de pics dans l'ODF, ce qui se traduit souvent par une recherche de maxima locaux.

5. Cette fonction peut aussi être appelée "novelty function" [Foote, 2000].

Tâche d'estimation d'onsets. Une tâche d'estimation des onsets a été introduite en 2005 dans « Music Information Retrieval Evaluation eXchange » (MIREX). Chaque année, quelques participants soumettent leur algorithmes d'estimation d'onset. Les méthodes qui donnent de bons résultats sont [Röbel, 2005] [Collins, 2005] [Zhou et al., 2011] [Böck et al., 2012a,b] [Eyben et al., 2010] [Lacoste et al., 2006] et [Degara et al., 2011]

De par la forte utilisation des onsets pour des tâches d'analyse automatique de la musique, on peut trouver des implémentations de fonctions de détection d'onset prêtes à être utilisées comme [Ellis, 2007].

Offset Si l'onset peut être assimilé au début d'un évènement musical (au début d'une note par exemple), l'offset en est sa fin. Il a beaucoup moins été étudié que son homologue. Son intérêt est moindre pour l'analyse du rythme, mais il est utile pour la transcription automatique de la musique. Nous pouvons citer les travaux de [Benetos et al., 2011], qui proposent une estimation conjointe de fréquences fondamentales multiples, d'onset et d'offset.

2.3.5 Représentations temporelles

La difficulté principale pour créer un descripteur rythmique est de modéliser correctement le temps. Faire des statistiques simples telle que des moyennes sur les analyses trame à trame d'un morceau n'est pas suffisant pour modéliser correctement les interactions temporelles entre les évènements sonores. C'est pourquoi dans la littérature des méthodes sont utilisées pour chercher à mettre en avant les périodicités du signal (partie 2.3.5.1), sa vitesse de variation sur plusieurs bandes de fréquence (partie 2.3.5.2), ses similarités à court-terme (partie 2.3.5.3) ou ses motifs rythmiques (partie 2.3.5.4).

2.3.5.1 Les fonctions de périodicité

Banc de filtres résonants. Une des premières fonction de périodicité est le banc de filtres résonants proposé par [Scheirer, 1998]. Scheirer utilise ces filtres en peigne comme résonateurs et analyse les périodicités du signal préalablement séparé sur six bandes de fréquences. Il utilise 150 résonateurs espacés logarithmiquement entre 60 et 180 bpm dont il somme les sorties sur les 6 bandes de fréquences. Cette représentation est ensuite utilisée pour estimer la position des battements et le tempo. [Klapuri et al., 2006] utilisent aussi ces bancs de résonateurs pour l'estimation de métrique.

Transformée de Fourier (FFT). Un des algorithmes les plus utilisés en traitement du signal pour l'extraction périodicités ou fréquences est la transformée de Fourier. On peut appliquer cette transformée directement sur différentes bandes de fréquence du signal, comme [Pampalk et al., 2002]. On peut aussi l'appliquer sur une fonction d'onset, comme [Holzapfel et al., 2011; Peeters, 2011] ou sur n'importe quelle représentation du signal comme le Flux Spectral [Klapuri et al., 2006] ou l'amplitude du spectre de puissance [Peeters, 2007].

Fonction d'auto-corrélation (ACF). C'est aussi une des représentations les plus communes pour montrer les différentes périodicités d'un signal. Elle se calcule comme :

$$ACF(\tau) = \int_{-\infty}^{\infty} x(t).x(t-\tau)dt$$

Nous montrons un exemple avec un motif classique de batterie sur la Figure 2.7. À gauche, nous présentons le motifs rythmique qui a permis de générer le signal. À droite, nous montrons l'auto-corrélation de sa fonction de détection d'onset. Les rythmes au-dessus des pics de l'auto-corrélation représentent les durées associées à ces pics. Le premier pic à environ $0,5$s correspond à la période de durée égale à une croche, le second pic à son double (soit la durée d'une noire), et ainsi de suite. Avec cette représentation, on voit pourquoi l'auto-corrélation est particulièrement adaptée à l'estimation des différents niveaux métriques.

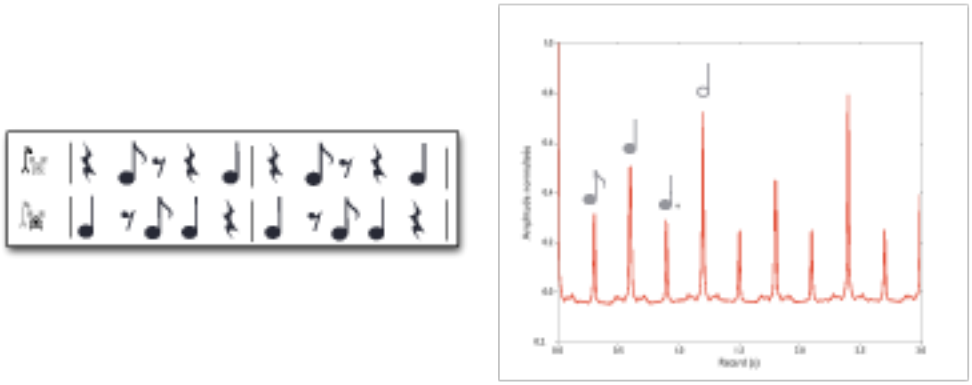

FIGURE 2.7 – *Un motif rythmique de batterie simple et son auto-corrélation. Les rythmes au-dessus de l'auto-corrélation représentent les durées associées à ces pics.*

L'auto-corrélation a été utilisée entre autres par [Goto, 2001 ; Gouyon et al., 2003b ; Peeters, 2011 ; Tzanetakis et al., 2002], et de très nombreuses variations de cette fonction sont possibles.

Une variante de l'ACF appelée « Narrowed Auto-Correlation Function » (NACF) a été introduite par [Brown et al., 1989]. Elle est calculée en une somme pondérée des auto-corrélation pour les retards multiples entiers k de τ (τ, 2τ, 3τ...) avec un poids diminuant quand k augmente. Cela permet une plus grande précision des périodicités, ce qui peut-être intéressant dans le cas où un signal possède des périodicités proches, mais ce qui n'est pas trop le cas dans le cadre de la détection du rythme où les périodicités sont relativement grandes.

Pour représenter le rythme, [Tzanetakis et al., 2002] propose un histogramme (« beat histogram ») basé sur une auto-corrélation améliorée, venant de [Tolonen et al., 2000]. Une fois la fonction d'auto-corrélation calculée, celle-ci est réduite à ses valeurs positives, puis étirée temporellement par un facteur 2, puis soustraite à elle-même. Cela permet de supprimer les effets des multiples entiers des périodicités premières, et d'ainsi améliorer la détection de pics ultérieure. L'idée peut être répétée autant de fois que nécessaire pour des périodicités supérieures (3, 4. . .).

Afin de proposer une représentation s'affranchissant en partie du tempo, [Dittmar et al., 2015 ; Eppler et al., 2014 ; Gruhne et al., 2009 ; Jensen et al., 2009 ; Völkel et al., 2010] ont proposé un ré-échantillonnage de la fonction d'auto-corrélation sur une échelle de temps logarithmique (« log-lag autocorrelation function » LLACF). Cette représentation est moins sensible aux petits changements de tempo. Corriger le tempo y est aussi plus facile, étant donné qu'un changement de tempo correspond à un simple décalage temporel du retard. Cette méthode se rapproche de la Transformée d'Échelle, que l'on verra par la suite.

Combinaison de la FFT et de l'ACF Une idée introduite par [Peeters, 2006] est de combiner les deux représentations précédentes. En effet, quand on cherche à trouver le tempo à partir des deux représentations, les problèmes principaux rencontrés sont les erreurs d'octave (le pic sélectionné pour représenter le tempo est le double ou la moitié du tempo réel). Il est intéressant de combiner l'auto-corrélation et la transformée de Fourier car les erreurs d'octave que l'on peut faire dans les deux cas évoluent de façon inverse l'une de l'autre. Notons f_0 la fréquence fondamentale associée au tempo. Avec la FFT, on risque de sélectionner des multiples de cette fréquence $kf_0, k \in Z$, alors qu'avec l'auto-corrélation, on risque de sélectionner des multiples de la période, soit $k/f_0, k \in Z$. Ainsi, lorsque l'on combine les deux fonctions, les erreurs ne se combinent pas et le pic principal est généralement celui du tempo.

Transformée d'Échelle. La transformée d'Échelle est définie par [Cohen, 1993] :

$$T_E(c) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} x(t) t^{-jc-\frac{1}{2}} dt$$

Elle possède une propriété très importante : l'invariance à l'échelle. Dans le cadre de la musique, cette invariance correspond à une quasi-invariance au tempo. Cette propriété est très intéressante pour la description du rythme. Elle a été utilisée dans le cadre de la modélisation du rythme par Holzapfel et al. [Holzapfel et al., 2009 ; Holzapfel et al., 2011]. Ils utilisent la transformée d'échelle pour modéliser l'auto-corrélation de la fonction de détection d'onset du signal. Les spectres d'échelle sont ensuite comparés entre eux au moyen d'une distance cosinusoïdale. Cette méthode donne actuellement les meilleurs résultats sur une tâche de classification de motifs rythmiques. La transformée d'échelle a aussi été utilsée par [Völkel et al., 2010] qui l'utilisent aussi pour s'affranchir du tempo. Ils proposent plusieurs représentation pour différencier des rythmes de musique latine américaine, dont l'une est basée sur la transformée d'échelle. Les résultats sont assez encourageants pour ce type de représentation. Cette transformée sera évoquée plus en détails dans la partie 6.

Autres. De très nombreuses autres représentations existent : la représentation en ondelettes [Smith et al., 2008 ; Tzanetakis et al., 2002], les histogrammes inter-onset [Gouyon et al., 2003a, 2002], la transformée à Q constant [Brown, 1991], ou le tempogramme [Cemgil et al., 2000] qui est la convolution du flux spectral avec les filtres en peigne de [Scheirer, 1998].

2.3.5.2 Modulation Spectrum

[Rodet et al., 2003 ; Worms, 1998] et Peeters proposent le Spectre de Modulation (Modulation Spectrum), sans le nommer explicitement, pour reconnaître automatiquement des extraits musicaux par leur empreinte sonore. Le spectre de modulation modélise l'évolution temporelle de contenu énergétique sur plusieurs bandes de fréquence par une transformée de Fourier. [McKinney et al., 2003] calculent le spectre de modulation sur un banc de 18 filtres Gammatone. Ils somment ensuite son énergie sur quatre bandes de fréquences, et utilise ces 18x4 descripteurs pour classifier un ensemble de données en cinq classes audio (classique, pop, parole, bruit, foule) et 7 genres musicaux. [Whitman et al., 2004] proposent les descripteurs Penny (Penny features) pour créer un système qui

comprenne les avis sur les productions musicales, et qui étiquette les chansons à partir de leur contenu audio. [Atlas et al., 2003] proposent un modèle joint de fréquence acoustique et de fréquence de modulation pour améliorer l'encodage audio. Ce modèle à 32kbit/s a été trouvé de meilleure qualité audio par des auditeurs que le format MP3 à 56 kbit/s.

2.3.5.3 Matrice similarité

La représentation par matrice de similarité a été introduite par [Foote, 1999]. Elle se calcule selon les étapes suivantes :

— l'audio est d'abord paramétré, avec une représentation spectrale [Foote et al., 2001], ou des MFCCs, ou encore des chromas [Bartsch et al., 2005]

— il en résulte une séquence de vecteurs de descripteurs dont les distances sont calculées deux à deux. Le résultat est présenté dans une matrice de similarité $S(i, j)$.

[Foote et al., 2001] proposent ensuite deux représentations dérivées de cette matrice d'auto-similarité, le « beat spectrum » et le « beat histogram ». Le « beat spectrum » (fonction du retard l) est calculé en trouvant les périodicités dans la matrice de similarité en prenant la somme diagonale correspondant au retard l :

$$B(l) = \sum_k S(k, k + l)$$

Le « beat histogram » (qui est le « beat spectrum » en fonction du temps) permet d'analyser la structure rythmique de l'extrait. Il est basé sur une mesure d'auto-corrélation de la matrice de similarité :

$$B(k, l) = \sum_{i,j} S(i, j)S(i + k, j + l)$$

Il a été ensuite proposé de multiples améliorations et utilisations de cette matrice d'auto-similarité, comme la convolution de cette matrice par un noyau de type « checkerboard kernel » [Foote, 2000] pour la segmentation de l'audio, ou le filtrage de cette matrice d'auto-similarité pour créer des résumés audio [Bartsch et al., 2005]. [Antonopoulos et al., 2007] utilisent cette représentation appliquée à l'ensemble d'un morceau ou à son résumé pour créer des « signature rythmiques ». Ces signatures sont ensuite comparées entre elles par un algorithme de Déformation Temporelle Dynamique ou « Dynamic Time Warping » (DTW) pour classer des corpus de musique traditionnelles grecques et africaines selon leur rythme prédominant.

2.3.5.4 Motifs rythmiques

Une autre approche intéressante est d'essayer d'estimer directement les motifs rythmiques d'un morceau. Le problème n'est pas aussi simple qu'il n'y paraît, et il est difficile de trouver des motifs sans connaître leur position temporelle ni leur tempo. C'est le paradoxe de l'œuf et de la poule : il est plus facile d'estimer les motifs lorsque l'on connaît leur position temporelle, et il est plus facile de trouver les positions temporelles des motifs lorsqu'on connaît leur forme. Pour résoudre ce problème, plusieurs approches ont été réalisées. La première consiste à synchroniser la recherche de motifs sur le tempo (et donc sur la mesure) [Dixon et al., 2004; Paulus et al., 2002]. La deuxième consiste à créer des motifs indépendants du tempo [Jensen et al., 2009]. La dernière consiste à créer des estimations

conjointes : des motifs rythmiques et de leurs positions [Tsunoo et al., 2009a,b], ou des motifs rythmiques et du tempo du morceau [Wright et al., 2008].

Motifs synchrones sur le tempo [Dixon et al., 2004] proposent des motifs de durée égale à une mesure. Pour trouver les instants de début et de fin des mesures, nécessaires à l'estimation de ses motifs rythmiques, ils utilisent l'algorithme BeatRoot [Dixon, 2001], dont ils corrigent manuellement les estimations. Le motif rythmique d'une mesure est l'amplitude de l'enveloppe du signal. Ils choisissent le motif dominant d'un morceau par partitionnement en k-moyennes : le motif dominant étant la moyenne géométrique de tous les motifs appartenant au groupe (cluster) principal.

[Paulus et al., 2002] estiment d'abord le tactus, le tatum et la durée de la mesure. Comme précédemment, un motif rythmique a la durée d'une mesure. Ils réduisent un motif rythmique à trois descripteurs (à 13 dimensions) représentant la brillance (centroïde spectral), le volume (logarithme de l'énergie du signal) et le timbre (MFCC). Ils mesurent la similarité entre différents motifs par DTW.

Motifs indépendants du tempo [Jensen et al., 2009] proposent une représentation des motifs rythmiques insensible aux petites déviations de tempo. Ils calculent d'abord l'auto-corrélation d'une fonction de détection d'onset sur des fenêtres temporelles de quatre secondes. Ils ré-échantillonnent ensuite ces auto-corrélations sur un axe de temps logarithmique. Ils montrent qu'un petit changement de tempo n'aura donc peu ou pas d'influence sur cette représentation.

Estimation conjointe des motifs et de leur position [Wright et al., 2008] s'intéressent aux motifs des claves dans la musique afro-cubaine. Ils proposent d'estimer conjointement les motifs et leurs positions. Les motifs rythmiques sont à priori connus et les auteurs proposent de considérer exhaustivement tous les tempos possibles et toutes les rotations possibles des motifs. Ils utilisent ensuite un score basé sur la corrélation croisée entre une fonction de détection d'onset adaptée à la détection de claves et toutes les possibilités de motifs rythmiques. Tous ces scores entrent ensuite dans un algorithme de programmation dynamique sensible aux rotations pour estimer réellement le tempo, les débuts de mesure et les motifs rythmiques.

2.4 Conclusion

Dans ce chapitre, nous avons montré qu'il existe de multiples définitions du mot rythme (partie 2.1). Nous choisissons la définition consistant à séparer le rythme en tempo, déviations et groupements (motifs rythmiques et métrique). Chacun de ces trois aspects fera l'objet d'un des chapitres suivants.

Nous définissons ensuite tous les concepts associés au rythme (partie 2.2). Nous insistons particulièrement sur les études perceptives et neurologiques montrant les phénomènes importants à prendre en compte lors de l'analyse automatique du rythme, comme l'ambiguïté de tempo.

Enfin, nous avons présenté un état de l'art exhaustif sur l'analyse automatique du rythme (partie 2.3). Nous remarquons que seulement un faible nombre de méthodes proposées s'inspirent de la perception.

Chapitre 3

Corpus d'évaluation

Contenu

3.1 Introduction

Dans ce chapitre nous décrivons les quatre corpus que nous avons utilisés tout au long de ce manuscrit pour évaluer nos modèles d'estimation du rythme.

Nous donnons tout d'abord un aperçu, dans la partie 3.1.1, de l'ensemble des corpus existants pour l'analyse automatique du rythme. Dans le cadre spécifique de tâches d'évaluation des déviations systématiques et des motifs rythmiques, les corpus existants ne sont pas totalement satisfaisants. C'est pourquoi deux des corpus d'évaluation que nous utiliserons (GTZAN-RHYTHM et EXTENDED BALLROOM) sont des contributions de cette thèse. Nous présenterons d'abord dans la partie 3.1.2 les critères nécessaires à la création d'un bon corpus d'évaluation.

Finalement, dans les parties suivantes, nous présenterons les quatre corpus utilisés dans ce manuscrit. Le premier, GTZAN-RHYTHM, (partie 3.3) est utilisé pour évaluer les méthodes d'estimation automatique de la métrique ou des déviations systématiques. Les trois suivants : BALLROOM, EXTENDED BALLROOM et CRETE sont utilisés principalement pour la classification en motifs rythmiques et sont décrits dans la partie 3.4.

3.1.1 Corpus existants

Il existe une multitude de corpus relatifs à la description des paramètres rythmiques à partir de l'audio. Nous présentons les principaux[1] dans la Table 3.1. Nous distinguons six types de tâches : l'estimation automatique du tempo, des battements, du 1er temps de la mesure, des déviations, de la métrique et des motifs rythmiques. Pour chacun des corpus listés, il est indiqué si les extraits audio sont accessibles et le type d'annotation disponible.

Comme on peut le constater dans la Table 3.1, de très nombreux corpus annotés en tempo existent. Une tâche plus spécialisée consiste à estimer automatiquement les instants du niveau métrique le plus important (tactus ou pulsation). Il est à noter que les difficultés des corpus varient beaucoup d'un corpus à l'autre. Récemment, [Holzapfel et al., 2012] ont créé un corpus (SMC) dont le but est de mettre en échec la majorité des algorithmes d'estimation automatique des temps, afin que ces algorithmes puissent continuer à s'améliorer. [Quinton et al., 2015] ont proposé très récemment un nouveau type d'annotation en métrique qui prend en compte le fait qu'il n'existe pas toujours un niveau métrique principal (tactus). Ce type d'annotation va certainement ouvrir des voix vers une estimation globale de tous les niveaux métriques, en s'affranchissant du tempo, ce qui, dans certains cas est plus logique du point de vue perceptif (par exemple lorsqu'il existe plusieurs tempo perçus pour un extrait musical [Peeters et al., 2014b]). Il n'existait pas de corpus ayant des annotations en déviations systématiques. C'est pourquoi nous avons créé le GTZAN-RHYTHM [Marchand et al., 2015a] (partie 3.3). Pour ce qui est de l'évaluation en motifs rythmiques, il existe deux corpus annotés : BALLROOM et CRETE. Nous avons mis à jour le corpus BALLROOM, afin d'avoir une meilleur qualité audio, une classe de rythme supplémentaire, et beaucoup plus de titres (voir partie 3.4.2).

1. Une liste plus exhaustive est disponible à l'adresse `http://www.audiocontentanalysis.org/data-sets/`.

Corpus		Tâche					
Nom	Accessibilité de l'audio	Tempo	Battements	1er temps	Déviations (swing)	Métrique	Motifs
ISMIR-2004 Songs	oui	✓	-	-	-	-	-
ISMIR-2004 Loop	oui						
Last-FM [Levy, 2011]	oui						
MIREX-2006	non	(✓)	✓	-	-	-	-
Hainsworth	oui						
Mazurka (MIREX)	non						
SMC / MIREX-2012	oui						
BALLROOM [Krebs et al., 2013]	oui	(✓)	✓	✓	-	-	-
Klapuri [Klapuri et al., 2006]	(oui)						
Isophonic-Beatles [Mauch et al., 2009]	oui						
RWC [Goto, 2006]	oui						
HJDB [Hockman et al., 2012]	non						
GTZAN-RHYTHM [Marchand et al., 2015b]	oui	(✓)	✓	✓	✓	-	-
GTZAN [Quinton et al., 2015]	oui	-	-	-	-	✓	-
BALLROOM [Cano et al., 2006]	oui	✓	-	-	-	-	✓
CRETE [Holzapfel et al., 2009]	(oui)						
EXTENDED BALLROOM [Marchand et al., 2016a]	oui						

TABLE 3.1 – *Principaux corpus existants pour l'analyse du rythme. Nous indiquons pour chaque corpus son nom, si l'audio est accessible, et s'il existe des annotations en tempo, battements, 1er temps des mesures, déviations, métrique et motifs. Les ✓ entre parenthèses signifient que l'on peut déduire l'information des autres annotations.*

3.1.2 Critères pour la création d'un bon corpus d'évaluation

Nous avons créé deux corpus d'évaluation au cours de nos travaux. Nous avons donc été amenés à nous interroger sur les critères permettant de former un bon corpus d'évaluation.

Concepts et limites.

Un point important est d'avoir une définition précise des concepts annotés, et d'être conscient des limites du corpus d'évaluation. Nous allons donc utiliser les travaux de [Peeters et al., 2012b] qui proposent des bonnes pratiques pour la description de corpus.

Contrainte de diffusion de l'audio.

Le principal intérêt de créer un corpus d'évaluation est de pouvoir comparer ses résultats à ceux de la littérature. Il est donc primordial de pouvoir le diffuser. Cela implique des contraintes fortes en matière d'extraits audio : il est possible d'utiliser

1. des extraits audio libres de droit,
2. des extraits audio que tout le monde peut récupérer facilement à l'identique,
3. ou des extraits audio que tout le monde possède déjà.

1. Libre de droit. La première possibilité (créer un corpus avec exclusivement de la musique sous licence libre) se révèle souvent insuffisante. Il existe des sites web ayant de la musique sous licence libre assez fournis comme `www.jamendo.com`, mais dont le nombre de morceaux spécifique à une tâche (comme l'estimation du facteur de swing par exemple) est assez limité. De plus, il est assez compliqué de créer un corpus d'évaluation à partir de morceaux relativement inconnus du grand public (ce qui est malheureusement encore le cas de la musique sous licence Creative Commons pour l'instant) : l'agrégation de méta-données est plus compliquée, mais surtout, la musique est moins représentative de la musique écoutée en général et de la production studio actuelle.

2. Récupérable à l'identique. La deuxième possibilité (donner les références permettant de récupérer les extraits musicaux à l'identique) est difficilement réalisable, mais possible si une attention particulière est portée sur les références/les outils disponibles pour construire le corpus. Deux choix s'offrent à nous : donner une référence précise des titres, ou donner un outils permettant de re-construire le corpus à l'identique. La première méthode est possible, mais demande beaucoup d'effort pour la personne cherchant à utiliser le corpus. Pour la plupart des titres de musique, il existe de multiples versions (studio, concert, voire plusieurs versions d'un CD à l'autre), sans compter les versions identiques avec des compressions audio différentes. Dans le cadre de l'estimation du swing, on a l'exemple de la Weimar Jazz Database utilisée par [Dittmar et al., 2015]. L'audio n'est pas téléchargeable directement, mais les transcriptions MIDI le sont et beaucoup d'efforts ont été réalisés pour aider à la diffusion de ce corpus : tous les morceaux sont clairement identifiés (titre, artistes, label, date d'enregistrement). Il est donc théoriquement possible de retrouver les enregistrements originaux. La deuxième méthode est d'utiliser un fournisseur de contenu. Certains

sites (comme www.7digital.com) proposent des extraits audio de 30 secondes pouvant être utilisés à des fins de recherche. Par contre, il est généralement interdit de les redistribuer. Il faut donc que chacun les télécharge individuellement, et il est difficile d'être sûr que ces extraits seront les mêmes dans quelques années. C'est pourquoi un effort particulier doit être fait dans les outils d'extraction du corpus afin d'assurer la reproductibilité des résultats. Le corpus d'évaluation EXTENDED BALLROOM que nous proposons suit ce principe. Nous décrivons précisément les mécanismes mis en place pour à la fois faciliter sa récupération et pour assurer son intégrité.

3. Déjà existants. La dernière possibilité (utiliser des extraits audio déjà largement diffusés) est peut-être la plus facile à réaliser en terme de diffusion d'audio. C'est la méthode choisie pour la création du corpus d'évaluation GTZAN-RHYTHM. Nous avons sélectionné l'ensemble de données nommé GTZAN publié par [Tzanetakis et al., 2002], qui a été utilisé un millier de fois pour l'évaluation de systèmes de classification en genre musical. Nous distribuons donc seulement les annotations que nous avons réalisé.

3.2 LEVY

Ce corpus est dérivé d'une expérience effectué par Last-FM et publiée par [Levy, 2011].

3.2.1 Audio (A)

Le corpus ne contient pas l'audio utilisé pour les expériences, seulement les annotations de tempo, le titre et l'artiste. Nous avons téléchargé des extraits audio de 30 secondes à partir du titre et du nom de l'artiste au moyen de l'API de 7-digital [2].

Nous ne sommes donc jamais sûr que l'audio que nous possédons correspond exactement aux annotations. Les informations du titre et de l'artiste peuvent être parfois ambiguës, ou plusieurs versions d'une même chanson peuvent exister. De plus, [Levy, 2011] a fait écouter un extrait de 30 secondes du titre, et nous ne savons pas lequel. Nos extraits choisis (par 7-digital) ne correspondent pas aux 30 premières secondes du morceau mais plutôt à 30 secondes représentatives de ce titre (on note que le refrain d'un morceau pop/rock est toujours présent). Nous ne pouvons qu'espérer que Levy a fait de même lorsqu'il a sélectionné les extraits pour son expérience.

Ces raisons font que nous ne sommes jamais sûr que notre audio, téléchargé au moyen de l'API de 7-digital, corresponde exactement à ce qu'ont écouté les annotateurs. Cependant, cette limitation s'équilibre avec le fait que, pour la musique populaire dont est composée le corpus, le tempo est très souvent constant. Donc l'audio que nous possédons est suffisant pour la tâche d'estimation du tempo.

2. http://developer.7digital.com/resources/api-docs

3.2.2 Annotations (B)

Origine des annotations (B1)

Pour ce corpus, [Levy, 2011] a fait appel à un grand nombre d'internautes, dans le cadre d'une expérience web perceptive. Dans cette expérience, il était demandé aux sujets d'écouter des extraits musicaux de 30 secondes, de choisir une classe de vitesse, de quantifier leur tempo (en bpm) et enfin de comparer le premier titre avec un autre extrait musical Nous nous intéressons uniquement à la partie annotation du tempo perceptif. L'annotation de tempo se faisait avec la barre d'espace.

Pour qu'une estimation de tempo soit prise en compte, il faut que l'auditeur ait tapé au moins 10 fois. S'il y a plus de 2 secondes entre les battements, le compteur est réinitialisé (ce qui limite le tempo minimal possible à 30 bpm). La moyenne de l'intervalle entre les battements est ensuite prise comme annotation de tempo.

Définitions des concepts (B21)

Les concepts annotés ne sont pas définis autrement que par les questions présentes sur l'interface web de l'expérience. Les annotateurs ont donc le choix entre trois classes *slow* (lent), *fast* (rapide) et *in-between* (entre les deux) pour la première question. Ils doivent taper le battement de l'extrait « Please tap the space bar along with the main beat of this track. ». Ils doivent enfin comparer l'extrait avec un autre selon les critères plus lent, à la même vitesse ou plus rapide.

Règles pour l'annotation (B22)

Aucune contrainte n'est définie pour l'annotation.

Annotateurs (B31)

Les annotateurs sont des auditeurs enregistrés de `last.fm` qui ont bien voulu donner un peu de temps pour l'expérience.

Fiabilité (B32)

Comme l'indique Levy, l'environnement dans lequel se fait l'expérience est sujet à caution. L'expérience est loin d'être une étude perceptive effectuée dans un environnement contrôlé. Il est impossible de s'assurer que les auditeurs ont bien écouté l'extrait musical avant de répondre.

Pour limiter les problèmes, l'expérience est restreinte aux seuls auditeurs inscrit sur `last.fm`. Levy a aussi mis en place un système de points et une page de classement représentant les meilleurs contributeurs, afin de motiver les gens à répondre à l'expérience. Pour limiter la tricherie, chaque réponse est liée à son annotateur, donc si un problème est découvert, il est facile supprimer toutes les annotations associées. Nous pouvons aussi noter que vu le peu d'intérêt qu'a la triche sur cette expérience, les annotations sont relativement fiables. Un autre point de l'expérience à noter, est que l'annotation se fait au moyen de la barre d'espace. La précision du tempo battu n'est donc pas très importante.

Cependant, les résultats de [Flocon-Cholet, 2012; Peeters et al., 2012a] ont montré que ce corpus était suffisamment fiable pour tester des algorithmes d'estimation de tempo perceptif.

3.2.3 Documentation (C)

Stockage (C2)

Les données étaient disponibles à l'adresse `http://users.last.fm/~mark/speedo.tgz` mais ne le sont malheureusement plus.

3.2.4 Perspectives

Cette étude a donc permis de réaliser une base de données assez conséquente, comme nous le montrons dans le Tableau 3.2. Nous possédons presque 4000 titres, annotés par 2000 auditeurs différents, pour un nombre total de presque 18000 annotations.

TABLE 3.2 – *Chiffres-clés du corpus de Levy.*

Nombre de titres	3698
Nombre d'annotateurs différents	1896
Nombre d'annotations de tempo	17884

Nous montrons dans la Figure 3.1 la répartition des genres musicaux des titres de la base. Nous avons extrait ces genres au moyen de l'API de 7-digital Pour les morceaux possédant plusieurs genres, nous n'avons pris que le genre principal. Nous possédons au final une base très majoritairement composée de musique pop/rock.

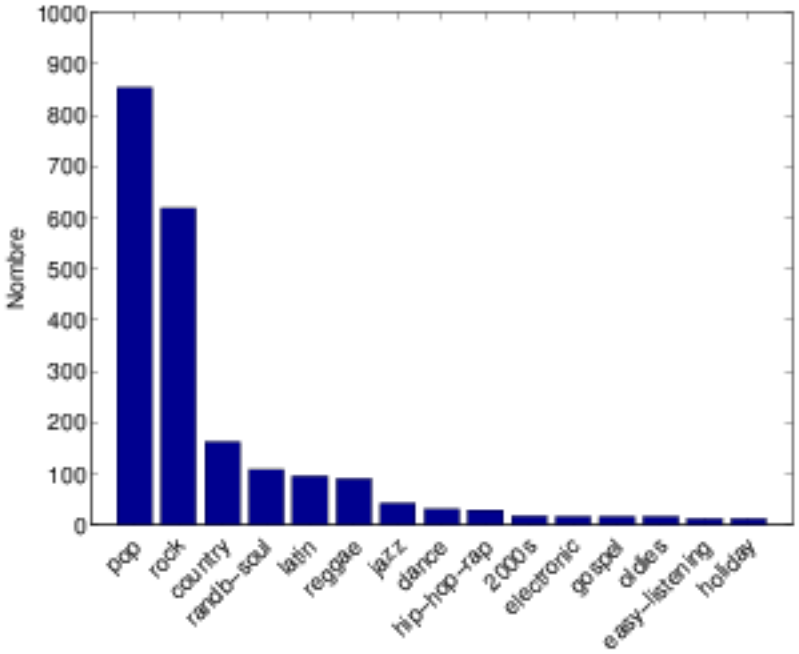

FIGURE 3.1 – *Genres musicaux du corpus, obtenu grâce à l'API de 7digital.com.*

Nettoyage

Le corpus LEVY est assez fiable pour tester une estimation de tempo perceptif lorsque les annotateurs sont d'accord entre eux [Peeters et al., 2012a]. Cependant, il s'est révélé inutilisable tel quel dans le cas où les auditeurs ne sont pas d'accord entre eux.

En effet, nous allons nous intéresser au cas où les auditeurs sont en désaccord, ce qui pose une question : les auditeurs ne sont-ils pas d'accord entre eux parce qu'il y a une ambiguïté de tempo (œ que nous aimerions), ou ne sont-ils pas d'accord parce qu'ils ont mal effectué l'expérience? Nous avons observé que dans un bon nombre de cas, c'est une erreur d'annotation qui est la cause du désaccord.

Ce problème n'est pas mis en valeur dans [Peeters et al., 2012a], car le fait de sélectionner les titres pour lesquels les auditeurs sont d'accord entre eux opère un filtrage des mauvaises annotations sur le corpus. Ce filtrage implicite n'est plus possible dans notre cas, étant donné que l'on s'intéresse justement au désaccord. Nous avons créé un outil permettant le nettoyage rapide du corpus dont nous montrons l'interfaœ sur la Figure 3.2.

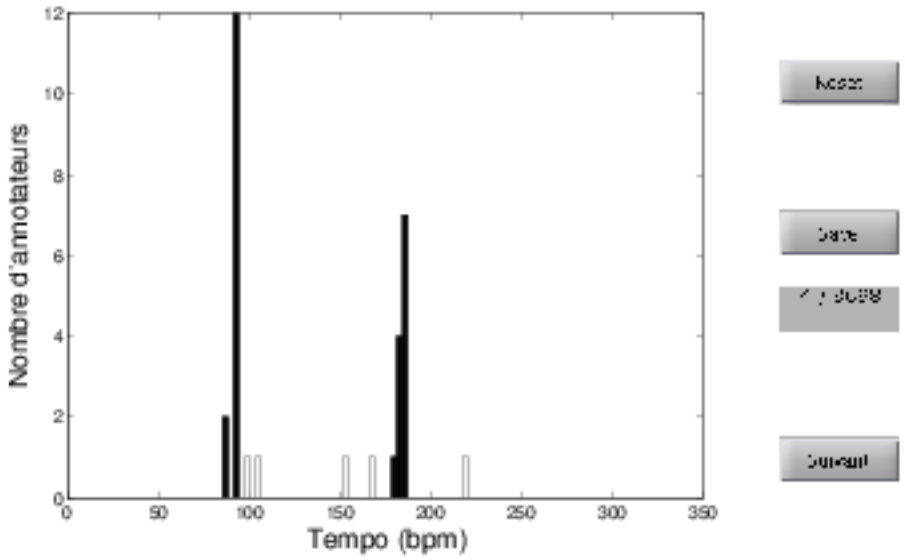

FIGURE 3.2 – *Capture d'écran de l'interface de nettoyage du corpus. L'histogramme est celui des annotations en fonction du tempo, pour un titre donné. Les annotations en noir seront conservées, les annotations transparentes seront supprimées.*

Sur cette figure, nous pouvons voir un exemple typique du travail que nous avons effectué pour chaque titre. Nous traçons pour chaque morceau du corpus les annotations de tempo sous forme d'histogramme. Puis nous sélectionnons manuellement celles que nous voulons conserver. Sur la figure, nous gardons les annotations en noir et nous supprimons les annotations transparentes. Dans cette figure, nous sommes dans le cas d'une ambiguïté de tempo, nous pouvons voir deux tempos principaux vers 90 et 180 bpm. Nous estimons que les 5 annotations transparentes ne doivent pas être associées à ces tempos principaux car les personnes qui les ont faites ne sont soit pas assez précises, soit n'ont pas compris la consigne.

Il est légitime de questionner une telle démarche, qui pourrait être vue comme une adaptation des données pour notre confort. Ce n'est pas vraiment

le cas, car ce nettoyage consiste simplement à ne conserver que les tempos principaux (qui sont des multiples l'un de l'autre) repérés par les annotateurs et supprimer les observations aberrantes et les titres pour lesquels aucun tempo ne semble prédominant. Compte-tenu du protocole expérimental de l'expérience (réalisée sur le web, aucun contrôle sur l'auditeur, précision floue due à la barre d'espace), nous pensons que cela est justifié.

Catégorisation $\mathcal{A}$ccord/$\mathcal{D}$ésaccord

Nous proposons d'utiliser l'écart inter-quartile des valeurs de tempo estimées pour déterminer à quelle classe ($\mathcal{A}$ccord/$\mathcal{D}$ésaccord) appartient un titre. L'écart interquartile (ou interquartile range, noté iqr) est une mesure de dispersion plus robuste dans notre cas que l'écart-type (qui est assez sensible aux valeurs extrêmes). L'iqr est calculé comme la différence entre le premier et le troisième quartile d'une série de mesures.

Pour chaque titre, nous calculons donc l'iqr du log_2[3] de tous les tempos annotés. Nous notons T_i le vecteur de toutes les annotations d'un titre i. Nous séparons le corpus en deux classe $\mathcal{A}$ccord/$\mathcal{D}$ésaccord selon le critère suivant :

— les titres de la classe $\mathcal{A}$ccord satisfont :

$$iqr\left(log_2\left(T_i\right)\right) < \tau = 0.2$$

— les titres de la classe $\mathcal{D}$ésaccord sont sélectionnés suivant :

$$iqr\left(log_2\left(T_i\right)\right) \geq \tau$$

Nous montrons dans la Figure 3.3 la distribution de l'iqr sur tous les titres du corpus, à gauche avant le nettoyage du corpus et à droite après. Nous représentons le seuil de séparation des classes $\mathcal{A}$ccord/$\mathcal{D}$ésaccord par la ligne pointillée verticale rouge. Nous observons bien l'effet du nettoyage : avant, une bonne partie des titres avaient une valeur comprise entre 0.1 et 0.9, ce qui les rendait peu facile à classer. Après nettoyage, la majorité des valeurs sont concentrées autour de 0 (classe $\mathcal{A}$ccord) et de 1 symbolisant un désaccord d'octave (classe $\mathcal{D}$ésaccord).

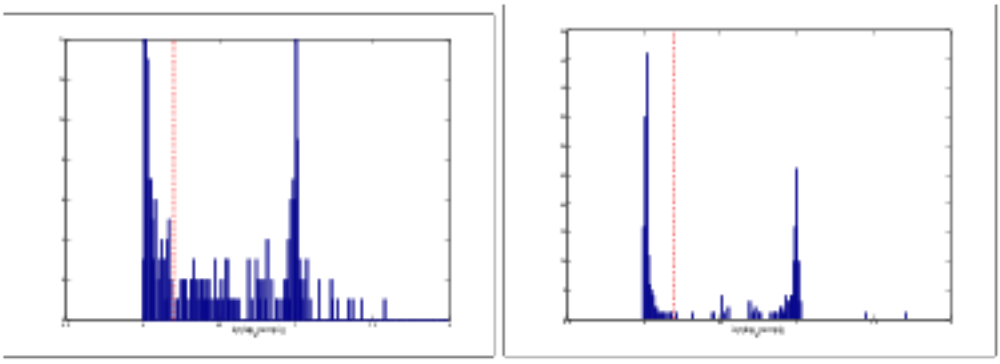

FIGURE 3.3 – *Distribution de l'$iqr(log_2(T_i))$. [Gauche] Avant nettoyage du corpus. [Droite] Après nettoyage. La ligne rouge verticale en pointillés représente le seuil τ de sélection de classe.*

3. Nous utilisons le log_2 des tempos annotés afin de mettre en évidence les erreurs d'octave. En effet, un tempo multiple de 2 d'un tempo de référence aura son log_2 supérieur de 1 : $log_2\left(2t\right) = 1 + log_2\left(t\right)$.

Nous montrons dans le Tableau 3.3 le nombre de titres utilisables à chaque étape de préparation de notre corpus. Nous voyons qu'il est drastiquement réduit. Malheureusement, les étapes de nettoyage et de sélection sont nécessaires pour avoir un corpus de travail fiable. Il nous reste donc 249 titres fiables, formant deux classes équilibrées : 134 titres sont dans la classe *Accord* et 115 titres dans la classe *Désaccord*. Ces titres vont donc former notre corpus d'étude, que nous utiliserons par la suite.

TABLE 3.3 – *Nombre de titres à chaque étape de préparation du corpus. En gras, le sous-ensemble du corpus qui va nous servir par la suite.*

		Nombre de titres
Corpus complet		3698
Après nettoyage		820
Avec 10 annotations minimum		**249**
dont	*Accord*	**134**
	Désaccord	**115**

3.3 GTZAN-RHYTHM

Une contribution de notre travail a été la création de l'ensemble d'évaluation GTZAN-RHYTHM, publié dans [Marchand et al., 2015a]. Ce corpus nous a permis d'évaluer quantitativement nos méthodes d'estimation de déviations systématiques décrites dans la partie 5.

Nous allons décrire ici le corpus GTZAN-RHYTHM : quels types d'annotations ont été réalisés et comment (partie 3.3.2), sous quelle forme est documentée et distribuée le corpus (partie 3.3.3) et enfin quelles sont ses possibles utilisations et ses limites (3.3.4).

Dans tout ce qui suit, on utilise les recommandations de [Peeters et al., 2012b] pour l'annotation de corpus en MIR. Les lettres entre parenthèses (comme par exemple « (B21) ») correspondent aux catégories de [Peeters et al., 2012b].

3.3.1 Audio (A)

Notre ensemble de données GTZAN-RHYTHM est basé sur l'ensemble de données bien connu GTZAN [Tzanetakis et al., 2002]. Cet ensemble a été créé pour évaluer la classification en genre musical. Il contient 1000 extraits audio de 30 secondes. Tous ces extraits sont distribués équitablement en 10 genres musicaux : blues, classique, country, disco, hip-hop, jazz, métal, pop, reggae et rock. Cet ensemble de données est très répandu, il existe plusieurs milliers de publications y faisant référence et comparant leur résultats d'estimation du genre musical sur ce même corpus. Cet ensemble est relativement bien adapté à notre tâche car il est constitué de suffisamment d'extraits ayant du swing pour permettre un entrainement et une évaluation.

3.3.2 Annotations (B)

Pour chacun des 1000 morceaux, nous avons annoté les positions de tous les battements et de tous les premiers temps des mesures. Pour chacun des morceaux ayant du swing, nous avons annoté les positions de toutes les croches.

Avec ces annotations manuelles, nous déduisons des données de plus haut niveau : le tempo, le ratio de swing, la métrique, et une mesure de confiance.

Origine des annotations (B1)

Les annotations en battements, premier temps et tatum ont été faites manuellement. Le tempo, le ratio de swing, la métrique et la mesure de confiance ont été calculés à partir de ces annotations manuelles.

Définitions des concepts (B21)

Battement : Nous utilisons la définition du battement que nous avons déjà proposée. C'est le niveau métrique sur lequel les auditeurs vont se synchroniser préférentiellement.

1er temps de la mesure : Nous avons annoté les premiers temps de chaque mesure (le niveau métrique juste au-dessus du tactus). Nous pouvons d'ores et déjà noter la très grande proportion de 4/4 (mesures de 4 noires) dans ce corpus. Le cas du reggae est particulier [4].

Swing : Pour chaque extrait ayant du swing, nous avons marqués les instants de toutes les croches présentes. Notons que musicalement, il n'y a pas systématiquement deux croches sur chaque battement, or seul les battements possédant deux croches sont intéressants pour le swing. Les battements comportant seulement une noire n'ont pas reçu de marquage de tatum (c'est le cas des deux premiers temps de l'exemple sur la Figure 3.5).

Tempo T : Le tactus précédemment annoté permet de déduire une information de tempo variable au cours du temps. En théorie, le tempo peut même changer au cours d'un morceau, mais en pratique, celui-ci est relativement constant pour nos (courts) extraits de 30 secondes.

En notant t_i l'ensemble des instants i des tactus, le tempo d'un extrait est calculé comme suit :

$$T = \text{moyenne}\left(\sum_i \frac{60}{t_i - t_{i-1}}\right) \tag{3.1}$$

Nous utilisons l'écart-type de la distribution $\frac{60}{t_i - t_{i-1}}$ pour modéliser la dispersion du tempo.

Ratio de swing S_r : Nous notons t_{s_i} l'ensemble des instants i des croches ayant du swing, t_{b_i} le battement précédant cette croche et $t_{b_{i+1}}$ le battement suivant cette croche. Nous avons représentés ces définitions sur la Figure 3.4. Le

4. Le reggae est caractérisé par une partie du signal situé à contre-temps. Ce sont souvent des guitares jouant staccato, doublées ou non par un piano ou un synthétiseur (le « skank »). Ces accords sont selon les définitions, soit situés sur les 2 et 4e temps d'une mesure 4/4, soit effectivement situé sur les contre-temps (le tempo est alors deux fois plus lent). Cela veut dire que deux tempos sont possibles lorsque l'on écoute un morceau de reggae. Dans la littérature, il semble que la première définition soit la plus utilisée (« It is true for both ska, rock steady, and for reggae that the rhythm is offbeat, ie on the second and fourth beat in a 4/4 measure » [Bennetzen et al., 1982]), même si elle va à l'encontre de l'envie que l'on a de garder un tempo plutôt lent pour du reggae. Nous avons choisi pour nos annotations cette définition, où les accords à contre-temps sont situés sur les deuxième et quatrième temps d'une mesure à quatre temps. Une autre caractéristique forte du reggae est le « kick » de la batterie sur le troisième temps (dans notre cas d'une mesure 4/4). Cette caractéristique est très utile pour distinguer le début d'une mesure, une fois les temps faibles repérés.

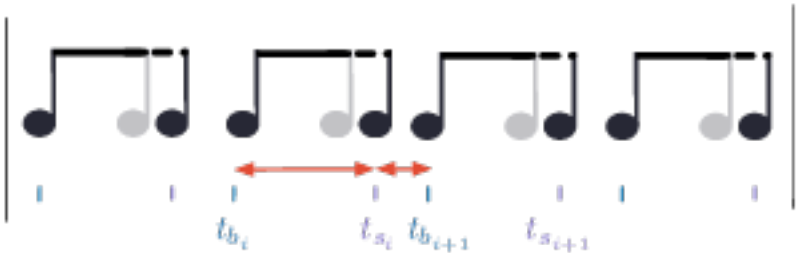

FIGURE 3.4 – *Définitions des t_{s_i} et t_{b_i}. Le ratio de swing est le rapport de la longueur de la flèche longue sur la longueur de la flèche courte.*

ratio de swing d'un extrait est calculé grâce à une médiane [5] comme suit :

$$S_r = \text{mediane}\left(\sum_i \frac{t_{s_i} - t_{b_i}}{t_{b_{i+1}} - t_{s_i}}\right) \tag{3.2}$$

De même, nous utilisons l'écart inter-quartile de la distribution $\frac{t_{s_i} - t_{b_i}}{t_{b_{i+1}} - t_{s_i}}$ pour modéliser la dispersion du ratio de swing.

Confiance : C'est le nombre de croches ayant du swing sur le nombre de temps de l'extrait.

Règles pour l'annotation (B22)

Les annotations ont été réalisées de façon semi-automatique, sur toute la durée des morceaux (30 secondes). Les positions des battements et des premiers temps ont d'abord été estimés automatiquement par la méthode de [Peeters et al., 2011], puis elles ont été corrigées manuellement une par une. Ensuite, pour chaque morceau ayant du swing, chaque temps a été scindé en deux afin de créer le tatum. Les tatums ont ensuite été ajustés globalement pour tout le morceau, puis individuellement corrigés afin de mieux correspondre localement à l'audio et au spectrogramme.

Annotateurs (B31)

Deux annotateurs ayant tous les deux une bonne expérience dans la musique et la recherche ont annotés l'ensemble de données. Chacun des annotateur a annoté 50% du corpus en battements et premiers temps. Ensuite l'un a vérifié le travail de l'autre et un consensus a été atteint en cas de désaccord. Le premier annotateur a annoté toute le corpus en tatum.

Une partie du corpus a été annotée par les deux annotateurs séparément afin d'en déduire des mesures de fiabilité (accord inter-annotateur) : 5% du corpus a été doublement annotée en battements (morceaux 95 à 99 de chaque genre), et 15% du corpus a été doublement annotée en premiers temps.

Fiabilité (B32)

La Table 3.4 présente les résultats d'accord entre les deux annotateurs pour les positions des battements, sur 50 morceaux.

5. Nous avons utilisé une médiane au lieu d'une moyenne car nous avons observé que cet estimateur semble plus robuste dans le cadre du ratio de swing.

TABLE 3.4 – *Accord inter-annotateurs sur la position des battements, sur un sous-ensemble de 5% du corpus* GTZAN *(50 morceaux, 5/100 morceaux de chaque genre).*

	F-measure	pScore	cmlC	cmlT	amlC	amlT	Inf. Gain
Score	90.70%	90.72%	83.56%	85.57%	92.29%	94.87%	3.99 bits

En ce qui concerne les positions des premiers temps des mesures, les deux annotateurs sont d'accords sur 150 extraits sur un sous-ensemble de 153 (soit 98% des extraits).

Outils d'annotation (B34)

Nous avons utilisé comme outil d'annotation Audiosculpt, qui permet de visualiser en même temps le contenu temporel (signal) et fréquentiel (spectral), d'écouter l'extrait, et de placer/déplacer graphiquement des marqueurs sur le signal. L'interface n'est pas limitante pour la tâche d'annotation que nous avons effectuée.

Un exemple de marquage est présenté sur la Figure 3.5. Les flèches noires indiquent les marqueurs de temps, et les flèches oranges les marqueurs de croches. Nous constatons qu'il n'y a pas une croche swinguée à tous les battements, et nous avons seulement été indiquée les croches audibles (ou visibles dans le spectrogramme). L'information de swing est donc assez clairsemée.

Les annotations de temps et mesures on été essentiellement faites à l'écoute. Audiosculpt permet en effet de placer des marqueurs et de les entendre (par un clic audio) lorsque l'on rejoue l'extrait. Ceci nous a permis de vérifier les positions de nos annotations. Les annotations en swing ont été essentiellement faites à l'aide du spectrogramme.

3.3.3 Documentation (C)

Identification du corpus (C1)

Les annotations sont disponibles à l'adresse : `http://anasynth.ircam.fr/home/media/GTZAN-rhythm`. La dernière version en date, publiée le 28/10/2015 a pour identifiant : v2_ismir2015latebreaking

Stockage (C2)

L'ensemble de données GTZAN-RHYTHM contient les dossiers et fichiers suivants :

xml (dossier) contient tous les fichiers d'annotations en format .xml. Ce dossier contient 3 sous-dossiers ('swing', 'no_swing' et 'ternary'). Les dossiers 'swing' et 'ternary' contiennent les annotations de tous les extraits ayant du swing ou ayant une mesure ternaire respectivement. Le dossier 'no_swing' contient le reste des annotations.[6]

jams (dossier) contient tous les annotations, sous un format de données différent, le format JAMS [Humphrey et al., 2014] (JSON Annotated Music Specification).

6. En pratique, il est à noter qu'aucun morceau n'est à la fois ternaire et swingué.

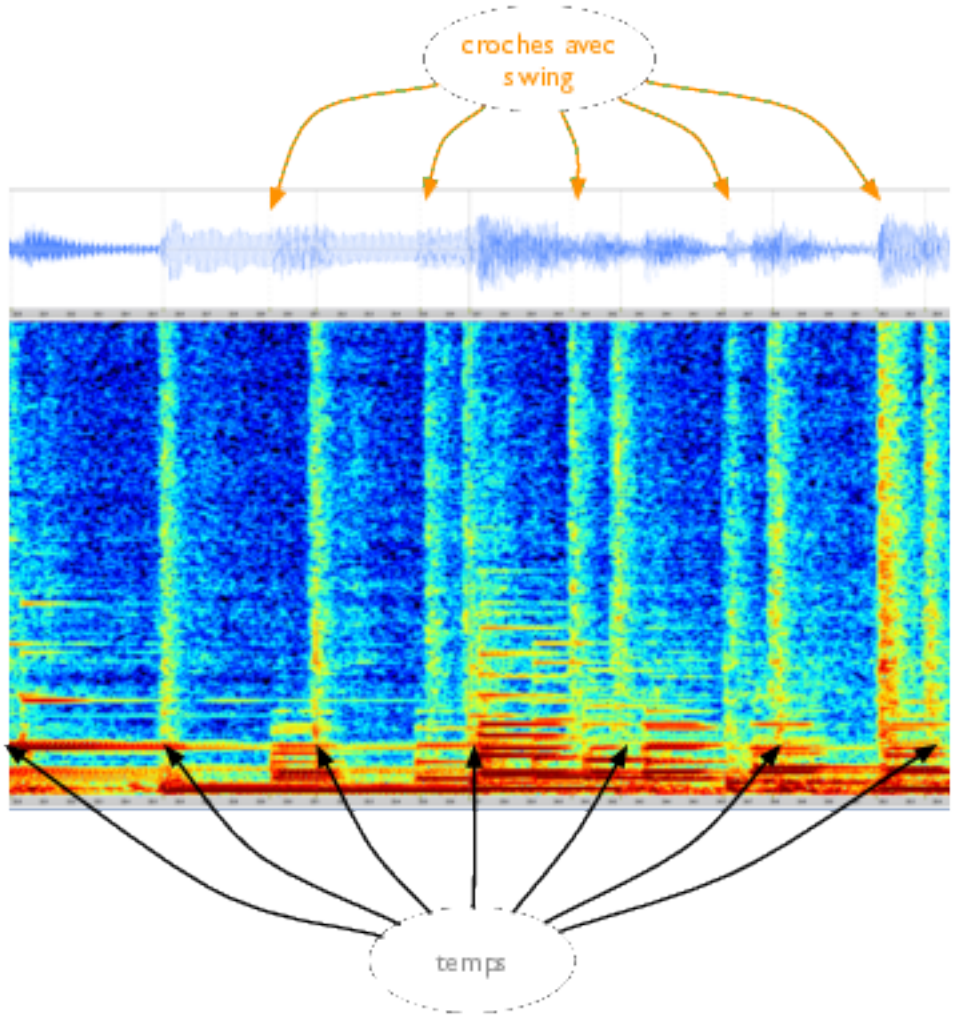

FIGURE 3.5 – *Exemple d'annotation : 22s à 25s de 'jazz.000059.wav'. De haut en bas, on voit le signal audio, puis le spectrogramme. Les traits verticaux en pointillés montre les annotations que nous avons faites. Les flèches noires indiquent des marqueurs de battements, et les flèches oranges des marqueurs des croches avec swing.*

GTZANindex.txt contient toutes les informations sur les extraits musicaux compilées par Bob Sturm [Sturm, 2013] (titre et artiste).

generate.py est un script python qui montre un exemple de lecture des fichiers d'annotations .xml. Ce script permet de générer les fichiers .jams ainsi que stats.csv.

stats.csv contient les informations de haut-niveau extraites des annotations brutes .xml et des métadonnées de Sturm. Ce fichier contient une ligne par extrait musical, contenant le nom du fichier, l'artiste, le titre, le tempo, le swing, la métrique...

Exemple d'un fichier d'annotation :

Ci-dessous, se trouve une partie d'un fichier XML d'annotation.

```
<segment time="24.8091256223" length="0" sourcetrack="0"  >
        <beattype id="1" pattern="0" measure="0" beat="1" tatum="1" />
</segment>
<segment time="25.190299" length="0" sourcetrack="0"  >
        <beattype id="1" pattern="0" measure="0" beat="0" tatum="1" />
</segment>
<segment time="25.3722095225" length="0" sourcetrack="0"  >
        <beattype id="1" pattern="0" measure="0" beat="1" tatum="1" />
</segment>
```

Chaque marqueur est encapsulé dans une balise *segment* dont l'instant est *time*. La plupart des autres balises ne sont pas pertinentes pour le sujet qui nous occupe, exceptée la balise *beat* qui vaut 1 lorsque le marqueur est un battement et 0 sinon (c'est-à-dire lorsque le marqueur indique donc une croche swinguée). La balise *measure* indique si le battement est le premier temps d'une mesure (1) ou non (0).

Annotations en fonction du genre

La Table 3.5 montre la distribution d'extraits possédant du swing en fonction du genre musical. La plupart des extraits ayant du swing sont du jazz et du blues (46 extraits pour chacun). Il y a aussi du swing dans le reggae (31 extraits), la country (17 extraits) et le rock (6 extraits). L'extrait classique classé comme swing est plus un cas particulier. C'est un morceau ayant un rythme de marche (ratio de swing 3:1). Il y a enfin 31 extraits ternaires, qui vont être considérés (dans nos expériences) comme des extraits ayant du swing (avec un rapport 2 :1). Au total, il y a donc 178 morceaux ayant des déviations systématiques et 822 n'en ayant pas.

3.3.4 Limites et perspectives

Concernant l'audio, Sturm a présenté dans [Sturm, 2013] les limites de l'ensemble GTZAN pour la classification en genre. Les principaux problèmes sont la présence de certains doublons et la sur-représentation de certains artistes dans certaines classes (comme le reggae, composé à 50% de Bob Marley). En ce qui concerne nos annotations en rythme, elles ont été faites manuellement, par deux annotateurs. La précision des instants annotés n'est peut-être pas parfaite, étant donné qu'elles ont été faites à l'écoute et à la main sur le spectrogramme. De plus, dans le cadre de l'estimation automatique de la position des temps/mesure, on peut noter que les extraits sont relativement faciles à analyser (percussions fortes,

TABLE 3.5 – *Distribution des extraits dans les classes Swing, Pas de swing et Ternaire en fonction du genre musical sur l'ensemble* GTZAN.

	Pas de swing	Swing	Ternaire
blues	50	46	4
classique	93	1	6
country	75	17	8
disco	98		2
hip-hop	100		
jazz	53	46	1
métal	94		6
pop	99		1
reggae	68	31	1
rock	92	6	2
Total	**822**	**147**	**31**

souvent du 4/4, ...), à part le classique et le jazz qui sont difficiles (beaucoup de rubato, métrique complexe).

Cependant, ces annotations sont tout à fait adaptée pour l'estimation du ratio de swing (150 titres possédant du swing, musique commerciale réelle). Le corpus permet de tester des algorithmes d'estimation de beat/downbeat (battement/1er temps des mesures) sur quelques cas compliqués comme la musique classique (beaucoup de rubato) et jazz (pulsation et mesure souvent difficiles à déterminer).

3.4 Estimation de motifs rythmiques

Dans cette partie, nous décrivons les trois ensembles de données que nous utiliserons pour l'évaluation des descripteurs de rythme. Le second, EXTENDED BALLROOM est l'une de nos contributions.

3.4.1 BALLROOM

Le corpus BALLROOM a été créé pour le concours de description du rythme d'ISMIR 2004 [Cano et al., 2006]. Il a été extrait du site web `www.ballroomdancers.com`[7] à ce moment-là. Il contient 698 extraits musicaux de 30 secondes, divisés en 8 genres musicaux, représentants différentes danses de salon : ChaChaCha, Jive, Quickstep, Rumba, Samba, Tango, VienneseWaltz (valse viennoise), Waltz (valse). Comme ces genres musicaux sont étroitement liés à leur rythme prédominant, on peut utiliser ces étiquettes de genre comme des étiquettes de rythme.

3.4.2 EXTENDED BALLROOM

Le corpus d'évaluation précédent a été extrait il y a plus de 10 ans maintenant. Vu son faible nombre d'extraits et la mauvaise qualité audio de ceux-ci, nous avons décidé de l'actualiser. Comme le site web ayant permis l'extraction existe encore et propose toujours un corpus de danses de salon annotée en tempo,

7. Ce site propose à la vente des CDs de danses de salon, et permet d'écouter des extraits de 30 secondes de chaque extrait avant de les acheter.

avec des extraits de 30 secondes pour chaque titre, nous avons pu extraire une nouveau corpus de celui-ci. De plus, nous avons annoté tous les différentes types de répétitions que nous avons pu trouver parmi les 4180 titres téléchargés.

Les avantages de l'ensemble de données EXTENDED BALLROOM sur le BALLROOM sont nombreux. La qualité des fichiers audio a été améliorée, le corpus est six fois plus volumineux, il possède 5 nouvelles classes rythmiques, et nous avons complété les méta-données avec des annotations des différents types de répétitions.

Nous allons maintenant décrire précisément ce nouveau corpus.

3.4.2.1 Audio (A)

Nous indiquons la distribution en classe de ce nouveau corpus dans le Tableau 3.6. Comme nous pouvons le constater, les différentes classes de rythmes sont passées de 87 extraits en moyenne à 444 extraits par classe (à part la classe valse viennoise, qui possède seulement 250 titres). Nous remarquons aussi l'apparition d'une nouvelle classe : Foxtrot. Il est à noter qu'en fait quatre autres nouvelles classes sont disponibles : Pasodole, Salsa, Slowwaltz, Wcswing. Cependant, comme elles possèdent relativement aux autres classes dix fois moins d'extraits (47 en moyenne par classe), nous avons choisi de ne pas les utiliser dans nos prorpes évaluations afin de garder un ensemble de données ayant des classes de tailles à peu près uniformes. Nous les publions quand même, car l'apprentissage et la classification sur des classes de tailles non-uniformes peut-être un problème en soit et donc intéresser certaines personnes.

TABLE 3.6 – *Répartition des classes de rythmes pour les deux ensembles de données* BALLROOM *et* EXTENDED BALLROOM.

Classe de rythme	BALLROOM	EXTENDED BALLROOM v1
Chacha	111	455
Jive	60	350
Quickstep	82	497
Rumba	98	470
Samba	86	468
Tango	86	464
Viennesewaltz	65	252
Waltz	110	529
Foxtrot		507
Pasodoble		53
Salsa		47
Slowwaltz		65
Wcswing		23
Total	**698**	**4180**

Il est à noter que le corpus BALLROOM n'est pas strictement inclus dans le corpus EXTENDED BALLROOM. Plusieurs albums et titres, originalement présents en 2004 ne sont désormais plus vendus. Seulement 343 morceaux sur les 698 que contenait le BALLROOM sont présent dans l'EXTENDED BALLROOM [8]. Regrouper

8. L'intersection des ensembles de données a été faite grâce aux ids des morceaux et les couples titre de l'album / titre du morceau. Nous n'avons pas utilisé le contenu audio pour cela.

les deux ensembles n'aurait que peu de sens au vu de la différence des qualités des fichiers audio.

3.4.2.2 Annotations (B)

Pour chacun des 4180 morceaux de l'ensemble EXTENDED BALLROOM, nous fournissons les annotations suivantes : tempo (en bpm), genre (dénoté aussi classe de rythme), artiste, titre de l'extrait, titre de l'album. Nous indiquons aussi pour chaque extrait ses liens avec les autres extraits de l'ensemble de donnée, basés sur leurs similarités. Nous définirons ci-dessous quatre types de similarités.

Origine des annotations (B1). Le tempo, la classe de rythme, l'artiste, le titre de l'extrait et de l'album sont tous extraits automatiquement du site web. L'annotation des similarités a été faite de façon semi-automatique. Les répétitions possibles ont été trouvées par l'algorithme Audio-Print[Ramona et al., 2013] et en trouvant les extraits ayant des titres similaires (au moyen d'une distance d'édition). Chacune des répétitions potentielles ont ensuite été vérifiées manuellement.

Définition des concepts (B21). Nous ne pouvons pas donner de définition précise du tempo et du genre (classe de rythme), étant donnée que ces valeurs ont été extraites automatiquement du site web. Ces valeurs sont destinées avant tout aux dans-eur/euse-s qui apprennent les danses de salon, on peut donc considérer qu'elles sont basées sur l'usage qu'ils/elles en font.

Nous définissons quatre catégories de répétitions entre extraits musicaux (inspirées par les travaux de [Sturm, 2013]) :

Répétition exacte : les domaines temps/fréquences entre les deux extraits sont très hautement similaires (différences d'encodage audio, ou présence d'un léger retard temporel).

Répétition temporelle : mêmes extraits, commençant à deux instants différents.

Répétition karaoké : deux extraits identiques, dont l'un ayant la voix chantée supprimée, ou remplacée par un instrument.

Répétition de version : les deux compositions sont les mêmes, mais jouées de façon différente, cela peut être une différence entre version studio et concert, ou des versions transposées de la même composition, ou la même chanson avec des instrumentations différentes.

La répartition de ces différents types de répétitions est présentée dans la Table 3.7.

TABLE 3.7 – *Répartition des différents types de répétitions sur l'ensemble de donnée* EXTENDED BALLROOM.

Type	Exact	Temporel	Karaoke	Version
#	248	16	12	257

Fiabilité (B32). Le tempo, le genre, le nom de l'artiste, le titre de la chanson et le titre de l'album sont tous extraits automatiquement du site web. Nous n'avons aucune explication sur comment ces annotations ont été faites. Cependant, les

annotations de tempo et de genre (classe de rythme) ont été utilisées par la communauté scientifique pendant plus de 10 ans, sans qu'aucune critique ne soit faite. De plus, lors des nombreuses écoutes que nous avons faites pour la créations des annotations de répétitions (1155 extraits sur les 4180), nous n'avons relevé aucune erreur de genre.

La fiabilité du contenu audio de ce corpus est un enjeu majeur pour sa diffusion, car nous ne pouvons pas distribuer le contenu audio avec les méta-données, pour des raisons de capacité de stockage et de droits d'auteurs. Nous fournissons donc un script Python, permettant de télécharger tous les extraits musicaux. Point plus important, ce script permet aussi, de vérifier que tous ces extraits sont identiques à ceux que nous avons utilisé. Cette vérification est faite au moyen de sommes MD5 (la somme MD5 de chaque extrait audio est fournie dans les métadonnées).

Finalement, comme les annotations en répétitions ont été faites de façon semi-automatiques (les extraits ayant un titre similaire ou une empreinte digitale similaire ont été vérifiés manuellement), il est possible que nous ayons oublié quelques répétitions. Cependant ces répétitions oubliées ne peuvent être que peu nombreuses étant donné que nous avons paramétré nos systèmes de détection automatique pour qu'ils sortent de nombreux faux positifs.

Outils d'annotation (B34). Nous référons le lecteur à la publication de [Ramona et al., 2013] pour une explication complète du système permettant la création 'd'empreintes digitales' d'un extrait audio, qui permettent de trouver les doublons d'un corpus musical.

3.4.2.3 Documentation (C)

Identification du corpus (C1). L'identifiant de ce corpus est Extended Ballroom v1.

Stockage (C2) Cette base a été rendue publique dans [Marchand et al., 2016a,b]. Elle est disponible à l'adresse `http://anasynth.ircam.fr/home/media/ExtendedBallroom`. Elle est distribuée sous forme de :

- Un fichier XML contentant toutes les méta-données (tempo, genre, artiste, titre de l'extrait, titre de l'album, répétitions, somme MD5).
- Un script Python capable de télécharger et vérifier l'intégrité de tout le contenu audio du corpus.
- Un fichier README décrivant le format XML.

3.4.2.4 Applications

En dehors de l'application directe liée à l'estimation automatique de classes de rythme, cet ensemble de données peut aussi être utile pour l'évaluation d'algorithmes d'estimation automatique du tempo. Nous apportons une contribution importante à ce corpus en fournissant la liste des extraits similaires entre eux, selon plusieurs catégories (répétitions exactes, temporelles, karaoke ou de version). Une des caractéristiques de ce corpus est le grand nombre de versions différentes que l'on peut trouver pour certains extraits (chaque version pouvant avoir des tonalités, des interprètes, des instrumentations ou des tempos différents). Ces annotations pourront sûrement être utiles pour l'évaluation d'algorithme de détection de version (cover song detection). Finalement, le grand

nombre d'extrait (plus de 4000) permet maintenant d'envisager de pouvoir utiliser des méthodes basées sur les réseaux de neurones profonds, car celles-ci ont besoin d'un grand nombre de données d'apprentissage.

3.4.3 CRETE

Le troisième et dernier ensemble de données pour la description du rythme est celui que nous appelons "CRETE".

Cet ensemble de données a été introduit par Holzapfel dans [Holzapfel et al., 2009] [Holzapfel et al., 2011]. Il contient 180 extraits de danses traditionnelles de l'île de Crête. Ces 6 danses sont *Kalamatianos*, *Kontilies*, *Maleviziotis*, *Pentozalis*, *Sousta* et *Kritikos Syrtos*. Dans ce jeu de données, les classes de rythme sont beaucoup plus difficiles à estimer que dans le corpus BALLROOM comme nous l'expliquerons dans la partie suivante.

3.4.4 Limites et perspectives

Distribution de tempo

Nous montrons dans les figures suivantes les distributions de tempo pour chaque classe de rythme des ensembles BALLROOM (Figure 3.6), EXTENDED BALLROOM (Figure 3.7) et CRETE (Figure 3.8[9]).

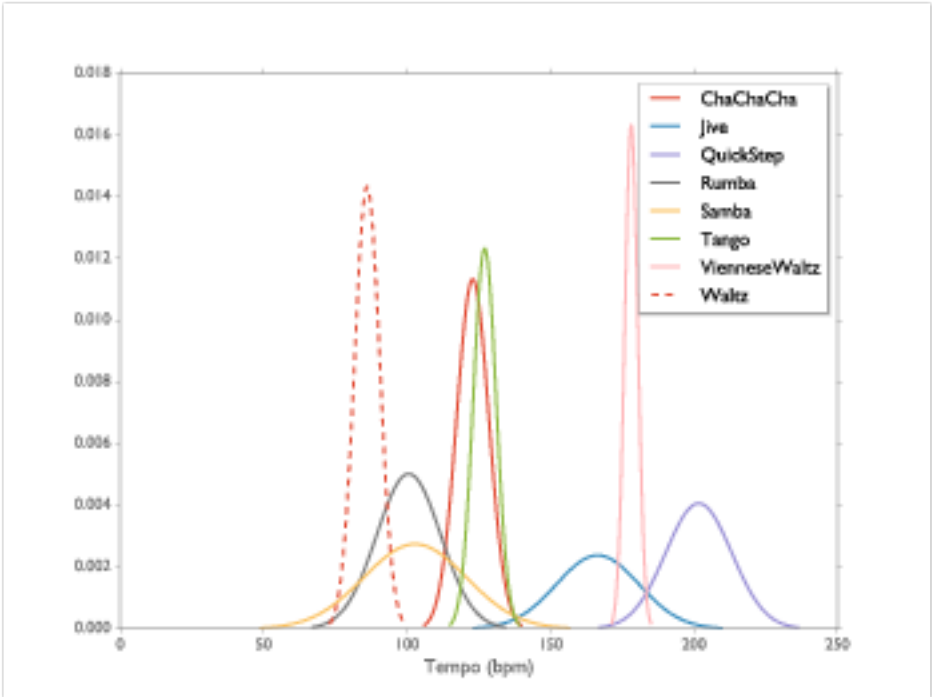

FIGURE 3.6 – *Distribution des tempo de l'ensemble de données* BALLROOM

Dans le cas de l'ensemble BALLROOM, nous remarquons que les classes Samba et Rumba ainsi que les classes Tango et ChaChaCha sont très proches en terme de tempo. Sinon, le tempo permet de définir sans trop d'ambiguïtés les trois autres classes de rythmes.

9. Cette troisième figure est équivalente à la Fig.3 de [Holzapfel et al., 2009], en divisant par deux les tempos annotés de 'kal' et de 'syrt'. L'auteur a donné les tempos à la noire, nous avons donné les tempos annotés.

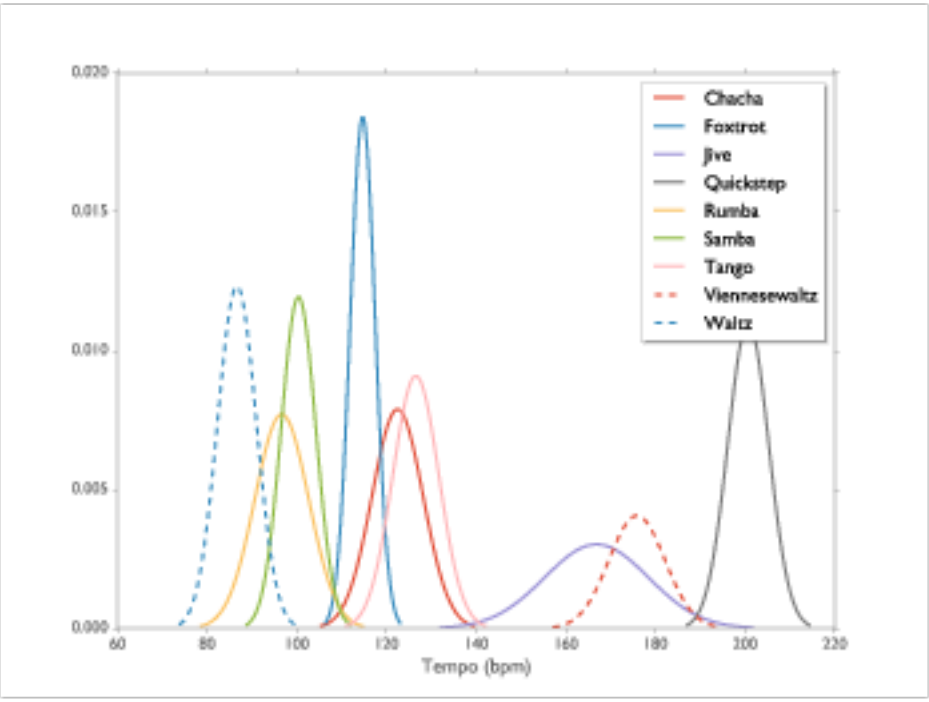

FIGURE 3.7 – *Distribution des tempo de l'ensemble de données* EXTENDED BALLROOM

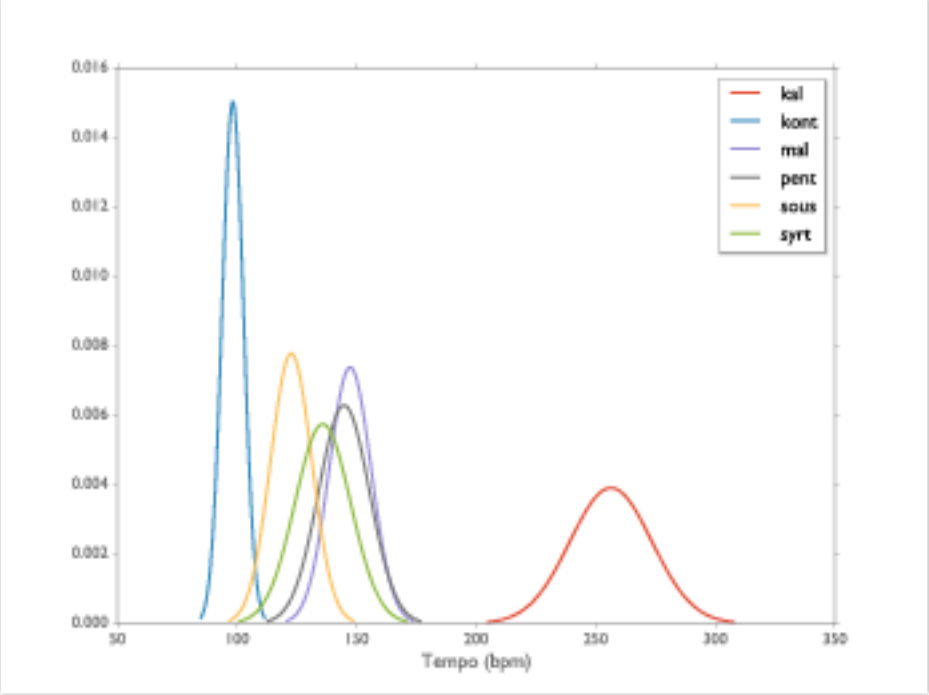

FIGURE 3.8 – *Distribution des tempo de l'ensemble de données* CRETE

Dans le cas de l'ensemble EXTENDED BALLROOM, nous sommes dans la même configuration (Samba/Rumba très proches en tempo et ChaChaCha/-Tango aussi). La nouvelle classe Foxtrot ajoutée est, d'un point de vue tempo, relativement indépendante des autres.

Dans le dernier cas de l'ensemble CRETE, trois classes de rythmes ont des tempos proches (mal, pent, syrt). Les classes kont, sous et kal se démarquent des autres.

D'un point vue du tempo seul, l'ensemble CRETE est donc relativement plus ambigu que les ensembles BALLROOM et EXTENDED BALLROOM. Nous remarquons déjà que montrer une bonne classification en rythmes sur ces trois ensembles ne montrera pas nécessairement que le descripteur utilisé est indépendant du tempo. Un descripteur très corrélé avec le tempo pourrait en effet donner de bons résultats de classification en rythmes sur ces ensembles de données. Une attention tout particulière devra donc être portée lorsque nous créerons des descripteurs ayant pour but de représenter les motifs rythmiques indépendamment du tempo.

Publications associées

Marchand, U., Fresnel, Q. et Peeters, G. (2015a). « GTZAN-Rhythm : Extending the GTZAN Test-Set with Beat, Downbeat and Swing Annotations ». *Late-Breaking-Demo Session of the 16th International Society for Music Information Retrieval Conference (ISMIR)*.

Marchand, U. et Peeters, G. (2016b). « The Extended Ballroom Dataset ». *Late-Breaking-Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR)*.

Chapitre 4

Tempo perceptif

Contenu

4.1 Introduction

4.1.1 Le tempo perceptif

Le tempo est un élément perceptif important pour la description du rythme. Il est défini comme la vitesse d'un extrait musical, exprimée en battement par minute (bpm). Dans ce chapitre nous allons parler uniquement de tempo perceptif, par opposition au tempo annoté sur la partition. Tout le monde est capable d'estimer le tempo d'un extrait musical en frappant les temps dans les mains par exemple.

Le tempo est donc une variable qui est définie par les auditeurs. Plusieurs études dont œlles de [Levy, 2011 ; Moelants et al., 2004] ont mis en évidence le phénomène d'ambiguïté de tempo [1]. En effet le tempo est avant tout une donnée perceptive : il arrive que plusieurs auditeurs auxquelles on demande de frapper le tempo d'un extrait frappent des tempos différents les uns des autres. Ces tempos sont souvent des multiples entiers les uns des autres. Le cas le plus courant est un auditeur frappant un tempo, et l'autre frappant deux fois plus vite : ce sont les fameuses « erreurs d'octave ».

La question à laquelle nous allons essayer de répondre dans ce chapitre est donc la suivante : quelles sont les caractéristiques du signal audio qui rendent la perception du tempo ambigüe ?

Nous avançons comme hypothèse que les auditeurs peuvent se baser sur différentes caractéristiques du signal pour inférer le tempo. Cela implique que si le signal possède des caractéristiques ambigües, alors différents auditeurs pourrons percevoir différents tempos. Par exemple, si la variation d'énergie du signal donne un tempo T_1, que les changements d'harmonie donnent un tempo T_2 (différent de T_1) et que la similarité à court-terme donne un tempo T_3 (différent de T_1 et de T_2) alors un premier auditeur pourra percevoir T_1, un autre auditeur T_2 et un dernier T_3. Nous allons donc proposer dans ce chapitre différentes façons de modéliser les relations entre les différentes caractéristiques du signal audio afin d'estimer si, pour un extrait musical donné, les auditeurs percevront le même tempo ou non.

Nous proposons un bref état de l'art sur l'estimation du tempo perceptif dans la partie 4.1.2. Puis nous proposons différents modèles de relations entre les différentes caractéristiques du signal dans la partie 4.2.2. Enfin nous les évaluons dans la partie 4.3.

4.1.2 État de l'art

Un très grand nombre de travaux existent sur l'estimation automatique du tempo dans la musique. Cependant, très peu s'attaquent au problème du tempo tel qu'il est perçu par les utilisateurs. [Xiao et al., 2008] émettent l'hypothèse qu'il y a un lien entre le timbre de l'extrait musical et son tempo perçu. [C.-W. Chen et al., 2009 ; Gkiokas et al., 2012 ; Hockman et al., 2010] proposent des systèmes permettant de réduire les erreurs d'octave. Seuls [Peeters et al., 2012a] estiment directement le tempo perceptif à partir de descripteurs basés sur des considérations perœptives.

[Xiao et al., 2008] émettent l'hypothèse que le tempo perçu est lié au timbre de l'extrait musical. Ils représentent donc chaque musique par un vecteur de MFCC. Ils utilisent ensuite un GMM à 8 gaussiennes pour modéliser les MFCCs

1. Une liste plus exhaustive est présente dans notre état de l'art, partie 2.2.2.

extraits et le tempo annoté. Pour chaque morceau de musique inconnu, une première estimation de tempo T_e est faite. Le modèle GMM sert ensuite à estimer les probabilités de T_e, $\frac{T_e}{2}$, $\frac{T_e}{3}$, $2T_e$, $3T_e$. Les auteurs sélectionne alors le tempo ayant la plus grande probabilité. Cette méthode donne des résultats similaires aux algorithmes d'estimation de tempo actuels.

[C.-W. Chen et al., 2009] proposent une méthode de correction des erreurs d'octave. Leur hypothèse est que le tempo est lié au caractère du morceau (par exemple : agressif signifiera souvent un tempo rapide, alors que romantique ou sentimental signifiera un tempo lent). Un algorithme Machines à Vecteurs de Support ou « Support Vector Machine » (SVM) basé sur une centaine de descripteurs apprend 4 classes de tempo (de lent à rapide). Puis suivant ces classes, le tempo estimé est multiplié par 2, divisé par 2, ou laissé inchangé. Cette méthode permet d'améliorer les résultats de beaucoup d'algorithmes de l'état de l'art.

[Hockman et al., 2010] proposent aussi de réduire les erreurs d'octave en apprenant des classes de tempo. Ils représentent chaque morceau par une série de descripteurs générés par jAudio [McKay et al., 2005]. L'apprentissage se fait au moyen de 6 algorithmes de classification et ils utilisent la séparation en classes pour corriger le tempo. Le point à noter de cet article est qu'ils n'utilisent pas de fonctions de périodicité pour ce problème mais une série de 80 descripteurs liés à la hauteur, à l'intensité et au timbre du morceau. L'algorithme de classification donne d'excellents résultats, mais la correction de tempo ne montre pas de résultats vraiment améliorés.

[Gkiokas et al., 2012] proposent aussi de réduire les erreurs d'octave en apprenant des classes de tempo. Ils représentent chaque morceau par un vecteur de périodicité et une classe (lent/modéré/rapide). Ils utilisent un modèle SVM pour apprendre la classe de tempo. Enfin ils estiment le tempo perceptif comme le pic prédominant de la fonction de périodicité, appartenant à l'intervalle de la classe estimée.

[Peeters et al., 2012a] proposent d'estimer directement le tempo perceptif au moyen de quatre indices acoustiques basés sur des considérations perceptives (variation d'énergie, similarité à court-terme, variation harmonique et alternance grave/aigu). Ils utilisent ensuite une régression GMM pour estimer directement le tempo perceptif. Cette méthode a permis une diminution des erreurs d'octave et une meilleure répartition de celles-ci sur l'échelle des tempos.

4.2 Méthode d'estimation *Accord*/*Désaccord*

À l'inverse des travaux précédents (voir partie 4.1.2), notre objectif n'est pas d'estimer le tempo perçu mais d'estimer si les auditeurs vont partager (*Accord*) ou pas (*Désaccord*) la perception du tempo.

4.2.1 Descripteurs audio

Pour cela, nous avons choisi quatre descripteurs audio. Il s'agit des descripteurs de [Peeters et al., 2012a] choisis pour leur relation avec la perception du tempo. Ces quatre descripteurs représentent les périodicités du signal suivant plusieurs points de vue différents.

$d_{onset}(\tau)$ représente les périodicités dues aux variations énergétiques du signal. Il permet de mettre en évidence les périodicités dues aux attaques du signal musical.

$d_{sim}(\tau)$ représente les périodicités dues aux similarités structurelles à court-terme. Il permet de mettre en évidence la quantité de changement et de répétition de plusieurs aspects musicaux : contenu fréquentiel, notes et rapport bruit/harmonicité.

$d_{spectral}(\tau)$ représente les périodicités dues à l'alternance entre les haute fréquences et les basses fréquences. L'hypothèse derrière ce descripteur est assez forte et très adaptée à la musique pop/ rock. Elle se résume comme : l'extrait musical possède quatre temps par mesure et les percussions suivent un motif bien défini, à savoir que la grosse caisse joue sur les premiers et troisièmes temps et que la caisse claire joue sur les deuxièmes et quatrièmes temps. Les basses fréquences (grosse caisse) et les hautes fréquences (caisse claire) doivent donc alterner à la vitesse de la moitié du tempo.

$d_{harmo}(\tau)$ représente les périodicités dues aux changements harmoniques. Ce descripteur est basé sur l'hypothèse que le taux de changements d'accords est directement proportionnel au tempo. Pour une mesure à quatre temps, le taux de changement d'accord vaut donc $\frac{1}{4}$ du tempo. contradictoires alors il y aura certainement désaccord.

Dans $d_{onset}(\tau)$, $d_{sim}(\tau)$, $d_{spectral}(\tau)$ et $d_{harmo}(\tau)$, τ représente soit le décalage d'une fonction d'auto-corrélation (en échelle logarithmique) soit la fréquence de la DFT.

4.2.2 Modèles de prédiction

Les modèles de prédiction *Accord/Désaccord* que nous allons proposer sont basés soit sur des résultats perceptifs, soit une mesure des relations qu'il y a entre ces descripteurs. Notre hypothèse est que si les descripteurs partagent des informations cohérentes alors il y a de grandes chances pour que les auditeurs soit d'accord entre eux sur le tempo et que réciproquement, si les descripteurs ont des informations contradictoires il soient en désaccord.

Nous proposons dans cette partie quatre modèles de prédiction de l'*Accord* entre auditeurs sur le tempo perceptif. Le schéma général de notre méthode d'estimation *Accord/Désaccord* est illustré sur la Figure 4.1. Nous extrayons les quatre descripteurs proposés dans la partie 4.2.1 à partir du signal audio, puis nous modélisons les relations entre ces descripteurs pour obtenir une classification *Accord/Désaccord* entre les auditeurs.

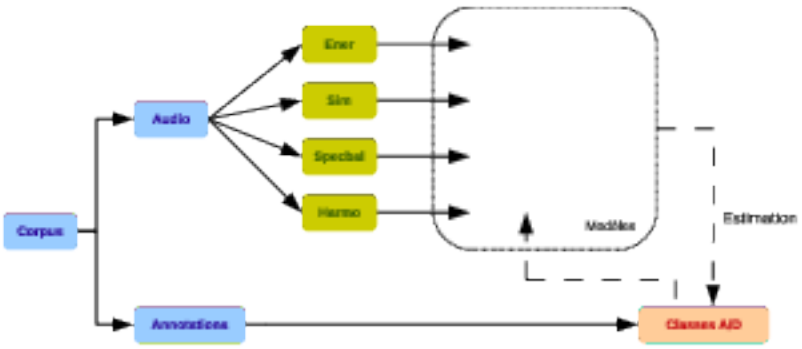

FIGURE 4.1 – *Schéma général de la classification Accord/Désaccord.*

Nos modèles seront entraînés et évalués sur la base LEVY décrite dans la partie 3.2. Nous avons extrait de cette base les annotations *Accord*/*Désaccord*.

Les quatre modèles sont les suivants. Le modèle (A) est basé sur les travaux de McKinney et Moelants. Il s'appuie sur l'existence d'un tempo perceptif préférentiel. Le modèle (B) Feature-GMM consiste simplement à essayer d'apprendre les classes *Accord*/*Désaccord* directement à partir des descripteurs. Le Modèle (C) Inform-GMM cherche à exploiter l'hypothèse que l'accord entre les auditeurs est lié à l'accord entre les descripteurs. Enfin, le modèle (D) Tempo-GMM exploite la même hypothèse, mais en apprenant les classes *Accord*/*Désaccord* à partir de quatre estimations de tempo, obtenues à partir des descripteurs pris indépendamment.

Modèle A (MM-Onset et MM-Sim). Ces premiers modèles sont basés sur deux hypothèses faites dans [McKinney et al., 2004] :

1. il existe un tempo préférentiel autour de 120 bpm (cette hypothèse a été confirmée par plusieurs travaux, comme nous avons pu le voir dans la partie 2).
2. si un extrait musical (dans notre cas, un des quatre descripteurs) contient un niveau métrique dont le tempo est proche du tempo résonnant (entre 110 et 170 bpm d'après les auteurs), le tempo perçu a de grandes chances de ne pas être ambigu. Si, par contre, les tempos principaux encadrent le tempo résonnant, le tempo perçu aura tendance à se séparer en plusieurs valeurs, et les auditeurs à ne pas être d'accords entre eux.

Nous illustrons le modèle MM dans la Figure 4.2. Pour chacun des deux descripteurs donnant les meilleurs résultats ($d_{onset}(\tau)$ et $d_{sim}(\tau)$ d'après [Peeters et al., 2012a]), nous détectons leurs deux pics principaux. Si l'un des deux pics appartient à l'intervalle $110 - 170$ bpm, nous classifions l'extrait dans la catégorie *Accord*, sinon nous le classifions dans la classe *Désaccord*. Le modèles obtenus sont notés respectivement MM-Onset et MM-Sim.

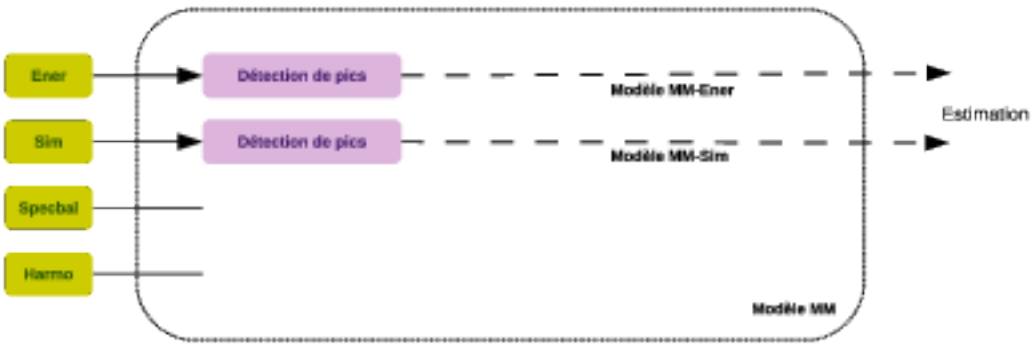

FIGURE 4.2 – *Schéma du modèle A (MM-onset et MM-sim).*

Pour ce modèle, τ représente les fréquences de la DFT. En effet, les descripteurs calculés en utilisant la DFT donnent de légèrement moins bon résultats lors d'une estimation de tempo que ceux issus de l'auto-corrélation, mais ils possèdent des pics principaux mieux définis et moins nombreux.

Nous montrons sur la Figure 4.3 un exemple de détection de pics sur $d_{onset}(\tau)$. La fonction d'onset est en bleu. Les croix rouges représentent les pics détectés, les deux gros points rouges sont les deux pics principaux. Le trait pointillé rouge vertical correspond au tempo prédominant (120 bpm). Les traits verts

pleins verticaux représentent l'intervalle [110−170] bpm. Si un pic se trouve dans cet intervalle, nous estimerons qu'il y a *Accord* entre les auditeurs. Dans le cas de la Figure 4.3, les deux pics principaux encadrent l'intervalle, nous estimons donc qu'il y a *Désaccord*.

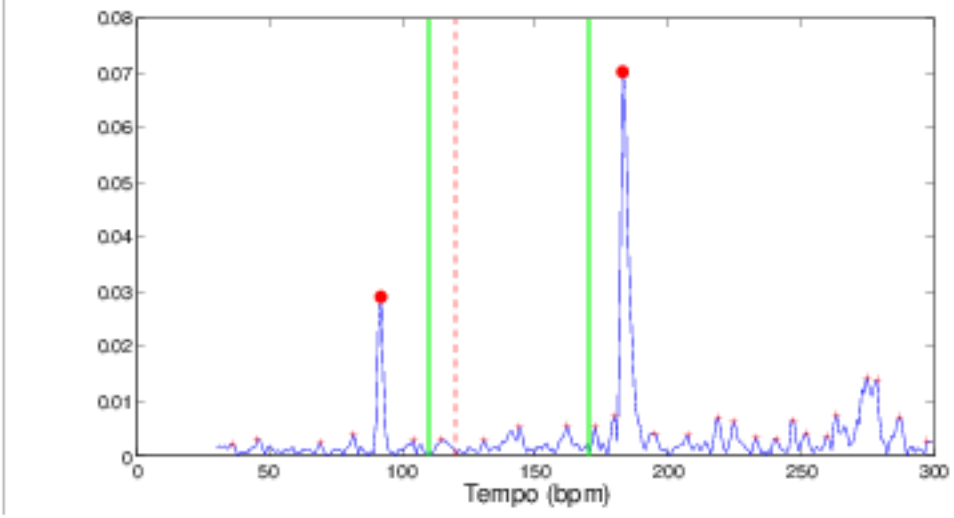

FIGURE 4.3 – *Exemple de détection de pics sur $d_{onset}(\tau)$. La fonction d'onset est en bleu. Les croix rouges représentent les pics détectés, les deux gros points rouges sont les deux pics principaux. Le trait pointillé rouge vertical correspond au tempo prédominant (120 bpm). Les traits verts pleins verticaux représentent l'intervalle [110 − 170] bpm. Si un pic se trouve dans cet intervalle, nous estimerons qu'il y a Accord entre les auditeurs. Dans le cas de la figure 4.3, les deux pics principaux encadrent l'intervalle, nous estimons donc qu'il y a Désaccord.*

Modèle B (Feature-GMM). Il s'agit du modèle le plus simple. Il est fortement inspiré de [Peeters et al., 2012a]. Dans cet article, les auteurs utilisent les quatre même descripteurs pour estimer le tempo à partir d'un modèle de régression GMM [2]. Nous faisons l'hypothèse que ces descripteurs sont aussi suffisants pour décrire l'accord entre auditeurs.

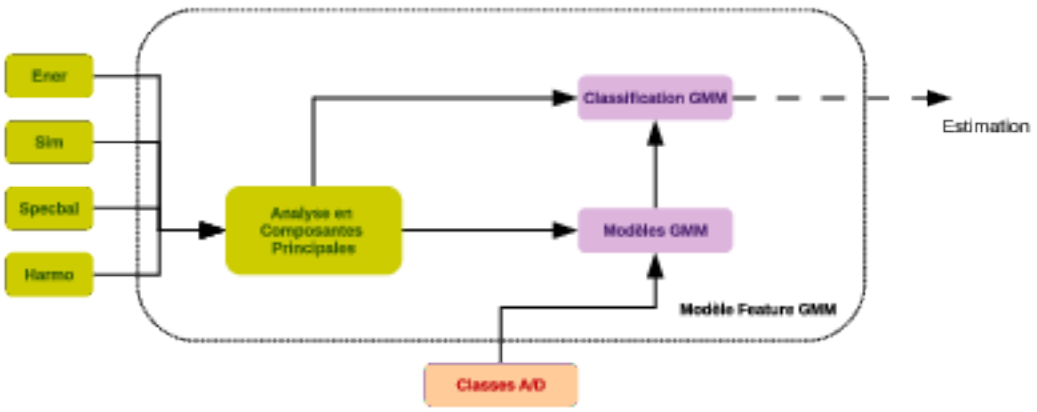

FIGURE 4.4 – *Schéma du modèle B (Feature-GMM).*

Nous illustrons le principe du modèle sur la Figure 4.4. Nous utilisons d'abord une Analyse en Composantes Principales ou « Principal Component

2. voir Annexe B pour la description des différentes méthodes de classification et de régression.

Analysis » (PCA) pour réduire la dimension des quatre vecteurs de 80 à 34 dimensions (nous gardons seulement les axes composant plus de 95 % de la variance). À partir de ce vecteur, nous entrainons deux Modèle de Mélange de Gaussiennes ou « Gaussian Mixture Model » (GMM) à 4 composantes avec une matrice de covariance pleine : un pour la classe *Accord* et un autre pour la classe *Désaccord*. Nous attribuons ensuite aux morceaux de musique inconnus la classe dont la probabilité à posteriori est la plus grande.

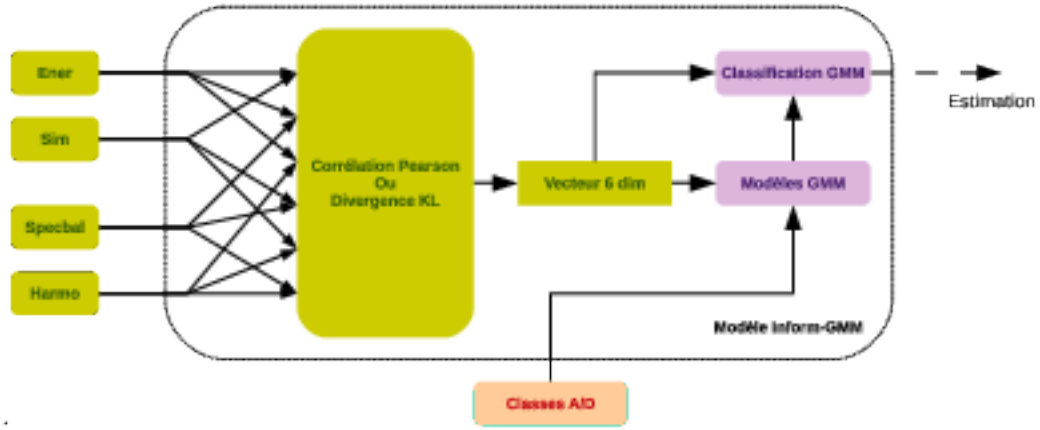

FIGURE 4.5 – *Schéma du modèle C (Inform-GMM).*

Modèle C (Inform-GMM). Dans ce modèle, nous nous intéressons aux corrélations entre les différents descripteurs. Nous rappelons que ces descripteurs représentent le signal audio sous différents points de vues. Notre hypothèse est que si deux vecteurs de descripteurs $d_i(\tau)$ possèdent la même information de périodicité alors les auditeurs vont être d'accord entre eux sur le tempo perceptif. Nous mesurons donc ici l'information partagée par les descripteurs :

$$\underline{C} = [c(d_{onset}(\tau), d_{sim}(\tau)),\ c(d_{onset}(\tau), d_{specbal}(\tau)), ...] \tag{4.1}$$

où c est une fonction qui mesure le partage d'informations entre deux descripteurs. Nous allons tester par la suite deux types de fonctions : la corrélation de Pearson et la divergence de Kullback-Leibler (KL) symétrisée. Nous utilisons ensuite deux GMMs entrainés sur le vecteur à 6 dimensions $\underline{C}$ pour estimer les classes *Accord* et *Désaccord*. Nous résumons ce modèle sur la Figure 4.5.

FIGURE 4.6 – *Schéma du modèle D (Tempo-GMM).*

Modèle D (Tempo-GMM et Tempo-SVM). Ce dernier modèle est basé sur la même hypothèse : si les quatre descripteurs partagent la même information alors il y aura plutôt *Accord* sur le tempo perceptif et sinon il y aura *Désaccord*. Dans ce modèle que nous avons illustré sur la Figure 4.6, plutôt que d'apprendre deux modèles GMM directement sur les descripteurs, nous allons le faire sur les quatre estimations de tempo obtenues par l'utilisation des quatre descripteurs pris indépendamment.

Nous estimons donc tout d'abord quatre tempos $\hat{t}_{ener}, \hat{t}_{sim}, \hat{t}_{spectral}, \hat{t}_{harmo}$ à partir des quatre descripteurs. Chaque algorithme d'estimation est une méthode de régression GMM. Ensuite nous utilisons le vecteur $[\hat{t}_{ener}, \hat{t}_{sim}, \hat{t}_{spectral}, \hat{t}_{harmo}]$ comme observation pour l'apprentissage de deux GMMs *Accord*/*Désaccord*.

Nous avons ensuite testé une seconde méthode de classification. Au lieu d'utiliser une classification GMM à partir des 4 tempos, nous utilisons une classification par SVM [3].

4.3 Évaluation

Nous utilisons pour nos entraînements et évaluations la base de données LEVY décrite dans la partie 3.2.

Protocole expérimental. Nous possédons un ensemble de données contenant des descripteurs et les annotations en classes *Accord*/*Désaccord*. Pour les trois derniers modèles (B, C, D) qui nécessitent une phase d'apprentissage, nous effectuons une validation croisée à cinq plis [4]. Nous obtenons donc un ensemble d'apprentissage sur lequel on va créer les deux GMM [5] à 4 gaussiennes (un pour chaque classe), et un ensemble de test, sur lequel nous estimons les probabilités d'appartenance à chaque classe. La probabilité à postériori la plus forte donne la classe estimée.

4.3.1 Résultats

Les résultats des modèles de classification sont présentés en terme de rappel moyen. Pour deux classes *Accord*/*Désaccord*, un rappel moyen de 50% équivaut à une estimation aléatoire.

Nous présentons les résultats de nos différents modèles dans la Figure 4.7. Les trois modèles MM-Onset, Feature-GMM et Inform-GMM (avec la corrélation Pearson) ne sont pas bons : ils ne catégorisent pas mieux qu'un classificateur aléatoire. Les deux modèles MM-Sim et Inform-GMM (utilisant la divergence de KL) sont encourageants et donnent des résultats un peu au dessus de l'aléatoire (57% et 56% respectivement). Les derniers modèles Tempo-GMM et Tempo-SVM obtiennent 70% et 75% de rappel moyen sur les deux classes *Accord* et *Désaccord*. Nous allons maintenant détailler les résultats pour chacun des modèles.

3. Nous avons obtenus les meilleurs paramètres $c = 1.59, \gamma = 0.001$ par recherche exhaustive sur une validation croisée à 5 plis.

4. La validation croisée à cinq plis consiste à séparer l'ensemble de données en cinq fois un ensemble d'apprentissage et un de test, qui contiendront respectivement $\frac{4}{5}$ des données et $\frac{1}{5}$ des données, de façon à ce que l'ensemble de test couvre toutes les données.

5. Dans notre étude, nous utilisons la bibliothèque Matlab développée par [Calinon, 2009] pour l'implémentation des modèles GMM.

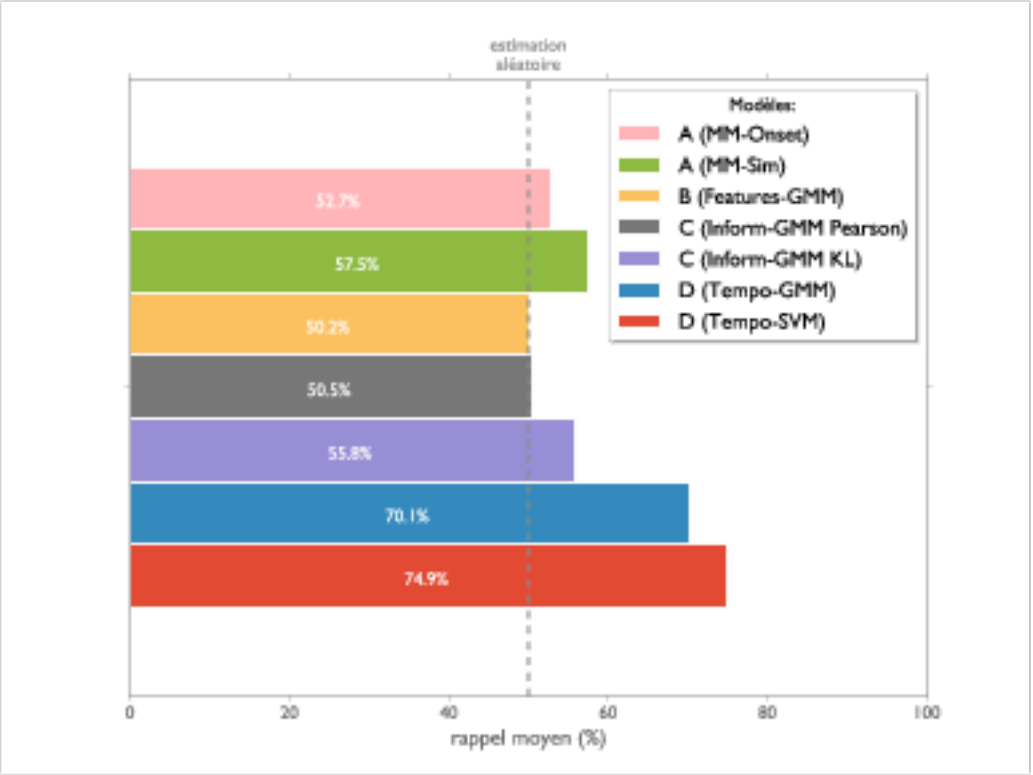

FIGURE 4.7

4.3.2 Analyse du modèle A (MM-onset et MM-sim)

Vérification du modèle de résonance de [Moelants et al., 2004]. [Moelants et al., 2004] supposent qu'il existe un tempo préférentiel autour de 120 bpm. Ils modélisent donc toutes les annotations de leur corpus par une fonction de résonance [Noorden et al., 1999] :

$$R = \frac{1}{\sqrt{(f_0^2 - f^2)^2 + \beta f^2}} - \frac{1}{\sqrt{{f_0}^4 - f^4}} \tag{4.2}$$

Nous avons testé cette hypothèse pour notre corpus. Nous montrons sur la Figure 4.8 l'histogramme de toutes les annotations selon leur tempo. La courbe rouge est le modèle de résonance décrit précédemment, dont les paramètres f_0 et β ont été trouvés par la méthode des moindres carrés.

Pour notre corpus LEVY, le tempo de résonance que nous avons trouvé est de 80 bpm, qui est donc assez différent des résultats obtenus par [Moelants et al., 2004]. Nous remarquons même un creux autour de 120 bpm, ce qui est contraire à leur hypothèse d'un tempo préférentiel à 120 bpm. Nous proposons plusieurs explications à ce résultat :

— Notre **corpus** est très différent du leur comme le montre la Figure 4.9. Le corpus de McKinney & Moelants est composé de morceaux également répartis en musique classique, country, dance, hip-hop, jazz, latin, reggea, rock/pop et soul. Le notre est composé à plus de 50 % de musique rock/pop, de presque 10 % de country et soul, de seulement 5 % de latin et reggea, et de pas ou peu de classique, dance, hip-hop et jazz.

— Les **auditeurs** n'ont pas les mêmes profils. Dans l'expérience de McKinney & Moelants, les 33 sujets ont une éducation musicale de 7 années en

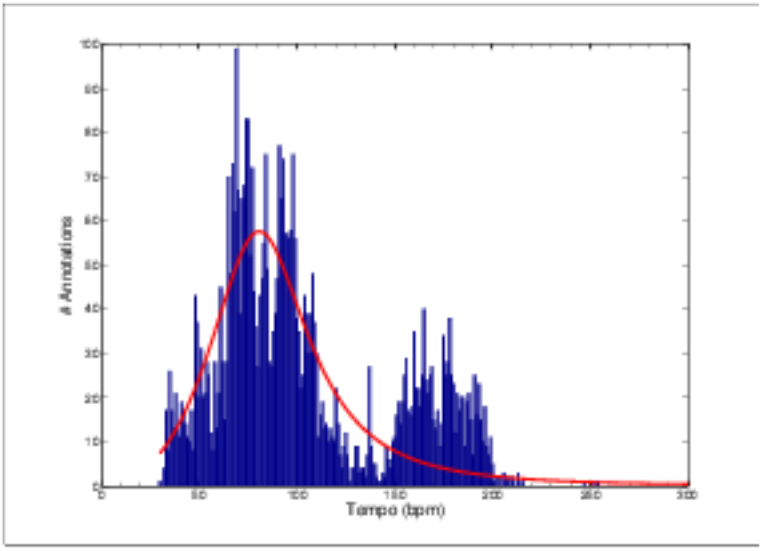

FIGURE 4.8 – *Modèle de tempo prédominant de [Moelants et al., 2004]. L'histogramme de toutes les annotations du corpus (des classes Accord et Désaccord) selon leur tempo est tracé en bleu. La gaussienne rouge représente le modèle de résonance de McKinney & Moelants.*

moyenne. Nous estimons que, dans notre cas d'expérience web, la majorité des auditeurs n'a pas nécessairement reçu d'éducation musicale.

— Enfin, les **protocoles** expérimentaux de création des corpus diffèrent énormément. Le notre est une expérience sur le web, pratiquement sans contrôle des auditeurs, tandis que le corpus de McKinney et Moelants a été créé dans de meilleures conditions.

L'hypothèse d'un tempo préférentiel autour de 120 bpm n'est donc pas réfutée, mais elle est seulement invalide sur notre corpus spécifique.

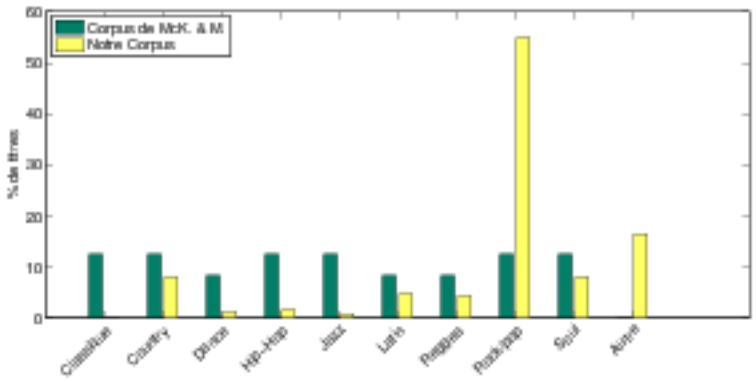

FIGURE 4.9 – *Comparaison de notre corpus avec celui de McKinney & Moelants en terme de genres musicaux. Nous représentons la distribution des genres musicaux du corpus de McKinney et Moelants, et en jaune la distribution des genres de notre corpus.*

Analyse des résultats des modèles MM-Onset et MM-Sim. Les deux modèles utilisant les hypothèses de [Moelants et al., 2004] ont un rappel moyen de 52.7%

(MM-Onset) et de 57.5% (MM-Sim), c'est-à-dire juste au niveau de l'aléatoire pour MM-Onset et juste au dessus de l'aléatoire pour MM-Sim. Ces résultats ne sont donc pas satisfaisant sur notre corpus.

Dans les expériences de [Moelants et al., 2004], l'hypothèse du tempo prédominant de 120 bpm fonctionnait bien sur l'une de leurs expériences, mais mal sur les deux autres. Les auteurs concluent que la perception du tempo peut être modifiée par d'autres paramètres comme la structure du morceau ou la répartition des accents. Ce modèles ne sont donc pas utilisables sur tous les types de corpus, et pas sur le notre en particulier. Nous pouvons ajouter quelques explications :

— le tempo préférentiel de 120 bpm ne correspond pas à notre corpus (Figure 4.2), dont la résonance est à 80 bpm. Nous avons donc essayé de changer l'intervalle signifiant l'accord à $[70-90]$ bpm au lieu de $[110-170]$ bpm mais cela n'a malheureusement pas permis d'améliorer les résultats, que nous montrons dans le Tableau 4.1.

— les différents descripteurs ne sont peut être pas appropriés pour ce type de modélisation. Ils le sont plus pour une estimation de tempo, car ils possèdent souvent beaucoup de pics à des multiples du tempo principal. Cela rend l'estimation de tempo plus robuste, mais cela gêne un peu notre détection de pic. Il est à noter que nous n'avons pas utilisé les descripteurs de balance spectrale et d'harmonicité, vu leur faibles performances en estimation de tempo seul (voir [Peeters et al., 2012a]).

TABLE 4.1 – *Résultats obtenus avec les modèles MM-Onset et MM-Sim en changeant l'intervalle de résonance à* $[70-90]$ *bpm.*

	*A*ccord	*D*ésaccord	Rappel moyen
Modèle MM-Ener	14.9%	87.8%	51.4%
Modèle MM-Sim	1.5%	88.7%	45.1%

4.3.3 Analyse du modèle B (Feature-GMM)

Ce modèle donne pour résultat un rappel moyen de 50.2%, c'est-à-dire l'équivalent d'une estimation aléatoire.

Nous avons essayé ce modèle car il donnait de bons résultats en estimation de tempo perceptif [Flocon-Cholet, 2012 ; Peeters et al., 2012a]. Dans ces études basées sur le même corpus de LEVY, les auteurs montrent qu'il est possible d'estimer de façon fiable le tempo perceptif à partir des quatre descripteurs comme entrée d'une Régression-GMM). La différence de nos travaux par rapport aux leurs est la sélection des morceaux du corpus. Dans [Flocon-Cholet, 2012 ; Peeters et al., 2012a], les auteurs sélectionnent les morceaux pour lesquels les auditeurs sont d'accords entre eux et suppriment donc du corpus tous les exemples ambigus, alors que notre méthode de sélection cherche justement les cas ambigus.

Le manque d'exemples d'apprentissage peut aussi être la cause du non-fonctionnement de cette méthode. Nous ne possédons en effet que 250 morceaux répartis en deux classes. L'information que l'on cherche (*A*ccord/*D*ésaccord des utilisateurs) n'est donc pas modélisable directement par les descripteurs, c'est pourquoi nous nous sommes intéressés à l'information partagée par ces descripteurs dans les deux modèles suivants.

4.3.4 Analyse du modèle C (Inform-GMM)

Nos deux modèles ont des rappels moyen de 50.5 % (Inform-GMM Pearson) et de 55.8 % (Inform-GMM KL), c'est-à-dire au niveau de l'aléatoire avec la corrélation de Pearson, et juste au-dessus de l'aléatoire avec la divergence de Kullback-Leibler symétrisée.

Nous montrons dans la Figure 4.10 les quatre descripteurs dans le cas *Accord* (à gauche) et dans le cas *Désaccord* (à droite). Nous voyons sur la figure de gauche que les trois premiers descripteurs possèdent bien leurs pics au même tempo. Par contre, dans la figure de droite, nous voyons que les descripteurs ont des pics à différents niveaux métriques ($d_{onset}(\tau)$ possède deux fois plus de pics que $d_{specbal}(\tau)$, qui possède deux fois plus de pics que $d_{sim}(\tau)$). Notre hypothèse semble donc correcte :si les descripteurs ne partagent pas les mêmes informations alors le tempo perçu sera ambigu. Nous pensons que les faibles résultats de ces modèles peuvent provenir des fonctions modélisant le partage d'information entre les descripteurs, qui ne mesurent pas exactement ce que l'on veut. La corrélation de Pearson, par exemple, prend en compte toute l'énergie commune entre deux descripteurs, même celle hors des pics principaux. Nous pouvons donc observer des corrélations fortes même pour deux descripteurs n'ayant aucun pics en commun. Nous proposons donc un dernier modèle (Tempo-GMM) pour modéliser les corrélations entre les descripteurs.

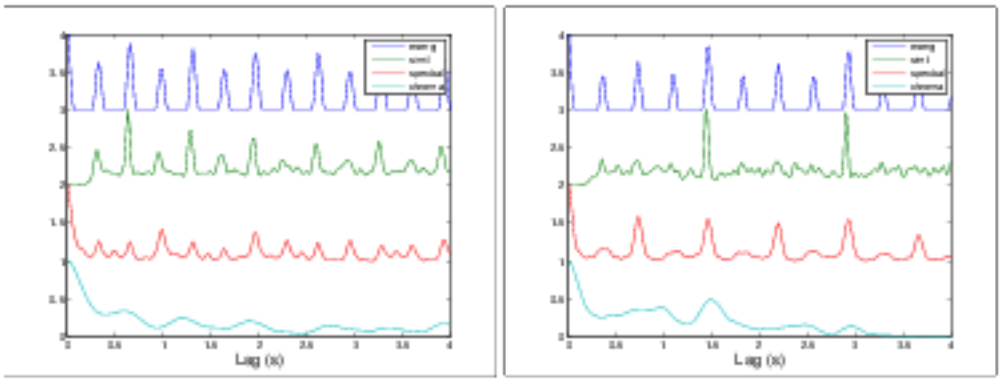

FIGURE 4.10 – *Gauche : De haut en bas, les descripteurs d'onset (noté energ dans la légende), de similarité, de balance spectrale et d'harmonicité (noté chroma), pour un morceau de la classe Accord. Droite : La même chose pour un morceau de la classe Désaccord.*

4.3.5 Analyse des modèles D (Tempo-GMM et Tempo-SVM)

Ces modèles donnent 70.1 % et 74.9% de rappel moyen. Ces résultats sont assez prometteurs et nous les analysons un peu plus en détails.

Nous présentons dans le Tableau 4.2, les résultats détaillés pour les modèles Tempo-GMM et Tempo-SVM Nous remarquons que le rappel moyen du SVM est meilleur de 5 % comparé à la classification par GMM. Par contre, les rappels des deux classes sont très déséquilibrés (87% et 44%), alors qu'ils sont beaucoup plus équilibrés avec un GMM (74% et 67%). C'est pourquoi nous préférons la méthode de classification par GMM.

Nous montrons sur la Figure 4.11 les relations entre les tempos estimés $\hat{t}_{ener}$, $\hat{t}_{sim}$ et $\hat{t}_{specbal}$. Les signes + rouges représentent les données de la classe *Accord* et les x bleues celles de la classe *Désaccord*. Nous observons bien que les éléments

TABLE 4.2 – *Résultats détaillés par classe pour les modèles Tempo-GMM et Tempo-SVM*

	Accord	Désaccord	Rappel moyen
Modèle Tempo-GMM	73.7%	66.5%	70.1%
Modèle Tempo-SVM	87.4 %	44.4 %	74.9 %

de la classe *Accord* sont majoritairement situés sur la diagonale (ce qui signifie que les tempos estimés par les différents descripteurs sont les mêmes), et que les éléments de la classe *Désaccord* sont majoritairement situés hors de la diagonale. Ces figures valident donc notre hypothèse : si l'information est partagée entre les descripteurs, les utilisateurs seront d'accord sur la perception du tempo.

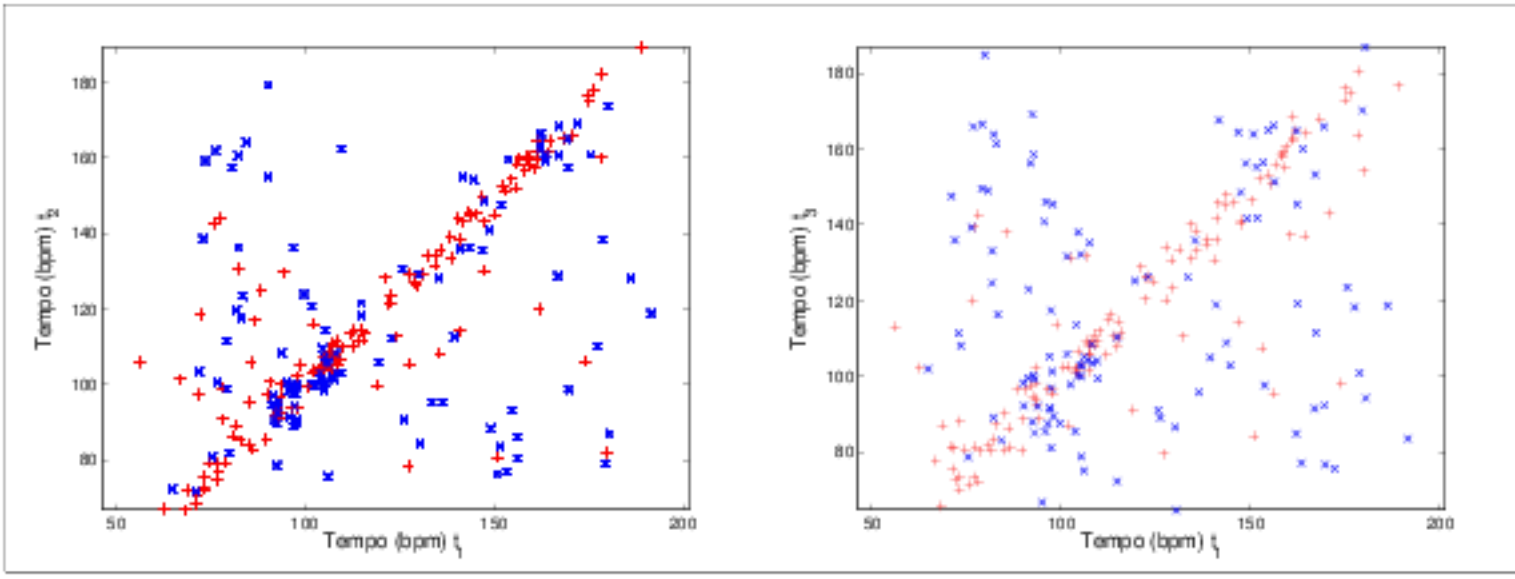

FIGURE 4.11 – *Figure de gauche : relation entre les tempos estimés* $t_1 = \hat{t}_{ener}$ *et* $t_2 = \hat{t}_{sim}$. *Figure de droite : relation entre les tempos estimés* $t_1 = \hat{t}_{ener}$ *et* $t_3 = \hat{t}_{specbal}$. *Les signes + rouges représentent les données de la classe Accord et les* $\times$ *bleues celles de la classe Désaccord.*

Nous notons que, comme dans [Flocon-Cholet, 2012 ; Peeters et al., 2012a], l'harmonicité est un descripteur qui fonctionne peu pour l'estimation du tempo perceptif. Nous pouvons le constater dans le Tableau 4.3 qui résume les performances des régressions GMM pour l'estimation de tempo individuelle sur chaque descripteur. C'est pour cela que nous n'avons pas tracé les relations entre $t_4 = \hat{t}_{harmo}$ et les autres $t_i, i = 1..3$. Les résultats du Tableau 4.3 ont été obtenus par une validation croisée à 5 plis lors de la phase d'estimation du tempo sur chaque descripteur. Nous y remarquons que les quatre descripteurs ont des performances décroissantes ($d_{onset} > d_{sim} > d_{specbal} > d_{harmo}$). Nous pouvons donc faire l'hypothèse que pour être efficaces en modélisation de tempo perceptif, les descripteurs doivent déjà avoir de bonne performances individuelles d'estimation de tempo dans le cas où les auditeurs sont d'accords entre eux.

TABLE 4.3 – *Performances individuelles pour chaque descripteur des régressions GMM pour l'estimation de tempo dans le cas où les auditeurs sont d'accords entre eux.*

Descripteur	Tempo trouvé
onset	85.9 %
sim	70.3 %
specbal	64.3 %
harmo	38.2 %

4.4 Conclusions

Dans ce chapitre, nous avons étudié la prédiction de l'accord entre auditeurs sur la perception du tempo. En effet, la perception du tempo peut ne pas être partagée pour un même morceau, les auditeurs pouvant se focaliser sur différents niveaux métriques. Cette perception dépend bien évidemment de l'auditeur mais nous faisons l'hypothèse que certaines caractéristiques du signal audio peuvent faciliter ou non la perception partagée du tempo. Pour étudier cela, nous avons utilisé les descripteurs de [Peeters et al., 2012a] (onset, similarité à court-terme, balance spectrale et changement d'harmonicité) pour modéliser les différentes caractéristiques du signal audio. Nous avons ensuite proposé quatre modèles de prédiction de l'*Accord*/*Désaccord* entre auditeurs sur base de ces descripteurs.

Un de ces modèles est basé sur les expériences perceptives de [Moelants et al., 2004]. Nous expliquons les mauvais résultats obtenus avec ce modèle par les différences de corpus, comme la distribution en genre, les profils des auditeurs et le protocole expérimental. Pour le corpus LEVY, le tempo de résonance se situe autour de 80 bpm au lieu de 120 bpm dans [Moelants et al., 2004].

Notre meilleure prédiction 75% repose sur un modèle utilisant la cohérence des quatre tempos estimés individuellement à l'aide des quatre descripteurs. Ce modèle valide notre hypothèse de départ : si l'information de tempo est partagée entre les descripteurs, les utilisateurs auront tendance à être d'*Accord* sur la perception du tempo.

Publications associées

Peeters, G. et Marchand, U. (2013). « Predicting agreement and disagreement in the perception of tempo ». *Proceedings of the 10th International Symposium on Computer Music Modeling and Retrieval*, p. 253–266.

— (2014b). « Predicting agreement and disagreement in the perception of tempo ». *Sound, Music, and Motion*. Springer, p. 313–329.

Chapitre 5

Estimation des déviations systématiques

Contenu

5.1 Introduction

5.1.1 Le swing

La musique est rarement jouée exactement telle qu'écrite sur sa partition. La plupart des interprétations d'une même musique sont différentes, c'est ce qui participe, entre autre, à la richesse du répertoire musical. Les musiciens dévient donc souvent de la partition afin de la rendre plus vivante. Ces déviations sont loin d'être aléatoires, et sont précisément choisies par les musiciens en fonction d'une multitude de paramètres comme, par exemple, leur connaissance du compositeur, l'époque à laquelle a été écrite l'œuvre ou leur propre sensation de comment elle doit être jouée.

Si on se focalise sur les aspects temporels de la musique, on peut citer plusieurs exemples de déviations systématiques. Il peut y avoir des changements de tempo locaux pour permettre plus d'expressivité et de naturel comme le *rubato* qui est particulièrement présent dans la musique romantique. *Tempo rubato*, qui signifie littéralement « temps volé », permet de jouer avec une rythmique plus expressive et plus libre que le cadre rigide de la partition. Un autre exemple provient de l'époque baroque, avant l'apparition des premiers piano forte. Pour accentuer certaines notes, les joueurs de clavecin, qui n'avaient pas la possibilité de jouer avec des nuances, pouvaient retarder certaines notes isolées, afin de les accentuer. Un troisième exemple est le swing, présent dans la musique jazz.

Le swing fait référence à un niveau temporel très précis : celui de la croche. Dans un morceau avec du swing, deux croches consécutives vont être jouées avec des durées inégales, selon un motif long-court. Le ratio de swing est le rapport entre la durée de la première croche (longue), et celle de la seconde croche (courte). Les ratios les plus présents dans la musique sont 1 : 1 (pas de swing), 2 : 1 (sensation ternaire), 3 : 1 (swing très prononcé). Ces différents ratios sont représentés sur la Figure 5.1. Pour [Laroche, 2001], le swing est défini comme le petit décalage du deuxième et du quatrième quart de temps (pour lui, un temps équivaut à une blanche). Il est aussi a noter que plus le tempo est rapide, plus les ratios de swing vont se rapprocher de 1 :1 [Friberg et al., 2002]. Le swing est une convention dans certains styles de jazz. Il peut ne pas être indiqué sur la partition, il est implicite et à l'appréciation du musicien. Il peut donc varier considérablement entre plusieurs interprètes, ou même au sein de la même performance. [Honing et al., 2008] ont montré cependant que les batteurs de jazz professionnels ont un contrôle extrêmement précis sur le ratio de swing qu'ils cherchent à atteindre.

FIGURE 5.1 – *Trois ratios de swing différents. De gauche à droite, les rythmes représentants les ratios de swing* 1 : 1 *(pas de swing),* 2 : 1 *(sensation ternaire ou « triple feel ») et* 3 : 1 *(swing très prononcé ou « hard swing »).*

5.1.2 État de l'art

Plusieurs études se sont intéressées au ratio de swing et particulièrement à sa relation au tempo. [Friberg et al., 2002] étudient le rapport des durées de la croche longue et de la croche courte dans des enregistrements de musique jazz. Les auteurs montrent que pour un tempo faible, le ratio de swing peut atteindre $3,5:1$ et que ce ratio décroît avec le tempo, pour atteindre $1:1$ à tempo élevé. Il montrent de plus que le ratio de swing varie linéairement avec le tempo. Ces résultats sont contredits par les travaux de [Honing et al., 2008] qui ne trouvent pas de relation linéaire entre le ratio de swing et le tempo. Plus généralement, [Desain et al., 1994] s'intéressent à la relation entre micro-déviations systématiques et tempo. Ils étudient pour cela la variance d'histogrammes inter-onset exprimée en échelle logarithmique, dans des enregistrements de musique classique possédant beaucoup de rubato. Ils montrent qu'en ce qui concerne les micro-timings, il n'existe pas de relation proportionnelle entre le tempo et les déviations systématiques de rythme.

Les travaux de [Friberg et al., 2002] montrent aussi que la durée de la croche courte ne descend pas en dessous de 100 ms, quel que soit le tempo. Cela suggère qu'il y a peut-être une limite physique au ratio de swing.

Plusieurs méthodes d'estimation automatique du ratio de swing existent dans la littérature. [Gouyon et al., 2003a] proposent un outil de modification du swing. La détection du ratio de swing se fait par deux méthodes. La première se base sur la moyenne du deuxième pic de l'histogramme d'intervalles inter-onset (le premier pic correspondant à l'auto-corrélation du signal avec lui même). Ce deuxième pic correspond à la durée de la croche courte. Ils estiment alors le ratio de swing comme $\frac{d_c+\Delta}{d_c-\Delta}$, d_c étant la durée de la croche théorique et Δ le décalage du pic de la croche courte par rapport à celui de la croche théorique. La deuxième méthode compare l'histogramme d'intervalles inter-onset à des modèles d'histogrammes prédéfinis pour différents rapports de swing et cherche celui qui correspond le mieux. Les auteurs proposent ensuite de modifier ce rapport de swing en utilisant une méthode basée sur le vocodeur de phase.

[Laroche, 2001] propose une estimation jointe du tempo, du premier temps de la mesure et du swing. Pour cela il détecte tout d'abord les instants où l'énergie d'une des bandes de fréquence du signal augmente très rapidement. Il compare les histogrammes de ces instants à un ensemble de fonctions de probabilités représentant conjointement tempo, instant du premier temps et ratio de swing. Une recherche exhaustive sur l'ensemble des triplets (tempo, instant du premier temps, swing) lui permet de trouver celui qui a la meilleure vraisemblance.

Parallèlement à nos travaux, [Dittmar et al., 2015] ont proposé deux méthodes d'estimation automatique des ratios de swing de la cymbale ride dans des enregistrements de jazz. La première est basée sur de la détection de pics dans la fonction de détection d'onset. La seconde est basée sur une représentation en log-temps (log-lag auto-correlation function) normalisée par le tempo annoté du morceau. Chaque morceau est ensuite comparé à une base de prototypes dans cette représentation et la meilleure correspondance est gardée.

5.1.3 Plan du chapitre

Dans ce chapitre, nous proposons plusieurs méthodes d'estimation automatiques du ratio de swing dans un extrait musical. Dans la partie 5.2, nous présentons les trois modèles d'estimation automatique du swing que nous avons

retenus. Dans la partie 5.3, nous présentons les différentes expériences nous permettant d'évaluer nos modèles, nous analysons et discutons les résultats.

5.2 Modèles d'estimation du swing

Dans cette partie, nous présentons trois méthodes d'estimation automatique du swing. Deux (ACF et PIC) sont des contributions originales, la troisième (LLACF) a été proposée par Dittmar et al.[Dittmar et al., 2015] parallèlement à nos travaux. Nous n'y avons fait que quelques modifications mineures afin de la comparer à nos méthodes. Toutes ces méthodes sont basées sur une représentation intermédiaire : l'auto-corrélation de l'onset du signal, dont les propriétés sont particulièrement intéressantes pour l'estimation automatique de déviations systématiques.

Nous présenterons d'abord dans la partie 5.2.1 l'extraction de l'auto-corrélation depuis le signal et en quoi cette représentation est intéressante pour l'étude de déviations systématiques. Puis nous nous intéresserons aux trois méthodes : ACF (partie 5.2.2), LLACF (partie 5.2.3) et PIC (partie 5.2.4).

5.2.1 Extraction de l'auto-corrélation

Méthode. Tout d'abord, nous calculons la fonction d'onset $o(t)$ de [Ellis, 2007] à partir du signal audio. Elle est ensuite découpée en trames de durée 16 secondes [1], avec un pas de 1 seconde [2] pour obtenir $o_u(t)$. Nous calculons sur chaque trame u l'auto-corrélation $R_u(\tau)$ normalisée par l'auto-corrélation à l'instant 0 de la fonction d'onset $o_u(t)$.

$$R_u(\tau) = \frac{1}{R_u(0)} \sum o_u(t) o_u(t-\tau) \tag{5.1}$$

Propriétés. La fonction d'auto-corrélation $R(\tau)$ de $o(t)$ permet de mettre en évidence les différents niveaux métriques d'un motif rythmique. On peut y voir toutes les périodicités temporelles présentes. Dans le cas d'un motif binaire simple composé de noires et de croches, la fonction d'auto-corrélation possèdera des pics principaux situés à la durée de la croche et à la durée de la noire (comme c'est le cas dans la Figure 5.2 à gauche).

Les déviations systématiques sont elles aussi visibles dans la fonction d'auto-corrélation. Dans le cas d'un motif simple composé d'une croche courte suivi d'une croche longue, la fonction d'auto-corrélation possède des pics principaux situés aux durées de la croche courte, de la croche longue et de la noire (comme dans la Figure 5.2 à droite).

Nous montrons un exemple sur la Figure 5.2. Pour un signal sans swing (à gauche), on voit un pic représentant le tactus (durée de la noire) à 0,45 secondes et un pic représentant le tatum (durée de la croche) à 0,22 secondes. Pour un signal avec swing (à droite), on voit toujours le pic représentant la durée de la noire à 0,45 secondes, mais le pic représentant la durée de la croche est scindé

1. Il est nécessaire de prendre une fenêtre longue afin d'augmenter la précision de l'algorithme. En effet, il n'y a pas forcement du swing sur chaque temps. Pour certains morceau, l'information de swing est assez disséminée, d'où la nécessité d'augmenter la taille de la fenêtre d'analyse.

2. Le pas de 1 seconde a été choisi arbitrairement sachant que le ratio de swing ne varie pas ou peu au cours d'un morceau. Étant donnée la taille de la fenêtre d'analyse, prendre un pas plus court n'aurait pas beaucoup de sens.

en deux pics : un pour la durée de la croche courte à 0,15 sec. et un pour la durée de la croche longue à 0,30 sec. L'objectif des différents modèles d'estimation automatique du swing va être de distinguer ces deux cas.

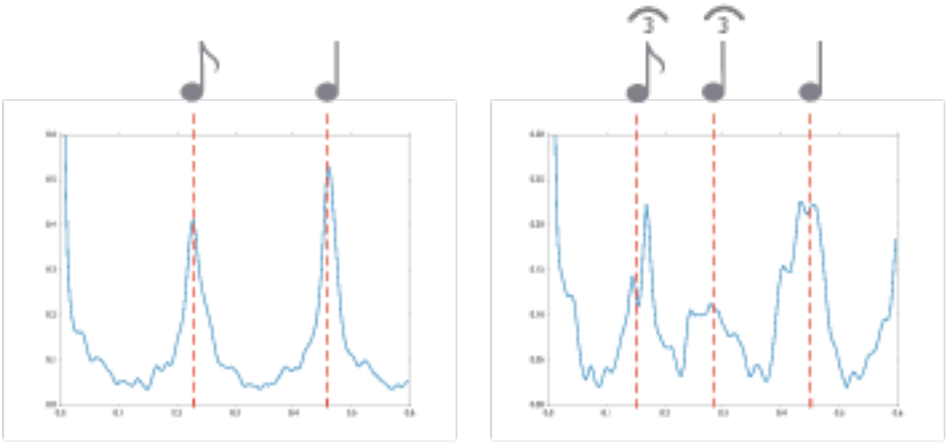

FIGURE 5.2 – *À gauche : fonction d'auto-corrélation d'un signal sans swing ('disco.00020.wav', première trame). À droite : fonction d'auto-corrélation d'un signal avec swing ('jazz.00099.wav', première trame).*

L'auto-corrélation sous diverses formes (log-lag auto-corrélation chez [Dittmar et al., 2015], inter-onset-histogramme chez [Gouyon et al., 2003a]) est très utilisée pour l'estimation automatique du swing ou plus généralement pour l'estimation de la métrique [Quinton et al., 2015], étant donnée qu'elle possède un pic pour chaque niveau métrique important. Elle est particulièrement adaptée à l'estimation de déviations systématiques (donc à l'estimation du swing) mais beaucoup moins adaptée pour repérer des déviations exceptionnelles. En effet, sur une longue fenêtre d'analyse (16 secondes dans notre cas), les rythmes se répétant sur toute la durée de la trame seront représentés par des pics bien défini, alors que les rythmes exceptionnels ne le seront pas. La robustesse de la fonction d'auto-corrélation est liée à la taille de la fenêtre d'analyse et à la précision temporelle des rythmes joués. Dans notre cas, cette précision est relativement faible (le tempo n'est pas parfaitement constant, et le ratio de swing varie beaucoup au sein d'un même extrait), cela nous force donc à prendre une fenêtre d'analyse grande. Prendre une fenêtre plus courte permettrait en théorie de voir des déviations moins systématiques que le swing, si l'enregistrement était suffisamment précis temporellement.

5.2.2 Modèle ACF (Auto-Correlation Function)

Ce premier modèle a pour objectif d'estimer un ratio de swing à partir d'une fonction d'auto-corrélation $R_u(\tau)$. Il est schématisé dans la Figure 5.3. L'idée de ce modèle est de comparer chaque $R_u(\tau)$ à une série de $R_{ideal}(\tau, T, s)$ représentant des prototypes d'auto-corrélation de signaux ayant un tempo T et un ratio de swing s fixés. Cette méthode permet de s'affranchir du tempo en estimant conjointement le tempo et le ratio de swing. Cette idée d'une estimation conjointe n'est pas nouvelle, il a déjà été proposé d'estimer conjointement plusieurs niveaux métriques ([Laroche, 2001] swing et tempo, [Peeters et al., 2011] tempo et mesure).

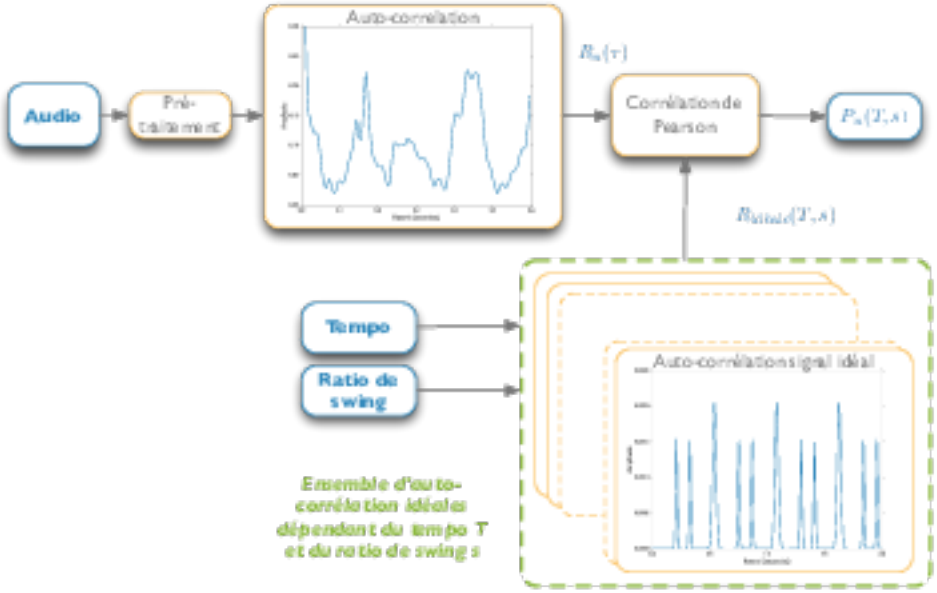

FIGURE 5.3 – *Schéma du modèle 'adaptation de motifs'.*

Base de prototypes. Nous avons créé un ensemble de prototypes d'auto-corrélations idéalisés, pour des tempos allant de 40 à 300 bpm (100 valeurs espacées logarithmiquement) et des ratios de swing allant de 1 à 4 (avec un pas de 0.05). Un exemple d'auto-corrélation idéalisée est montrée dans la Figure 5.4. Une fonction d'auto-corrélation prototype pour le tempo T et le swing s est notée $R_{\text{ideal}}(\tau, T, s)$. Le retard de cette fonction d'auto-corrélation est noté τ.

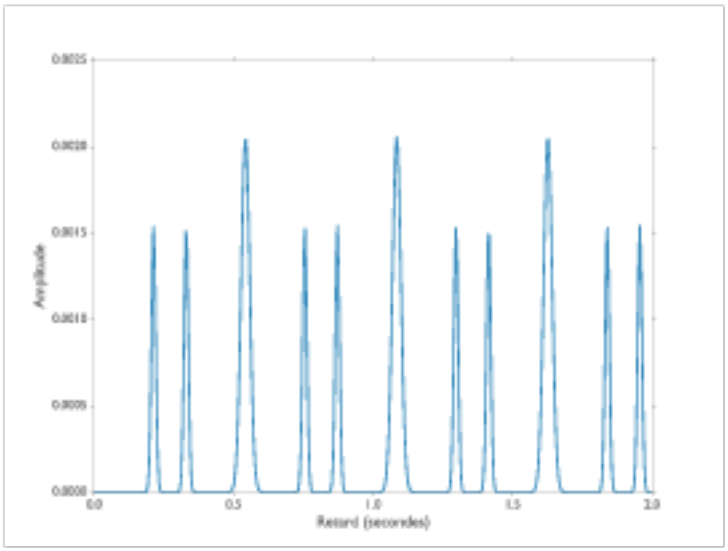

FIGURE 5.4 – *Une auto-corrélation idéalisée pour un tempo $T = 111bpm$ et un ratio de swing $s = 1.6$.*

Comparaison. Notre approche consiste à comparer l'auto-corrélation extraite à celle de tous les prototypes et de sélectionner le tempo et le ratio de swing qui

correspondent le mieux. Chacune de ces fonctions d'auto-corrélation est comparée à toutes les fonctions de la base, grâce au coefficient de corrélation de Pearson (comme proposé par [Dittmar et al., 2015]).

Nous obtenons pour chaque trame u la probabilité $P_{\mathrm{u}}(T,s)$ que la trame ait un ratio de swing s et un tempo T.

$$P_{\mathrm{u}}(T,s) = \frac{\sum_\tau \left(R_{\mathrm{u}}(\tau) - \overline{R_{\mathrm{u}}(\tau)}\right)\left(R_{\mathrm{ideal}}(\tau,T,s) - \overline{R_{\mathrm{ideal}}(\tau,T,s)}\right)}{\sqrt{\sum_\tau \left(R_{\mathrm{u}}(\tau) - \overline{R_{\mathrm{u}}(\tau)}\right)^2}\sqrt{\sum_\tau \left(R_{\mathrm{ideal}}(\tau,T,s) - \overline{R_{\mathrm{ideal}}(\tau,T,s)}\right)^2}}$$

5.2.3 Modèle LLACF (Log-Lag Auto-Correlation Function)

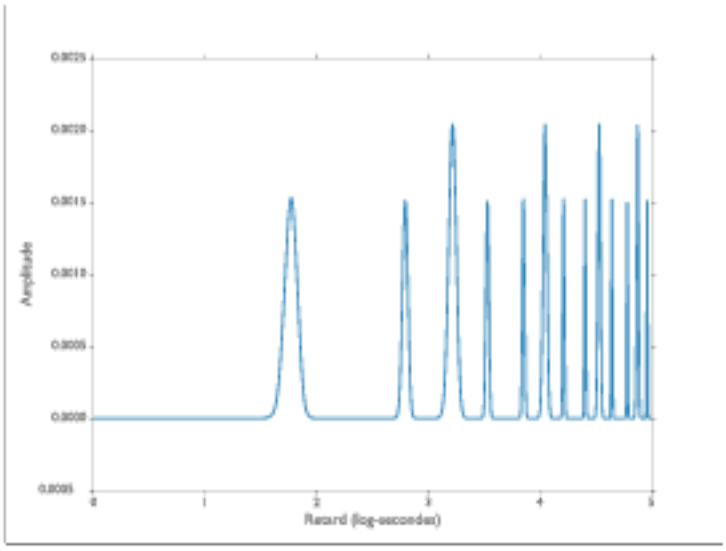

FIGURE 5.5 – *Une auto-corrélation en log-retard idéalisée (204bpm, ratio de swing 2,4).*

Cette méthode a été publiée en même temps que nos travaux. [Dittmar et al., 2015] proposent d'utiliser la fonction d'auto-corrélation exprimée en retard logarithmique $R_{\mathrm{u}}(\log\tau)$. L'avantage de cette représentation est qu'un changement de tempo correspond à un simple décalage du retard τ. Cet avantage est utilisé pour rendre la représentation indépendante du tempo. Ceci est obtenu en décalant la fonction d'auto-corrélation proposée grâce à la valeur du tempo annoté.

De même que dans la méthode précédente, nous allons créer une base de prototypes idéaux, dont un exemple est montré dans la Figure 5.5.

Pour chaque trame de signal, on obtient les $R_{\mathrm{u}}(\log\tau)$ à partir des $R_{\mathrm{u}}(\tau)$ en re-échantillonant les τ sur un axe logarithmique[3]. Chaque $R_{\mathrm{u}}(\log\tau)$ est ensuite comparé à la base de prototypes au moyen de la corrélation de Pearson comme précédemment.

5.2.4 Modèle PIC (Adaptation aux pics)

La troisième méthode est toujours basée sur l'auto-corrélation du signal. Elle est résumée par la Figure 5.6. Nous en décrivons les étapes dans la partie suivante. Elle a fait l'objet de la publication [Marchand et al., 2015b].

3. Ce re-échantillonnage est présenté dans l'annexe A.

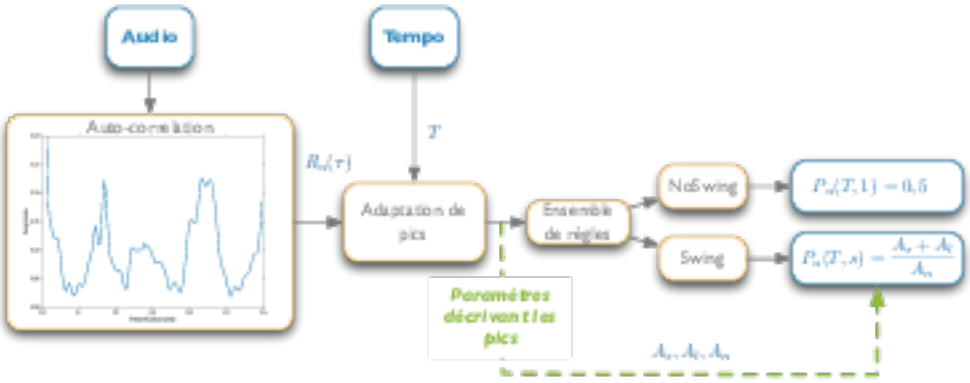

FIGURE 5.6 – *Schéma du modèle PIC.*

Tout d'abord l'auto-corrélation ainsi qu'une valeur de tempo sont fournies à un algorithme d'adaptation aux pics (partie 5.2.4.1). Il en sort des paramètres décrivant les pics de la fonction d'auto-corrélation. À partir de ces paramètres, nous appliquons un ensemble de règles (décrites dans la partie 5.2.4.2) pour détecter s'il y a du swing ou non. Afin d'homogénéiser notre modèle avec les deux autres et de pouvoir les comparer entre eux, nous en déduisons finalement une matrice de probabilités $P_u(T, s)$.

Cette méthode est similaire aux travaux de [Gouyon et al., 2003a], mais possède quelques différences. Notre représentation est basée sur une fonction continue de la fonction d'onset du signal, alors que Gouyon se base sur une version discrétisée (l'histogramme des intervalles inter-onset). De plus, notre recherche des pics représentant les croches de swing est plus robuste car elle est effectuée à la fois pour la croche courte et pour la croche longue, alors que Gouyon ne s'intéresse qu'à la durée de la croche courte. Cette méthode est de surcroit différente de la première méthode proposée par Dittmar [Dittmar et al., 2015], qui consiste à chercher des pics directement dans la fonction d'onset. Nous les cherchons dans une représentation plus stable : l'auto-corrélation de la fonction d'onset.

5.2.4.1 Adaptation de pics

Cette partie décrit comment, à partir du signal d'auto-corrélation $R_u(\tau)$ et d'un tempo T, notre méthode extrait des paramètres décrivant les pics importants de l'auto-corrélation. L'objectif va être de trouver plusieurs pics importants dans $R(\tau)$: un correspondant à la noire, et deux correspondant aux croches courtes et longues du motif de swing.

La durée d'une croche peut-être dérivée de la position des pics dans la fonction d'auto-corrélation $R(\tau)$. Comme $R(\tau)$ est en pratique souvent assez bruitée (comme on peut le constater sur la Figure 5.2 par exemple), les algorithmes de détection de pics cherchant les maximums locaux ne fonctionnent pas. Nous proposons d'utiliser une méthode de régression dans laquelle le signal cible est approché par une forme prédéfinie de courbe. Nous utilisons pour cela une régression au travers de la méthode des moindres carrés, grâce à l'algorithme de Levenberg-Marquardt [Levenberg, 1944 ; Marquardt, 1963], avec une forme de

courbe gaussienne [4] :

$$f(\tau) = A \exp\left(\frac{-(\tau-\mu)^2}{2\sigma^2}\right)$$

Une régression fonctionne mieux lorsqu'il y a peu de paramètres à adapter et lorsque la quantité de signal à adapter est faible. Nous avons donc pris le parti d'aider l'adaptation de pic en lui fournissant les intervalles de recherches où doivent théoriquement se trouver les pics. Pour cela, nous déduisons la durée théorique d'une croche d_c à partir du tempo T : $d_c = \frac{60}{2T}$. En l'absence de swing, un pic doit être présent à d_c secondes (correspondant à une croche sans swing). En présence de swing, ce pic est scindé en deux pics : un entre 0 et d_c, et un autre entre d_c et $2d_c$. Dans tous les cas, un pic doit être présent à $2d_c$ (durée théorique d'une noire). Comme, à priori, on ne sait pas si on se trouve dans un cas où il y a du swing, ou dans un cas où il n'y en a pas, on va chercher des pics dans les quatre intervalles mentionnés (et représentés sur la Figure 5.7) :

— Le premier correspond à une croche sans swing : $d_c \pm \frac{d_c}{2}$.
— Le second correspond à la croche courte (en cas de présence de swing) : $\left[\frac{d_c}{2}, d_c\right]$ [5].
— Le troisième correspond à la croche longue : $\left[d_c, \frac{3d_c}{2}\right]$.
— Le dernier correspond à la noire : $2d_c \pm \frac{d_c}{6}$.

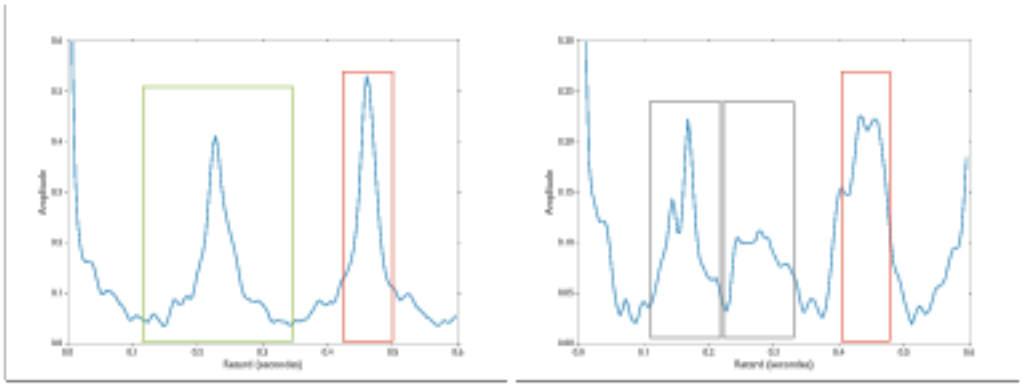

FIGURE 5.7 – *Visualisation des intervalles de recherches des différents pics de la fonction d'auto-corrélation. À droite, le signal possède du swing et à gauche non. Chaque intervalle est représenté par un rectangle. L'intervalle de droite sur chacune des figures correspond à la noire, les deux intervalles de gauche sur la figure de droite sont les intervalles de recherche des croches ayant du swing. Ces intervalles ne sont pas représentés sur la figure de gauche pour des raisons de clarté. L'intervalle de gauche de la figure de gauche correspond à celui d'une croche sans swing (cet intervalle n'est pas non plus représenté sur la figure de droite).*

Pour chacun de ces intervalles, l'algorithme de détection de pic va nous renvoyer trois valeurs correspondant aux paramètres du pic estimé : son amplitude A, son écart-type σ (sa largeur) et son espérance μ (sa durée estimée). Dans le cas où l'algorithme de régression ne trouve aucun pic dans l'intervalle, ces trois valeurs vaudront 0.

Les paramètres du pic correspondant à la croche sans swing sont notés A_c, σ_c et μ_c. Les paramètres de la croche courte et de la croche longues sont notés

4. Nous aurions aussi pu utiliser un polynôme du second degré, mais les paramètres des gaussiennes sont sémantiquement plus intéressants : A est directement l'amplitude du pic, μ est sa position et σ est proportionnel à sa largeur à mi-hauteur.

5. Cet intervalle implique que le ratio de swing est compris entre 1 et 3.

respectivement A_s, σ_s, μ_s et A_l, σ_l, μ_l [6]. Les paramètres de la noire sont notés A_n, σ_n et μ_n.

5.2.4.2 Ensemble de règles

Grâce aux douze paramètres estimés ci-dessus, nous proposons un ensemble de règles qui permettent de décider si du swing est présent ou non dans la trame. Les critères choisis sont essentiellement des critères qui mesurent la fiabilité des sorties de l'algorithme de régression [7].

L'ensemble de règles choisies pour décider si une trame possède du swing ou non peut être résumé à : « Il existe des fonctions gaussiennes approchant le pic de la croche courte et celui de la croche longue. », ce qui se traduit mathématiquement par :

- des amplitudes strictement positives $A_s > 0$, $A_l > 0$.
- des largeurs petites $\sigma_s < \frac{d_c}{4}$, $\sigma_l < \frac{d_c}{4}$.
- des positions dans les bons intervalles $\frac{d_c}{2} < \mu_s < d_c < \mu_l < \frac{3d_c}{2}$ (Cette règle est nécessaire car la régression peut trouver des gaussiennes dont le centre n'est pas dans l'intervalle souhaité.).

Si toutes ces conditions sont satisfaites, **alors** la fonction d'auto-corrélation de la trame est classifiée comme ayant du swing. Pour chaque trame classifiée comme ayant du swing, on peut estimer son ratio de swing $s_r = \frac{\mu_l}{\mu_s}$.

Approche arbre de décision. Cet ensemble de règles précédent a été mis au point empiriquement. Ces règles donnent de bons résultats, comme on pourra le voir plus tard. Cependant, on peut se demander si cet ensemble de règles est optimal et si le seuil choisi pour la largeur des pics ($\sigma_s < \frac{d_c}{4}$) est le bon. Nous avons utilisé une méthode d'apprentissage automatique afin de créer un ensemble de règles de décision (apprentissage par arbres de décision, algorithme CART, critère de réduction d'entropie).

Nous avons trouvé les seuils légèrement plus optimaux :

$$0,0211 < \frac{\sigma_s}{d_c} < 0,2219 \quad \text{au lieu de} \quad \sigma_s < \frac{d_c}{4}$$

$$0,0725 < \frac{\sigma_l}{d_c} < 0,2358 \quad \text{au lieu de} \quad \sigma_l < \frac{d_c}{4}$$

$$0,4247 d_c < \mu_s < 0,8748 d_c \quad \text{au lieu de} \quad 0,5 d_c < \mu_s < d_c$$

$$1,0465 d_c < \mu_l \quad \text{au lieu de} \quad d_c < \mu_l < \frac{3d_c}{2}$$

Cependant, le peu d'amélioration que nous apporte ces nouveaux seuils nous a conduit à garder les règles énoncées précédemment, car nous pensons que ces nouvelles règles sont un peu sur-apprises et spécifiques pour notre ensemble de données, et nous préférons garder des règles un plus générales et plus simples.

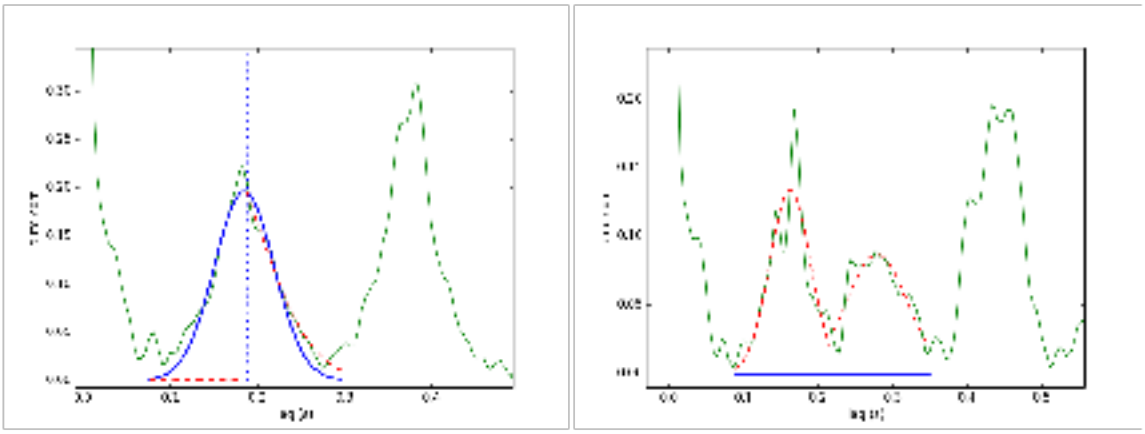

FIGURE 5.8 – *À gauche : fonction d'auto-corrélation d'un signal sans swing ('disco.00020.wav', première trame). À droite : fonction d'auto-corrélation d'un signal avec swing ('jazz.00099.wav', première trame). La ligne verticale pointillée correspond à la durée théorique d'une croche. L'auto-corrélation $r(\tau)$ est la ligne fine verte. La ligne épaisse et bleue correspond à l'approximation du pic correspondant à une croche non-swinguée. Les lignes rouges en pointillés correspondent à l'approximation des pics correspondant aux croches courtes et longues du swing.*

5.2.4.3 Illustrations

La détection et l'estimation de swing expliquées précédemment sont illustrées dans la double figure 5.8. Sur ces deux figures, la ligne continue verte correspond à la fonction d'auto-corrélation. La ligne pointillée verticale correspond à la durée théorique de la croche d_c. La ligne épaisse et bleue correspond à la fonction gaussienne approximée du pic correspondant à une croche non-swinguée. Les lignes rouges en pointillés correspondent aux fonctions gaussiennes approximées des pics correspondants aux croches courtes et longues du swing. Grâce aux fonctions approximées et à l'ensemble de règles décrit ci-dessus, notre algorithme va décider qu'il n'y a pas de swing à gauche, et qu'il y a du swing à droite. De plus, dans le cas où du swing est présent (figure de droite), le ratio de swing estimé est $s_r = \frac{0,28}{0,16} = 1,75$ (pour un swing annoté de 1.73).

5.3 Évaluation

Dans cette partie nous proposons différentes expériences pour mesurer et comparer les performances de nos modèles d'estimation automatique du ratio de swing. Dans les deux premières expériences, nous évaluons la capacité de modèles à détecter si une trame, puis un extrait possèdent du swing ou non, indépendamment de sa valeur. Dans la troisième expérience, nous proposons une évaluation de la précision des ratios de swing estimés. Dans la dernière expérience, nous nous intéressons aux relations entre le ratio de swing, le tempo et l'artiste. Pour toutes nos expériences, nous utilisons l'ensemble de données GTZAN-RHYTHM qui a été créé spécifiquement pour cette tâche et qui a été décrit dans la partie 3.3.

6. s pour 'short' et l pour 'long'
7. D'autres critères ont été testés, comme une mesure de symétrie entre les croche courte et longue par rapport à d_c, sans succès.

F-mesure, précision, rappel... Tous nos résultats sont présentés en terme de f-mesure moyenne. La f-mesure, pour une classe, est définie comme suit :

$$\text{F1} = 2.\frac{\text{Pre.Rec}}{\text{Pre} + \text{Rec}}$$

où Pre est la précision, et Rec le rappel. La précision Pre, pour une classe donnée, est définie comme le nombre de trames correctement attribuées à cette classe sur le nombre total de trames attribuées à cette classe. Le rappel Rec, pour une classe donnée, est défini comme le nombre de trames correctement attribuées à cette classe sur le nombre de trames appartenant à cette classe.

Estimation de tempo. Une estimation de tempo $\hat{T}$ est dite correcte si $\frac{T-\hat{T}}{T} < 4\%$.

5.3.1 Détection du swing

Dans cette première expérience, nous évaluons la capacité des trois modèles à estimer correctement si une trame possède ou non du swing, indépendamment de la valeur exacte du ratio de swing. C'est donc une expérience de classification[8] à deux classes *Swing* et *NoSwing*. Pour rappel, l'ensemble de données GTZAN-RHYTHM que nous avons annoté possède 178 extraits ayant du swing, et 822 n'en ayant pas. Chacun des modèles fournit, pour chaque trame u, une matrice de probabilité $P_u(T, s)$ (avec T le tempo et s le ratio de swing). Une trame est estimée comme appartenant à la classe *Swing* si elle a un ratio de swing $\hat{s}$ estimé strictement supérieur à 1. L'idée est donc de convertir la matrice de probabilité $P_u(T, s)$ en une liste de swings estimés $\hat{s}(u)$ et de tempos estimés $\hat{T}(u)$.

Méthode A. Dans un premier temps, nous utilisons directement le maximum de $P_u(T, s)$ à chaque trame. Le swing estimé $\hat{s}(u)$ pour chaque trame u est le ratio de swing pour lequel P est maximale :

$$\hat{T}, \hat{s} = \text{argmax}_{T,s}\left(P_u(T, s)\right)$$

Méthode B : t_{constant}. Dans un second temps, nous faisons l'hypothèse que le tempo reste constant au cours de l'extrait musical[9]. Le tempo T_{constant} est automatiquement déduit des probabilités $P_u(T, s)$: nous le trouvons en prenant le tempo le plus probable au niveau de l'extrait musical.

Méthode C : $t_{\text{estimé}}$. Dans un troisième temps, nous fixons le tempo $\hat{T}$ à la valeur du tempo estimé par la méthode de [Peeters et al., 2011]. Le tempo est également considéré constant.

Méthode D : $t_{\text{annoté}}$. Dans un dernier temps, nous fixons le tempo $\hat{T}$ au tempo annoté $T_{\text{annoté}}$. Le tempo est également considéré constant.

8. voir Annexe B pour la description des différentes méthodes de classification et de régression.

9. Cette hypothèse est possible sur l'ensemble GTZAN-RHYTHM où les tempos ne changent pas trop au cours des morceaux car les extraits sont courts. Elle ne sera pas forcement valide sur d'autres ensembles de données.

Une fois que nous avons le tempo, nous prenons pour chaque trame le ratio de swing associé ayant la plus forte probabilité :

$$\hat{s} = \text{argmax}_s \left(P_u(T, s | T = T_{\text{fixé}}) \right)$$

Nous notons $\hat{s}(u)$ le ratio de swing estimé en fonction de la trame u pour un extrait musical, (et $s(u)$ la vérité-terrain). Lorsque le ratio $\hat{s}(u)$ vaut 1, la trame est dite appartenir à la classe *NoSwing*. Dans le cas contraire ($\hat{s}(u) > 1$), elle appartient à la classe *Swing*.

Résultats. Nous présentons les résultats de l'expérience dans la Figure 5.9. Sans source d'information extérieure relative au tempo (A), c'est la méthode LLACF qui fonctionne le mieux avec une f-mesure moyenne de 59% (contre 35% et 26% pour ACF et PIC). L'hypothèse seule d'un tempo constant (B : déduction automatique du tempo T_{constant} des probabilités) ne suffit pas à améliorer notablement les résultats des différentes méthodes, à part pour la méthode PIC qui passe de 25,7% à 30,4%(+5%). En prenant un tempo constant et estimé par [Peeters et al., 2011] (C), la méthode PIC (78%) est loin devant les deux autres méthodes (54 et 55%). En prenant un tempo constant et annoté (D), PIC (91%) est encore bien supérieure aux deux autres méthodes (56 et 58%).

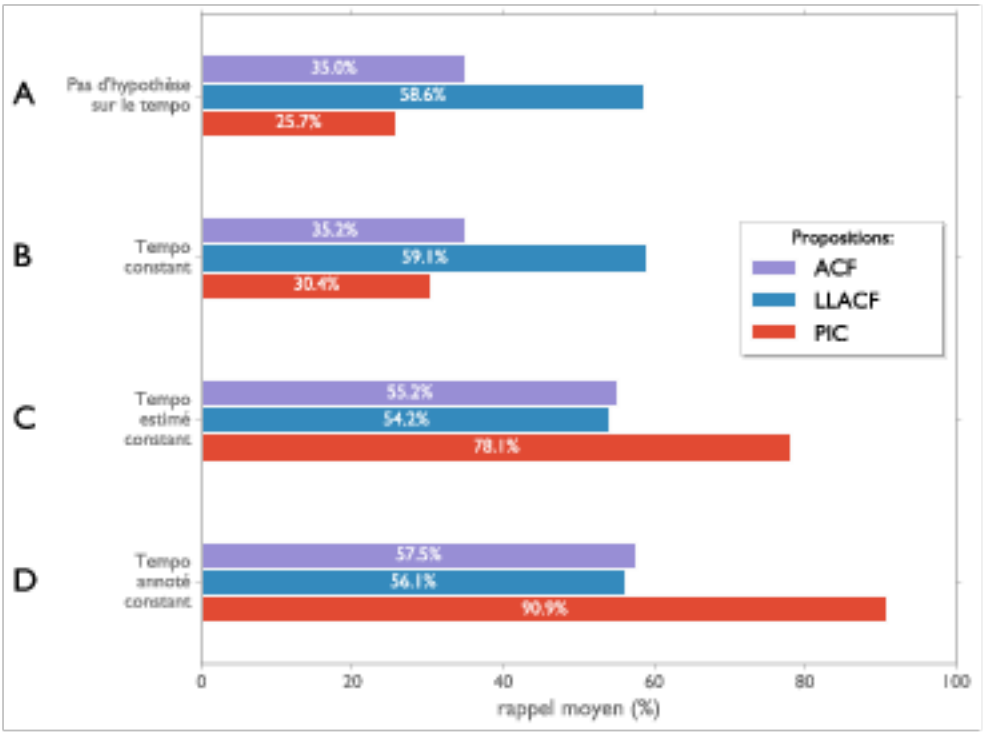

FIGURE 5.9 – *Détection du swing à la trame. Les résultats sont indiqués en f-mesure moyenne.*

Discussion. L'hypothèse d'un tempo constant est donc insuffisante pour améliorer nos méthodes d'estimation du swing. Cette absence d'amélioration peut s'expliquer par la grande similarité qu'il peut y avoir entre les différentes trames d'un même extrait. En général, si une méthode se trompe sur l'estimation du swing sur une trame d'un extrait, elle se trompera de la même façon sur toutes les autres trames. Une modélisation plus poussée de l'évolution du tempo et du

swing au cours du temps (par exemple en utilisant un décodage Vitterbi) n'apporte pas non plus d'améliorations, pour la même raison.

Les deux méthodes ACF et LLACF sont très similaires et obtiennent des résultats semblables, sauf dans les premiers cas (A et B) où la méthode LLACF est bien meilleure (environ 60% contre 35%). Par rapport à la méthode ACF, La méthode LLACF utilise une représentation de l'auto-corrélation en log-temps. Cela permet de transposer une variation de tempo en simple décalage temporel, ce qui permet donc d'avoir une meilleure corrélation entre tempo proches. La méthode LLACF est donc plus à même de se synchroniser sur le bon tempo que son homologue. Lorsque l'on rajoute l'information de tempo, la méthode LLACF perd donc son avantage sur la méthode ACF, et les résultats redeviennent similaires.

Pour conclure, nous notons une très nette supériorité de notre méthode PIC sur celle de [Dittmar et al., 2015] dès qu'une information de tempo est apportée. Avec un tempo estimé, notre méthode PIC surpasse de 24% la méthode LLACF, et avec le tempo annoté elle surpasse de 35% (91% contre 56%).

Résultats détaillés par classe et par genre musical. Comme les chiffres de la Figure 5.9 ne suffisent pas à eux seuls à expliquer pourquoi les résultats des méthodes ACF et LLACF sont si faibles, nous les détaillons dans un second tableau. Dans le Tableau 5.1, chaque colonne représente les résultats sur un sous-ensemble du GTZAN-RHYTHM. Chaque genre du GTZAN-RHYTHM est divisé en trames avec swing et sans swing (classes *Swing* et *NoSwing*). Pour chaque sous-ensemble, nous indiquons : le nombre de trame, le pourcentage de bonne estimation de tempo, et les rappels des deux classes *Swing* et *NoSwing* obtenus pour quatre expériences (LLACF avec tempo estimé, LLACF avec tempo annoté et PIC avec tempo estimé/annoté)[10].

TABLE 5.1 – *Résultats détaillés de la détection du swing. Chaque colonne représente les résultats sur un sous-ensemble du* GTZAN-RHYTHM. *Chaque genre du* GTZAN-RHYTHM *est divisé en trame avec swing et sans swing. Pour chaque sous-ensemble, nous indiquons : le nombre de trame, le pourcentage de bonne estimation de tempo, et les* ***rappels*** *des deux classes Swing et NoSwing obtenus pour quatre expériences (LLACF avec tempo estimé, LLACF avec tempo annoté et PIC avec tempo estimé/annoté).*

	blues		**classical**		**country**		**disco**		**hiphop**	
	Swing	*NoSwing*	*Swing*	*NoSwing*	*Swing*	*NoSwing*	*Swing*	*NoSwing*	*Swing*	*NoSwing*
# trames	700	700	92	1308	350	1050	28	1372	0	1400
Estimation du tempo (%)	66	74	41	65	52	57	100	94	0	97
LLACF / $T_{estimé}$	80.3	60.1	63.0	41.9	82.3	62.8	100.0	70.4	0.0	58.4
LLACF / $T_{annoté}$	99.7	57.7	100.0	37.5	97.7	56.3	100.0	69.6	0.0	55.6
PIC / $T_{estimé}$	63.0	93.7	56.5	99.6	45.4	99.0	100.0	99.3	0.0	99.9
PIC / $T_{annoté}$	93.0	94.9	92.4	99.6	82.6	97.2	100.0	99.3	0.0	99.9

	jazz		**metal**		**pop**		**reggae**		**rock**	
	Swing	*NoSwing*	*Swing*	*NoSwing*	*Swing*	*NoSwing*	*Swing*	*NoSwing*	*Swing*	*NoSwing*
# trames	644	756	84	1316	14	1386	448	952	112	1288
Estimation du tempo (%)	37	52	83	62	100	79	59	62	62	82
LLACF / $T_{estimé}$	71.0	31.5	100.0	53.8	100.0	33.8	77.5	66.3	90.2	62.6
LLACF / $T_{annoté}$	93.0	28.7	100.0	60.2	100.0	23.1	86.2	62.1	100.0	58.1
PIC / $T_{estimé}$	23.6	99.2	63.1	99.8	0.0	100.0	52.0	94.5	47.3	98.2
PIC / $T_{annoté}$	59.6	98.7	77.4	99.6	100.0	97.8	84.2	90.2	82.1	99.6

10. Nous avons volontairement choisi de ne pas donner les résultats de certaines configurations car ceux-ci ne présentent que peu d'intérêt et alourdissent la lecture : les méthodes ACF et LLACF donnent des résultats très similaires, il est donc inutile de présenter les deux, et les estimations à base de tempo variable et de $T_{constant}$ ne fournissent pas de bons résultats.

Avant de discuter ces résultats, nous rappelons qu'un rappel de la classe *Swing* faible signifie une grande proportion de faux-négatifs (trames estimées comme n'ayant pas de swing alors qu'elles en ont). Un rappel de la classe *NoSwing* faible signifie une grande proportion de faux-positifs (trames estimées comme ayant du swing alors qu'elles n'en ont pas).

La méthode LLACF possède, en général, un faible rappel pour la classe *NoSwing* (environ 50% en moyenne), quelque soit le genre. Cela signifie donc que cette méthode produit énormément de faux-positifs. Inversement, la méthode PIC a, en moyenne, ses rappels plus faibles pour la classe *Swing* (48% pour $T_{\text{estimé}}$ et 80% pour $T_{\text{annoté}}$) que pour la classe *NoSwing* (99% et 98% en moyenne). Cela signifie que la majorité des erreurs de cette méthode sont des faux-négatifs. Ces résultats montrent que les deux méthodes font deux types d'erreurs différentes : LLACF produit majoritairement des faux-positifs et PIC des faux-négatifs. Le phénomène peut s'expliquer par le fait qu'un pic représentant la croche (non-swinguée) n'est pas toujours présent dans la fonction d'auto-corrélation. Dans ce cas-là, la méthode PIC n'a que très peu de chances de trouver des formes de pics, alors que la méthode LLACF va juste donner une probabilité plus forte au motif qui correspond le mieux. En l'absence d'un pic pour la croche forte, le meilleur prototype a de grandes chances d'être un motif avec du swing.

Le deuxième point à noter de ce tableau est qu'utiliser un tempo plus fiable (passer du $T_{\text{estimé}}$ au $T_{\text{annoté}}$), permet d'augmenter significativement le rappel de la classe *Swing* (+20/25% en moyenne), quelque soit la méthode. Cela permet donc de diminuer grandement la proportion de faux-négatifs. Pour la méthode PIC, il est indispensable d'avoir une bonne estimation de tempo pour détecter correctement la classe *Swing* : une mauvaise estimation du tempo conduisant souvent à une non-détection du swing. Pour la méthode LLACF, n'ayant déjà que peu de faux-négatifs, utiliser le tempo annoté n'améliore (en f-mesure) que peu les résultats. Cependant, en terme de rappels (Tableau 5.1), la tendance est plutôt positive : on voit cependant une faible baisse des rappels de la classe *NoSwing* et une forte augmentation des rappels de la classe *Swing*.

Pour l'instant, la méthode donnant de meilleurs résultats est donc la PIC, avec le tempo annoté. À part pour les extraits jazz, le rappel est toujours supérieur à 80% pour les deux classes *Swing* et *NoSwing*. On retrouve ici ce qui a été mentionné par les annotateurs du GTZAN-Rhythm à savoir que les extraits de la catégorie jazz sont très complexes rythmiquement.

5.3.2 Généralisation à un extrait complet

L'objectif de cette deuxième expérience est de voir s'il est possible d'améliorer les estimations locales (à la trame) du ratio de swing en tenant compte des informations des trames de tout l'extrait. C'est la même expérience que précédemment, mais avec une évaluation faite au niveau de l'extrait, au lieu d'être faite au niveau de la trame. Concrètement, le problème est le suivant : connaissant une liste de ratios de swing au cours du temps $\hat{s}(u)$, est-ce que l'extrait doit être classifié comme *Swing* ou *NoSwing* ? Par exemple, une liste de ratios $[1; 1; 1; 1; 2.4; 1; 1; 1; 3; 1; 1]$ provient très probablement d'un extrait n'ayant pas de swing du tout. Les deux valeurs différentes de 1 sont sûrement des erreurs d'estimation car il est peu probable que seules deux trames au milieu de l'extrait swinguent, alors que le reste de l'extrait non. Nous proposons deux méthodes de généralisation à un extrait : une méthode dite « médiane », et une méthode d'apprentissage machine notée SVM.

Méthode médiane. L'extrait est classifié comme *Swing* si la médiane de $\hat{s}(u)$ est strictement supérieure à 1 (c'est-à-dire si la moitié au moins des trames swinguent).

Méthode apprentissage SVM. Nous proposons une expérience simple d'apprentissage machine. Premièrement, nous extrayons un ensemble de descripteurs à partir de $\hat{s}(u)$, puis nous les utilisons pour entraîner un algorithme de classification supervisé afin d'estimer les classes *Swing* et *NoSwing*.

Nous proposons l'ensemble de descripteurs suivant, extrait à partir des informations de swing trame à trame $\hat{s}(u)$ et de leur probabilités associées notées $P_u(T, s|T = \hat{T}(u), s = \hat{s}(u))$:

— le pourcentage de trames ayant du swing dans l'extrait.
— moyenne, écart-type, médiane, écart interquartile des $\hat{s}(u)$ parmi les trames ayant du swing ($\hat{s}(u) > 1$).
— moyenne, écart-type, médiane, écart interquartile des probabilités associées $P_u(T, s|T = T_{\text{constant,estimé,annoté}}, s = s(u)$.

Nous utilisons ces descripteurs en entrée d'un algorithme de classification automatique de type SVM avec un noyau RBF. Nous trouvons les meilleurs paramètres du SVM c (paramètre de coût) et γ (coefficient du noyau) par recherche exhaustive. Nous cherchons la meilleure moyenne des f-mesures par classe (*Swing* et *NoSwing*) par une validation croisée à 10 plis.

Résultats. Nous présentons les résultats des trois méthodes (ACF, LLACF, PIC) pour trois tempos différents (constant, estimé, annoté) en terme de f-mesure moyenne. Les résultats en rouge (les trois résultats du haut de chaque groupe) sont les rappels de l'expérience précédente (Figure 5.9). Les groupes bleus et violets (respectivement les trois lignes centrales et les trois lignes du bas de chaque groupe) montrent les f-mesures moyennes évaluées au niveau d'un extrait complet, avec les deux méthodes de généralisation (médiane et SVM) proposées dans la partie 5.3.2

Tout d'abord, la méthode médiane n'améliore les résultats que dans seulement deux cas sur neuf. Cependant, il faut noter que ces deux cas (PIC, $T_{\text{estimé}}$ et $T_{\text{annoté}}$) sont les meilleurs résultats jusqu'à présent et les améliorer, même légèrement (78,1% à 79.1% et 90.9% à 92%), n'est pas négligeable. La méthode SVM permet elle, une nette amélioration de tous les résultats. Cette amélioration est limitée (environ +2%) pour les deux cas que nous venons de citer (PIC, $T_{\text{estimé}}$ et $T_{\text{annoté}}$), mais elle est énorme dans les autres cas (jusqu'à +33% pour LLACF, $T_{\text{annoté}}$).

Les meilleurs résultats sont évidemment obtenus pour les expériences utilisant le tempo annoté. La méthode PIC possède toujours les meilleurs performances (93% contre 89.3% pour LLACF et 78% pour ACF). Les meilleurs résultats obtenus sans information annotée sont à 80.4% pour la PIC, suivi de près par LLACF (78.6%).

5.3.3 Ratio de swing

Dans cette expérience, nous mesurons la précision des valeurs du ratio de swing fournis par nos trois modèles (ACF, LLACF et PIC) et nos deux méthodes d'apprentissage (médiane et SVM). Attention, dans cette partie, ce n'est plus une classification par SVM mais une régression SVM que nous notons SVM-R. Nous utilisons pour la corrélation de Pearson pour comparer nos résultats à ceux de

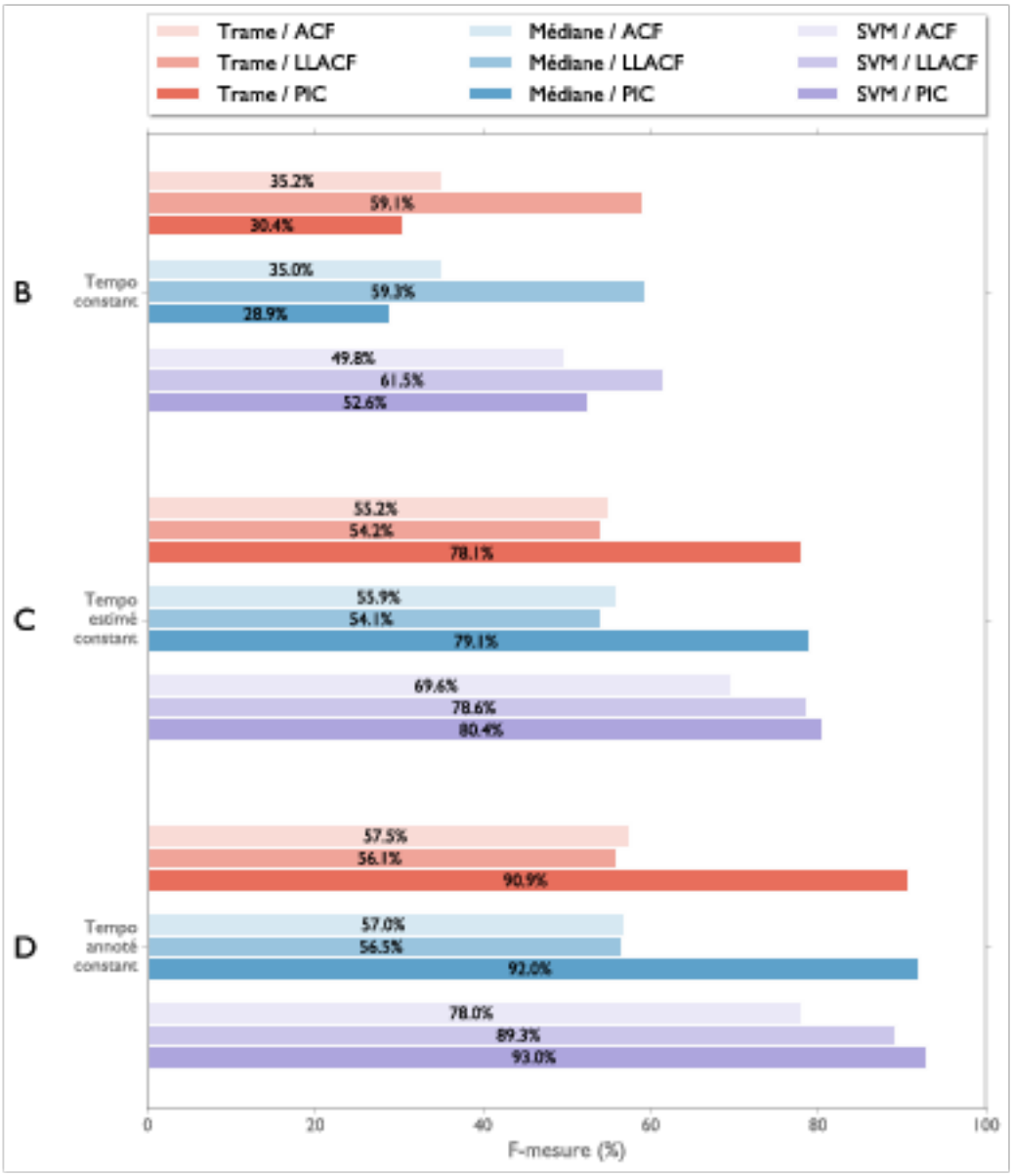

FIGURE 5.10 – *Généralisation de la détection du swing à un extrait. Nous présentons les résultats des trois méthodes (ACF, LLACF, PIC) pour trois tempos différents (constant, estimé, annoté) en terme de f-mesure moyenne. Les résultats en rouge (les trois résultats du haut de chaque groupe) sont les rappels de l'expérience précédente (Figure 5.9). Les groupes bleus et violets (respectivement les trois lignes centrales et les trois lignes du bas de chaque groupe) montrent les f-mesures moyennes évaluées au niveau d'un extrait complet, avec les deux méthodes de généralisation (médiane et svm) proposées dans la partie 5.3.2*

[Dittmar et al., 2015]. Si $\hat{s}$ est la liste des n ratios de swing estimés et s la vérité-terrain, la corrélation de Pearson vaut :

$$r = \frac{\sum_{i-1}^{n}(\hat{s}_i - \bar{\hat{s}})(s_i - \bar{s})}{\sqrt{\sum_{i-1}^{n}(\hat{s}_i - \bar{\hat{s}})^2}\sqrt{\sum_{i-1}^{n}(s_i - \bar{s})^2}}$$

Résultats. Nous présentons les résultats de cette expérience dans la Table 5.2. Nous cherchons à évaluer la précision de l'estimation du ratio de swing. La première ligne donne le nombre d'extraits sur lesquels ont été calculés les résultats. Les précisions sont indiquées en terme de corrélation de Pearson. Nous comparons deux méthodes médiane et SVM-R entre elles. À titre indicatif, les résultats de l'état de l'art sont indiqués, même si ceux-ci ont été obtenus sur un corpus différent.

TABLE 5.2 – *Précision des ratios de swing estimés. La première ligne donne le nombre d'extraits sur lesquels ont été calculés les résultats. Les précisions sont indiquées en terme de corrélation de Pearson. Nous comparons deux méthodes médiane et SVM-R entre elles. À titre indicatif, les résultats de l'état de l'art sont indiqués, même si ceux-ci ont été obtenus sur un corpus différent.*

Corpus :		GTZAN-RHYTHM (#=177)		Weimar Jazz Database (#=42)
Méthode :		médiane	SVM-R	[Dittmar et al., 2015]
$T_{\text{estimé}}$	ACF	0.40	0.34	
	LLACF	0.56	0.44	
	PIC	0.40	0.61	
$T_{\text{annoté}}$	ACF	0.24	0.47	
	LLACF	0.40	0.66	0.9
	PIC	0.59	0.77	

Nous remarquons dans le Tableau 5.2 qu'estimer le ratio de swing grâce à la médiane donne en général des valeurs moins précises. La méthode donnant de meilleurs résultats est encore la méthode PIC avec $T_{\text{annoté}}$ (coefficients de 0.59 avec la médiane et 0.77 avec SVM-R). Il est difficile de comparer ces résultats avec l'état de l'art [Dittmar et al., 2015] qui obtient 0.9 de corrélation, car nous n'utilisons pas les même corpus d'évaluation. De plus, [Dittmar et al., 2015] possède un corpus de seulement 42 exemples contre 177 pour nos résultats.

Discussion. Nous avons tracé dans la Figure 5.11 le ratio de swing estimé en fonction du ratio swing annoté, pour la méthode PIC, $T_{\text{annoté}}$, en haut avec la médiane et en bas avec SVM-R. Nous observons que la méthode la plus simple (prendre la médiane des valeurs de swing des différentes trames d'un morceau) est en général plus précise pour l'estimation du ratio de swing : les points sont regroupés autour de la diagonale. Cependant cette méthode se trompe plus souvent en estimant qu'il n'y a pas de swing, et fait des erreurs plus grandes. La méthode SVM-R possède une meilleure corrélation de Pearson, mais les valeurs estimées sont beaucoup plus regroupées entre 1.5 et 2. Ceci est certainement dû à un manque d'exemples d'apprentissage ayant des ratios de swing très faibles et très élevés. Nous préférons donc la méthode médiane car elle produit des résultats plus précis, des erreurs plus franches et elle est beaucoup plus simple à mettre en œuvre et comprendre.

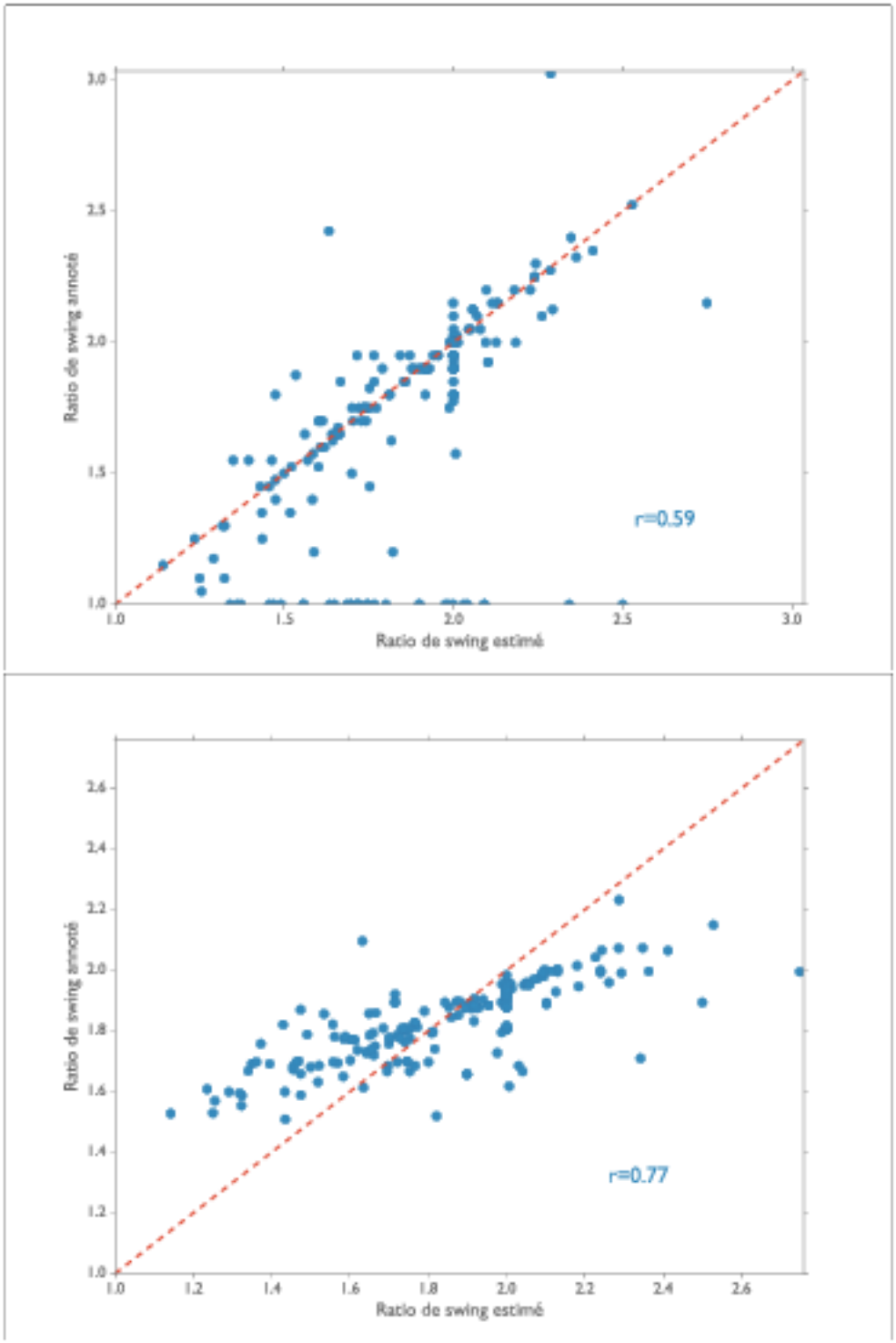

FIGURE 5.11 – *Swing estimé en fonction du swing annoté, pour la méthode PIC, $T_{annoté}$. En haut avec la médiane et en bas avec un svm.*

Nous pouvons aussi conclure que la corrélation de Pearson n'est pas la mesure la plus adaptée pour comparer des ratios de swing. Nous l'avons utilisée afin de pouvoir comparer nos résultats à [Dittmar et al., 2015] et il serait intéressant d'utiliser une mesure plus adaptée à notre problème.

5.3.4 Ratio de swing en fonction du tempo et de l'artiste

Nous avons tracé dans la Figure 5.12 le ratio de swing annoté en fonction du tempo annoté pour un sous-ensemble du test-set (seul les extraits blues et jazz ayant du swing ont été sélectionnés). Nous avons aussi indiqué les artistes correspondants.

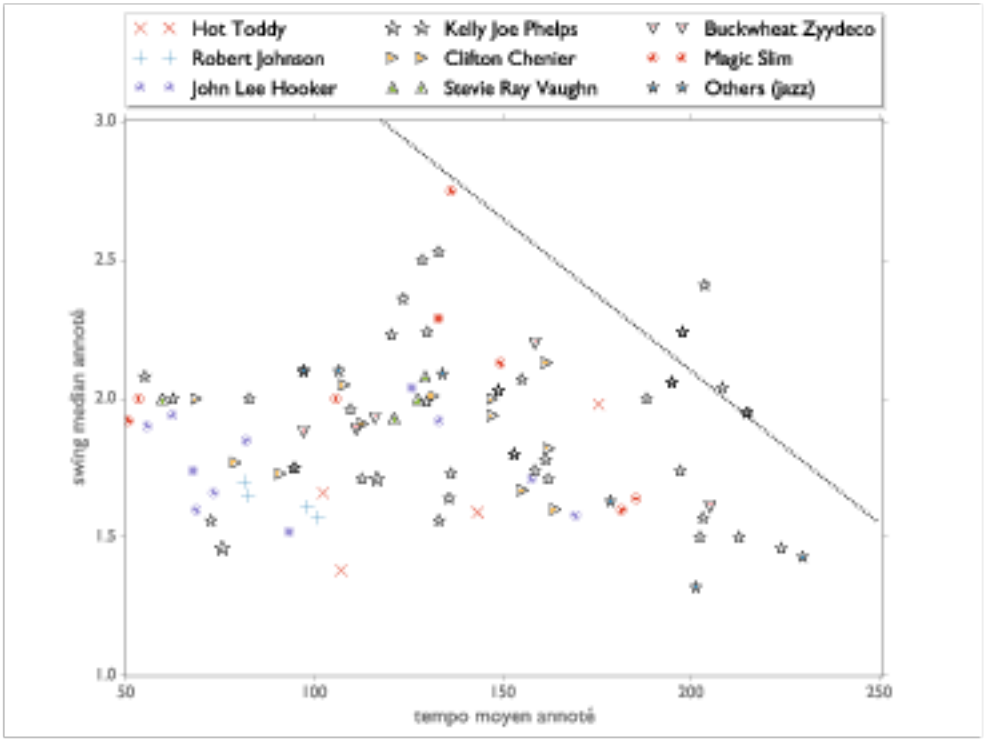

FIGURE 5.12 – *Ratio de swing annoté en fonction du tempo. Chaque différent marqueur correspond à un artiste différent. La ligne pointillée correspond à la relation linéaire entre swing et tempo provenant de l'expérience de Friberg [Friberg et al., 2002].*

Premièrement nous testons l'hypothèse d'un facteur de swing variant linéairement avec le tempo. Cette hypothèse a été suggérée par l'expérience de Friberg et al. [Friberg et al., 2002], confirmée par [Dittmar et al., 2015], mais infirmée par d'autres études [Desain et al., 1994; Honing et al., 2008]. La régression linéaire utilisée dans la Figure 1 de [Friberg et al., 2002] est tracée (en pointillés) dans la Figure 5.12. Nous pouvons voir que pour nos données, le ratio de swing n'a pas de relation linéaire avec le tempo : tous les points de l'ensemble de donnée sont loin de l'hypothèse linéaire de Friberg, et ils ne montrent pas la même tendance non plus. La Weimar Jazz Database (voir Figure 5 de [Dittmar et al., 2015]) est, quant à elle, plus en accord avec l'hypothèse de Friberg. Il faut cependant noter que nos données sont assez différentes, surtout du point de vue du tempo (50-200 bpm pour le GTZAN-RHYTHM, 150-320 bpm pour la Weimar Jazz Database, 120-320 bpm pour les 6 enregistrements de [Friberg et al., 2002]). Notre corpus possède beaucoup plus de tempos faibles et peu de tempos élevés. Il est

donc difficile de comparer nos résultats aux autres expériences, et nos résultats ne viennent donc ni infirmer, ni confirmer l'hypothèse d'un tempo variant linéairement pour les tempos élevés (> 150 bpm). En revanche, nos résultats semblent plus en accord avec les résultats de [Honing et al., 2008] disant qu'il n'y a aucune relation linéaire entre ratio de swing et tempo, lorsque le tempo est faible (< 100-150 bpm).

On voit de plus qu'il y a une légère préférence pour le ratio de swing 2 : 1 (« triple feel ») dans le GTZAN-Rhythm : plus de la moitié des exemples ont un ratio de swing compris entre 1.9 et 2.1. Cette tendance confirme les expériences perceptives de [Essens et al., 1985] et les observations neurologiques [Abecasis et al., 2005], qui ont montrés la prépondérance du ratio 2 par rapport aux autres ratios temporels dans la perception du temps dans la musique.

Enfin, nous testons s'il existe une relation entre l'artiste et le ratio de swing. Nous voyons dans la Figure 5.12, que certains artistes ont tendance à jouer avec un swing faible (Hot Toddy, Robert Johnson) ou avec un swing élevé (Magic Slim). Tous les extraits de Stevie Ray Vaughan ont le même ratio de swing, mais aussi le même tempo. Il est donc difficile, à partir de ces données, de dire si le swing est une caractéristique de l'artiste ou du tempo. Il semble aussi difficile de prédire quel artiste est en train de jouer uniquement à partir du ratio de swing. Une corrélation beaucoup plus marquée était présente dans les travaux de [Dittmar et al., 2015] (voir Figure 5).

5.4 Conclusions

Dans ce chapitre, nous avons étudié les déviations systématiques de position des évènements dans la musique jazz : le swing. Nous avons proposé plusieurs méthodes d'estimation de ces déviations systématiques, toutes basées sur l'auto-corrélation de la fonction d'onset du signal. Cette fonction d'auto-corrélation est très utile pour l'estimation du swing (ou plus généralement de la métrique) car elle possède un pic pour chaque niveau métrique important. La première méthode, notée ACF, compare l'auto-corrélation d'un morceau à une base de prototypes des couples swing/tempo. La seconde, notée LLACF, fait de même mais en utilisant l'auto-corrélation en échelle logarithmique de décalages. La dernière, notée PIC, modélise les pics de la fonction d'auto-corrélation correspondants aux différents niveaux métriques.

Nous avons comparé ces trois méthodes pour une tâche de classification de morceaux en *Swing/NoSwing* (selon que le morceau possède ou non du swing) et pour une tâche d'estimation du ratio de swing. Nous avons montré que notre méthode PIC donne de meilleurs résultats que l'état de l'art (LLACF [Dittmar et al., 2015]) en détection de swing et en estimation de son ratio.

En dépit des résultats moins bons d'une représentation temporelle en échelle logarithmique (comme utilisée dans la méthode LLACF) pour la détection du swing, cette représentation logarithmique permet une estimation du tempo beaucoup plus robuste que l'échelle linéaire (comme utilisée dans la méthode ACF).

Finalement, nous avons montré qu'il n'existe pas, pour notre corpus, de relation linéaire claire entre d'une part tempo et swing et d'autre part tempo et artiste. Ceci est contraire aux expériences de [Dittmar et al., 2015 ; Friberg et al., 2002].

Publications associées

Marchand, U. et Peeters, G. (2015b). « Swing ratio estimation ». *Proceedings of the 18th International Conference on Digital Audio Effects (Dafx)*. Trondheim, Norway.

Chapitre 6

Estimation des motifs rythmiques

Contenu

6.1 Description des motifs rythmiques

Dans cette partie, nous proposons plusieurs descripteurs permettant de modéliser les motifs rythmiques d'un signal musical. Les motifs rythmiques représentent la manière dont les évènements sonores sont groupés entre eux (voir partie 2). D'un point de vue pratique, nous pouvons représenter les motifs rythmiques comme sur la Figure 6.1. Un motif rythmique est un ensemble de rythmes simples (noire, croche, ...), indépendants de la valeur du tempo et pouvant prendre plusieurs dimensions (différents instruments ou différentes bandes de fréquences).

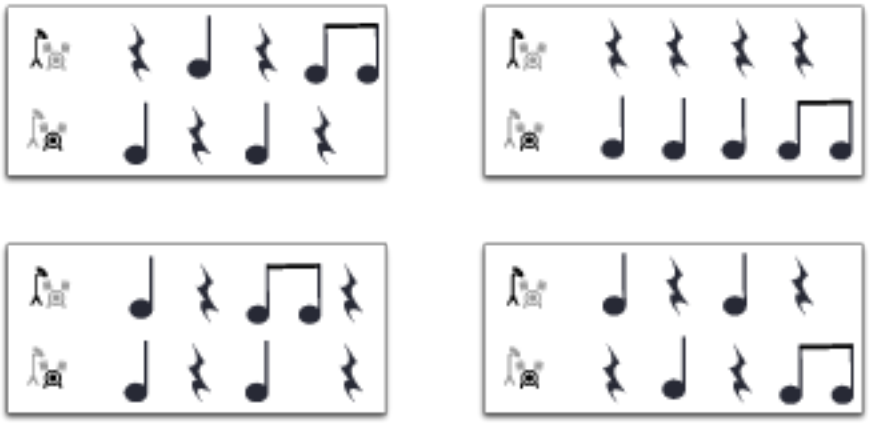

FIGURE 6.1 – *Quelques exemples de motifs rythmiques.*

Nous cherchons à créer une représentation qui modélise ce type de motifs rythmiques. Nous devons donc respecter deux invariances. La première est l'invariance aux décalages temporels. En effet, deux motifs rythmiques joués à des instants différents restent identiques (un exemple sera montré plus tard dans la Figure 6.5). L'instant de début d'un motif rythmique n'a aucune importance. La seconde est l'invariance au tempo. En effet, deux motifs rythmiques joués à des tempo différents sont identiques. Nous allons proposer dans cette partie des descripteurs qui prennent en compte ces deux invariances.

La tâche d'évaluation que nous allons utiliser est une tâche de classification d'extraits musicaux en classes de motifs rythmiques. Nous utilisons pour cela trois corpus dont les classes annotées sont des danses de salons (BALLROOM et EXTENDED BALLROOM) ou des danses traditionnelles de l'île de Crête (CRETE). Chacune des danses possède un motif rythmique prédominant : c'est celui-ci qui permet aux danseurs/danseuses de reconnaître la danse et de faire les bons pas. Nous assimilons donc ces classes de danses à des classes de motifs rythmiques.

6.2 Modulation Scale Spectrum

Dans cette partie, nous proposons une représentation du contenu musical adaptée à l'étude des motifs rythmiques que nous appelons Modulation Scale Spectrum (MSS). Cette représentation est basée sur deux outils : la transformée d'échelle et le spectre de modulation. Nous allons d'abord décrire ces outils, avant de proposer notre descripteur dans la partie 6.2.4 et de l'évaluer dans la partie 6.3.

6.2.1 La transformée d'échelle

La transformée d'échelle (ou Scale Transform) est un cas particulier de la transformée de Mellin qui a été introduite par [Cohen, 1993]. L'échelle y est définie comme un attribut physique du signal, comme la fréquence ou le temps. Avec la transformée d'échelle, il est possible de voir le contenu d'échelle d'un signal comme il est possible de voir le contenu fréquentiel d'un signal grâce à la transformée de Fourier. La transformée d'échelle T_E est définie par :

$$T_E(c) = \frac{1}{\sqrt{2\pi}} \int_0^\infty x(t) t^{-jc-\frac{1}{2}} dt \tag{6.1}$$

où c est la variable d'échelle, t le temps et x un signal.

On peut remarquer qu'il est aussi possible d'interpréter la transformée d'échelle comme la transformée de Fourier du signal ré-échantillonné exponentiellement et atténué par $e^{\frac{1}{2}t}$. En effet, un changement de variable ($t \to e^t$) rend l'équation 6.1 équivalente à :

$$T_E(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x(e^t) e^{\frac{1}{2}t} e^{-jct} dt \tag{6.2}$$

Ce ré-échantillonnage exponentiel possède des similarités avec les représentations du signal où le temps est exprimé en échelle logarithmique [Dittmar et al., 2015 ; Jensen et al., 2009] qui permettent de s'affranchir dans une certaine mesure du tempo.

La transformée d'échelle possède une propriété particulièrement intéressante pour la description du rythme : l'invariance aux changements d'échelle, ce qui signifie dans le cadre d'un signal audio, une quasi-invariance au tempo. La transformée d'échelle a déjà été utilisée pour la représentation du rythme par [Holzapfel et al., 2009 ; Holzapfel et al., 2011].

Invariance aux changements d'échelle

Nous savons déjà que la transformée d'échelle peut se voir comme la transformée de Fourier d'un signal ré-échantillonné et atténué (voir Eq. 6.2). En définissant g telle que :

$$g(t) \stackrel{\text{déf}}{=} x(e^t) e^{\frac{t}{2}} \tag{6.3}$$

La transformée d'échelle de x se résume à la transformée de Fourier de g (notée $T_F[g]$), c'est-à-dire :

$$T_E[x] = T_F[g] \tag{6.4}$$

Ensuite, nous appliquons un changement d'échelle [1] α à x que nous normalisons : $x_\alpha(t)$ est ainsi définie. Nous proposons de plus de définir $g_\alpha(t)$.

$$\begin{aligned} x_\alpha(t) &\stackrel{\text{déf}}{=} \sqrt{\alpha} x(\alpha t) \\ g_\alpha(t) &\stackrel{\text{déf}}{=} \sqrt{\alpha} x(\alpha e^t) e^{\frac{t}{2}} = x(e^{t+\ln\alpha}) e^{\frac{t+\ln\alpha}{2}} = g(t+\ln\alpha) \end{aligned} \tag{6.5}$$

1. Remarque : si $\alpha < 1$, on a une expansion, et si $\alpha > 1$, on a une compression de l'échelle.

En calculant la transformée d'échelle du signal x_α étiré/compressé :

$$\begin{aligned} T_E[x_\alpha](c) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \sqrt{(\alpha)} x(\alpha t) t^{-jc-\frac{1}{2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \sqrt{\alpha} x(\alpha e^t) e^{\frac{t}{2}} e^{-jct} dt \\ &= T_F[g_\alpha(t)] \\ &= T_F[g(t + \ln \alpha)] \\ &= \alpha^{jc} T_F[g(t)] \end{aligned} \tag{6.6}$$

Le passage de l'avant dernière ligne à la dernière ligne se fait grâce à la propriété de la transformée de Fourier suivante :

$$\begin{aligned} x(t) &\rightleftharpoons T_F[x](\omega) \\ x(t + \Delta) &\rightleftharpoons e^{j\Delta\omega} T_F[x](\omega) \end{aligned} \tag{6.7}$$

Donc, la modification d'échelle de x_α se transforme en un décalage temporel dans g_α par l'étape de re-échantillonnage exponentiel. Celui-ci se traduit ensuite par un décalage de phase par l'étape de transformée de Fourier. Nous obtenons finalement :

$$T_E[x](c) = \alpha^{jc} T_E[x_\alpha](c) \tag{6.8}$$

Mathématiquement, si un signal $x(t)$ a pour transformée d'échelle $T_E(c)$, le même signal étiré ou compressé dans le temps aura un spectre d'échelle identique en module :

$$\begin{aligned} x(t) &\rightleftharpoons T_E[x](c) \\ \sqrt{\alpha}\, x(\alpha t) &\rightleftharpoons e^{jc \ln \alpha} T_E[x](c) \end{aligned} \tag{6.9}$$

Cela signifie que deux signaux identiques, mais dont l'un est la version étirée de l'autre, auront le même module de transformée d'échelle.

Quasi-invariance au tempo

Deux signaux identiques dont l'un est la version étirée de l'autre ont donc le même module de transformée d'échelle. Cette propriété est illustrée sur la Figure 6.2 où nous montrons deux signaux dont l'un est la version étirée de l'autre (en haut) et leurs spectres d'échelle qui se superposent parfaitement (en bas).

Dans le cadre de l'analyse du rythme, deux rythmes joués à des tempos proches n'auront pas des signaux étirés l'un de l'autre. Ce ne sont pas les signaux, mais les points d'attaques des évènements qui subissent une transformation d'échelle dans ce cas. Les différents signaux représentant les évènements rythmiques ne sont finalement que translatés selon une échelle logarithmique. C'est pourquoi il est important de réduire les évènements rythmiques à leur instant de début (fonction de détection d'onset).

Nous montrons sur la Figure 6.3 deux signaux audio de deux rythmes identiques joués à deux vitesses différentes (100 et 120 bpm) et leur spectre d'échelle après la chaîne de traitement du MSS (signal, onset, auto-corrélation, transformée d'échelle). Nous voyons que les deux spectres, grâce à l'action de l'onset, et de l'auto-corrélation, sont quasiment identiques.

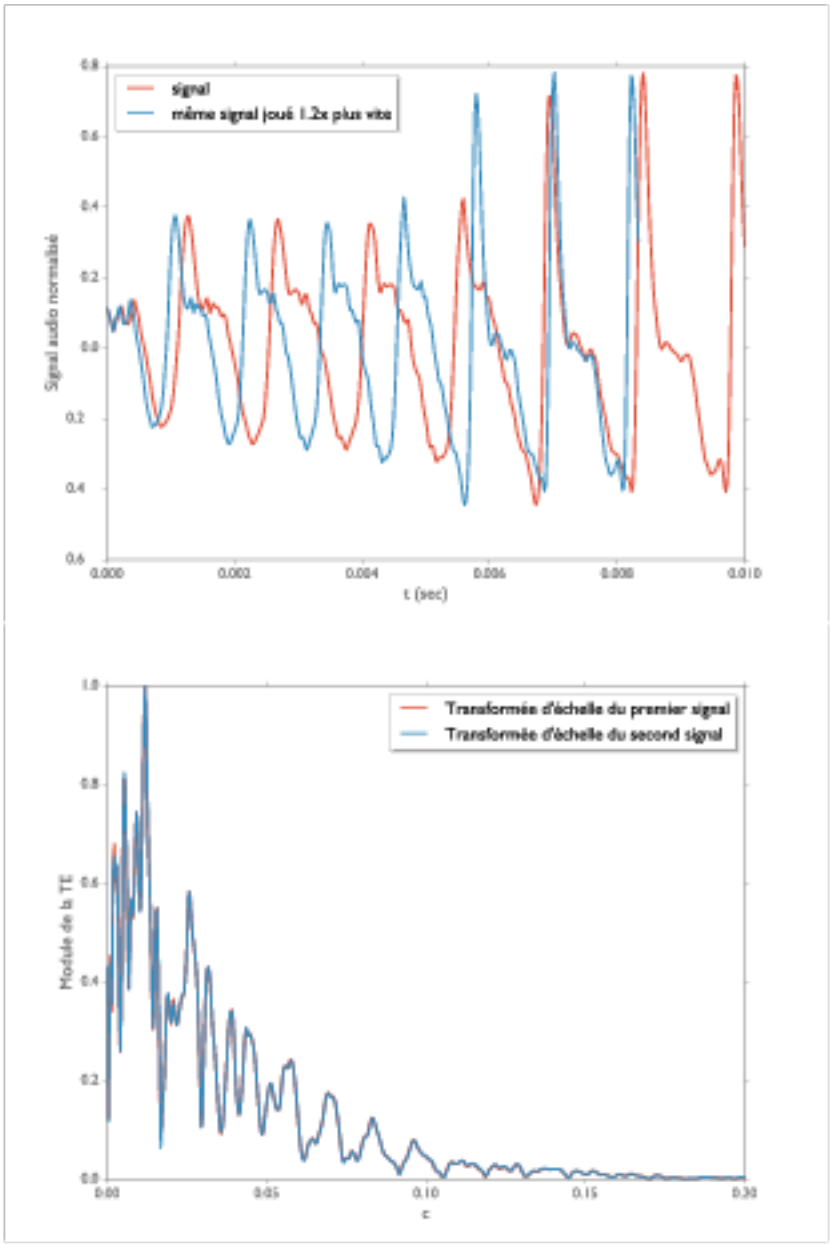

FIGURE 6.2 – *Deux signaux dont l'un est la version étirée de l'autre (en haut) et modules de leur transformées d'échelle (en bas).*

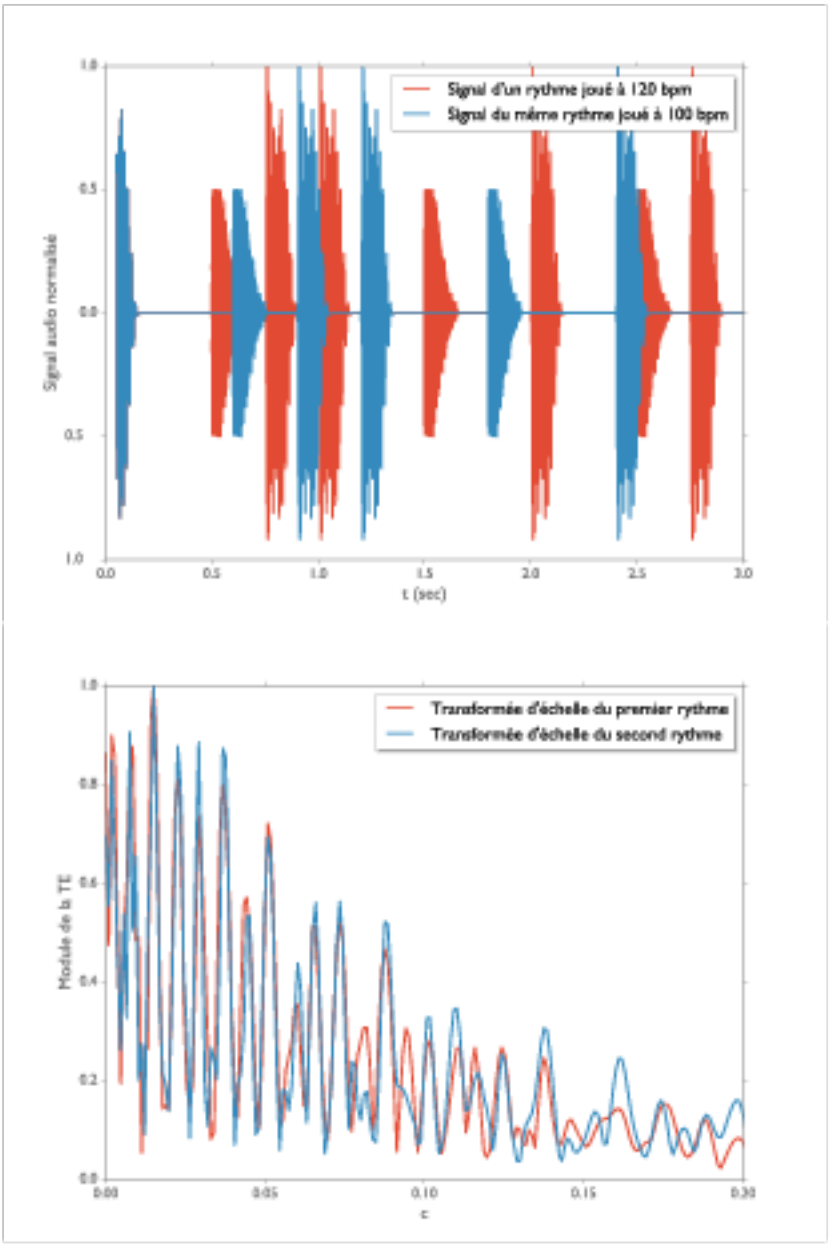

FIGURE 6.3 – *Deux signaux audio de rythmes joués à des vitesse différentes et modules de leur transformées d'échelle.*

6.2.2 Le spectre de modulation

Le spectre de modulation ou « Modulation Spectrum » a déjà été utilisé sous diverses appellations par [Worms, 1998] puis [Rodet et al., 2003] (identification audio par fingerprint), [McKinney et al., 2003] (classification en genre), [Whitman et al., 2004] (étiquetage automatique), et [Atlas et al., 2003] (encodage audio). C'est une représentation compacte décrivant les caractéristiques spectro-temporelles d'un extrait musical.

Le spectre de modulation $X(\omega, \Omega)$ d'un signal $x(t)$ est défini par :

$$\begin{aligned} x(\omega, \tau) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty x(t) h(\tau - t) e^{-j\omega t} dt \\ X(\omega, \Omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty |x(\omega, \tau)| e^{-j\Omega\tau} d\tau \end{aligned} \tag{6.10}$$

La première étape est une transformée de Fourier à Court-Terme (TFCT). Dans la seconde étape, une deuxième transformée de Fourier (TF) est calculée, non pas sur les fréquences de la première (ce qui reviendrait à faire une transformée de Fourier inverse), mais sur l'axe des trames temporelles. Cette seconde TF permet ainsi de décrire la vitesse de variation de l'amplitude des différentes bandes de fréquence (c'est-à-dire la modulation).

6.2.3 Modulation Scale Spectrum

Nous proposons le MSS comme tirant profit des propriétés des deux représentations précédentes. La transformée d'échelle permet une bonne description du contenu rythmique grâce à sa propriété d'invariance à l'échelle, mais ne permet pas une description du timbre (contenu fréquentiel). Le spectre de modulation permet quant à lui une description compacte du timbre, mais n'est procure pas d'invariance au tempo.

Notre contribution a été d'utiliser simultanément les deux approches afin de combiner les propriétés du spectre de modulation (très bonne description compacte du timbre) et de la transformée d'échelle (invariance au tempo). Le MSS est défini par la première étape de l'équation 6.10, suivi d'une seconde étape modifiée :

$$\begin{aligned} x(\omega, \tau) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty x(t) h(\tau - t) e^{-j\omega t} dt \\ T_E(\omega, c) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty |x(\omega, e^\tau)| e^{\frac{1}{2}t} e^{-jc\tau} d\tau \end{aligned} \tag{6.11}$$

Nous avons remplacé la seconde étape, qui était une transformée de Fourier, par une transformée d'échelle. Cela permet de décrire la vitesse de variation de l'amplitude des différentes bandes de fréquences de manière indépendante du tempo.

Le MSS permet donc de combiner description compacte du timbre et invariance au tempo.

6.2.4 Application à la description des motifs rythmiques

Nous proposons ici une implémentation spécifique du MSS qui permet de décrire le contenu rythmique d'un signal. Les étapes successives sont résumées dans la Figure 6.4. Nous les décrivons et les justifions dans cette partie.

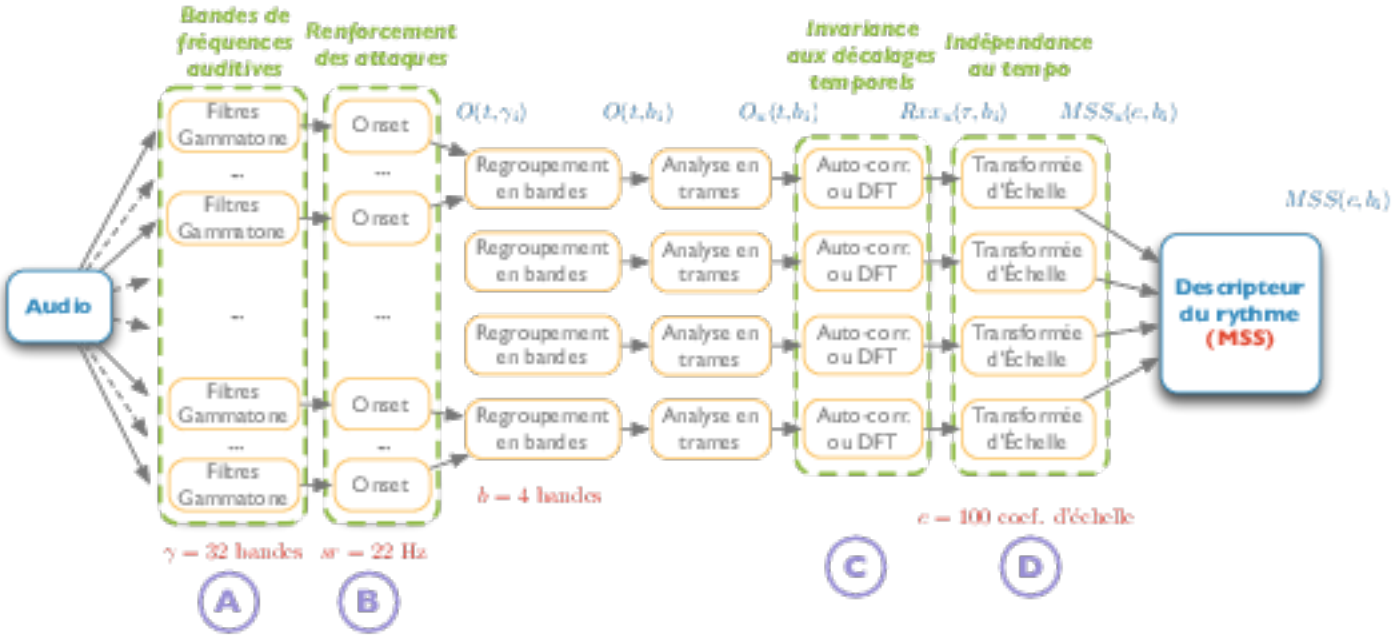

FIGURE 6.4 – *Le Modulation Scale Spectrum pour l'extraction des motifs rythmiques.*

(A) Bandes de fréquences auditives. Tout d'abord, le signal est séparé en différentes bandes de fréquences en utilisant des filtres gammatone. [Patterson et al., 1992] ont montré que la réponse impulsionnelle de la fonction gammatone d'ordre 4 est une très bonne approximation des formes des filtres auditifs humains. Remplacer la TFCT du spectre de modulation par des filtres gammatones n'est pas nouveau, œla a déjà proposé par [McKinney et al., 2003]. Dans notre cas, nous utilisons pas 18 filtres comme proposé, mais un nombre variable : γ filtres. D'un point de vue pratique, nous avons utilisé une implémentation efficace de ces filtres proposé par [Ma et al., 2007], dont le code est disponible en ligne [2]. Nous utilisons des filtres gammatone passe-bandes du 4ème ordre, centré sur des fréquences espacée logarithmiquement entre 26 à 9795 Hz Nous obtenons donc un signal séparé en plusieurs bandes de fréquences $x(\gamma, t)$ à partir du signal $x(t)$.

(B) Renforcement des attaques. Ensuite la fonction d'onset $o(\gamma, t)$ est calculée sur la sortie de chaque filtre $x(\gamma, t)$. L'utilisation d'une fonction d'onset permet d'améliorer la détection des attaques dans le signal, ce qui est souhaitable pour la description du rythme. En effet, même si les évènements rythmiques sont loin d'être cantonnés aux attaques fortes du signal, ils restent en pratique majoritairement issus de celles-ci (c'est-à-dire issus des éléments percussifs de la musique). Dans ce manuscrit, nous ne comparerons pas les différentes fonctions d'onset qu'il aurait été possible d'utiliser ici. Nous avons choisi et utilisé la fonction proposée par [Ellis, 2007], que nous avons paramétrée pour obtenir une fréquence d'échantillonnage sr plutôt basse ($< 50Hz$), afin de s'adapter aux fréquences des rythmes que l'on recherche [3].

2. `http://staffwww.dcs.shef.ac.uk/people/N.Ma/resources/gammatone/`

3. Comme vitesses extrêmes, nous prenons une double croche jouée à 300bpm, et une blanche à 40bpm. En fréquence, ces rythmes correspondent à un intervalle $0.33Hz < sr < 20Hz$.

Puis nous réduisons la quantité d'information disponible en regroupant certaines bandes entres elles. En effet, d"un point de vue général, il est préférable d'avoir des descripteurs ayant la dimension la plus faible possible. De plus regrouper les bandes permet d'être plus robuste aux petits changements fréquentiels. Les γ bandes de fréquences sont regroupées en b bandes. La méthode de regroupement est linéaire : les $\frac{\gamma}{b}$ premières bandes sont moyennées pour former la nouvelle première bande γ_1, puis les $\frac{\gamma}{b}$ bandes suivantes forment la γ_2 et ainsi de suite. Il n'est en effet pas nécessaire de grouper les bandes logarithmiquement, étant donné que les filtres gammatone le sont déjà. Cette façon de grouper permet de conserver les bandes perceptives formées par les filtres gammatone.

Ensuite, une analyse en trames temporelles u est réalisée. Les trames ont une durée de 8 secondes et un pas de 1 seconde. Nous avons choisi des trames d'une durée assez longue, afin qu'elle contiennent toujours au moins quelques mesures musicales (dans 8 secondes tiennent 3 mesures de 4 temps à 40 bpm). Vu les fréquences des phénomènes que nous étudions, prendre un pas d'avancement inférieur à 1 seconde ne présente pas d'intérêt. Nous obtenons donc $o(u, b, t)$

(C) **Invariance aux décalages temporels.** Dans le cadre de l'analyse du rythme, l'invariance aux décalages temporels est indispensable : deux rythmes similaires commençant à des instants différents sont semblables, comme le sont les rythmes présentés sur la Figure 6.5. Nous proposons de comparer deux moyens de rendre

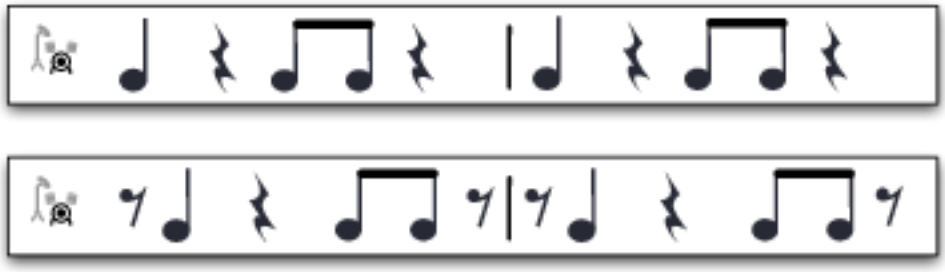

FIGURE 6.5 – *Rythmes devant être considérés comme identiques.*

notre descripteur invariant aux décalages temporels. Le premier est d'utiliser une fonction d'auto-corrélation et le second est de prendre le module d'une transformée de Fourier. Le descripteur issu du calcul utilisant l'auto-corrélation (notée $R_{xx}(u, b, \tau)$) de chaque $o(u, b, t)$ est appelé MSS ACF. Le descripteur issu du calcul utilisant le module de la transformée de Fourier discrète est appelé MSS DFT Nous proposons aussi de fusionner les sorties du MSS ACF et du MSS DFT par une technique de fusion tardive proposée plus bas. Ce descripteur sera appelé MSS ACF/DFT.

(D) **Invariance au tempo.** Enfin, une transformée d'échelle est calculée sur chaque fonction d'auto-corrélation $R_{xx}(u, b, \tau)$ (ou de chaque module de transformée de Fourier). Cette étape, avec la précédente, est l'une des plus importante de la création du descripteur. Elle permet de le rendre quasi-invariant au tempo, comme nous l'avons expliqué dans la partie 6.2.1. Nous obtenons une transformée d'échelle pour chaque bande de fréquence b et chaque trame u : $T_E(u, b, c)$ (où c est le coefficient d'échelle).

D'un point de vue pratique, la transformée d'échelle peut-être implémentée de deux façons. Dans la première, proposée par [Williams et al., 2000], elle est calculée de façon discrète (le détails des équations peut-être trouvé dans leur annexe). C'est cette méthode qui est utilisée par [Holzapfel et al., 2011 ; Prockup et al., 2015]. Dans la seconde méthode proposée par [De Sena et al., 2007], la transformée d'échelle est vue comme la transformée de Fourier d'un signal ré-échantillonné exponentiellement. Après le ré-échantillonnage (que nous présentons en annexe A), la transformée d'échelle peut donc être calculée de façon efficace par une transformée de Fourier rapide (FFT). C'est l'implémentation qui a été utilisée dans [Holzapfel et al., 2009] et dans nos travaux.

Afin d'obtenir une représentation plus adaptée au signal dans son ensemble, $T_E(u, b, c)$ est moyennée sur toutes les trames u. Cette moyenne est notée $T_E(b, c)$. Finalement, la dimension du descripteur est réduite. [Holzapfel et al., 2009] a constaté que la majorité de l'information est présente dans les premiers coefficients d'échelle. Nous gardons les C premiers coefficients sur chaque bande de fréquence. Notre descripteur a donc finalement une dimension de $b \times C$.

6.2.5 Classificateurs

Dans les expériences suivantes, nous utilisons deux classificateurs : un de type SVM [4] et un de type algorithme des k plus proches voisins ou « K-Nearest Neighbors algorithm » (KNN) modifié. Le KNN utilise une distance cosinusoïdale pour calculer la distance entre deux MSS.

KNN modifié

Nous ne prenons pas la classe majoritaire parmi les n voisins, mais nous pondérons chaque classe des voisins par un poids w_k, dépendant des valeurs des distances des voisins d_k et de la distance du voisin suivant le plus éloigné d_{n+1} :

$$w_k = 1 - \frac{d_k}{d_{n+1}} \tag{6.12}$$

Cette pondération a été choisie car elle donne dans la pratique de bien meilleurs résultats.

Distance Cosinusoïdale

Pour comparer deux MSS entre eux nous utilisons de préférence une distance cosinusoïdale plutôt qu'une distance euclidienne. Pour la comparaison de spectre en MIR, [Foote et al., 2002 ; Holzapfel et al., 2009] ont déjà montré que cette distance donne de meilleurs résultats que la distance euclidienne. Nous allons justifier pourquoi ci-dessous.

La distance cosinusoïdale est définie par :

$$d = 1 - c_{sim} \tag{6.13}$$

où c_{sim} est la similarité cosinus, elle-même définie (pour deux spectres T_{Ei} et T_{Ej}) par :

$$c_{sim} = \frac{T_{Ei}(c) \cdot T_{Ej}(c)}{|T_{Ei}(c)||T_{Ej}(c)|} \tag{6.14}$$

4. voir Annexe B pour la description des différentes méthodes de classification et de régression.

Cette métrique est invariante à la multiplication par un coefficient d'un ou des deux vecteurs. Elle mesure l'angle entre les deux vecteurs T_{Ei} et T_{Ej}.

Cette distance cosinusoïdale est meilleure qu'une distance euclidienne pour la comparaison de MSS. Elle est en effet moins sensible à la variation de l'amplitude des pics du spectre que la distance euclidienne. Elle est par contre beaucoup plus sensible à la variation de leurs positions. Autrement dit, tant que les pics sont situés à la même position, leur amplitude importe peu. Nous en déduisons donc que :

— si l'amplitude des fréquences varie, la distance cosinusoïdale variera plus lentement que la distance euclidienne,
— si les fréquences sont dilatées ou décalées, la distance cosinusoïdale variera plus vite que la distance euclidienne.

Cette distance cosinusoïdale est donc meilleure qu'une distance euclidienne pour la comparaison de MSS car nous nous intéressons plus à l'existence des pics (représentant les rythmes) qu'à leurs amplitudes.

Fusion tardive

Pour combiner plusieurs descripteurs, deux possibilités s'offrent à nous. La première, appelée fusion précoce (« early-fusion » en anglais) consiste à combiner les descripteurs avant l'étape d'apprentissage. La seconde, appelée fusion tardive (« late-fusion » en anglais) consiste à fusionner les modèles résultants des différents apprentissages.

Il n'y a pas vraiment de méthode meilleure qu'une autre. Dans la littérature, la fusion tardive est plus efficace dans certains cas [Snoek et al., 2005] et la fusion précoce l'est dans d'autres [Gunes et al., 2005]. Nous avons choisi la méthode de fusion tardive parce qu'elle est plus facile à mettre en œuvre pour des descripteurs de tailles et d'échelles différentes.

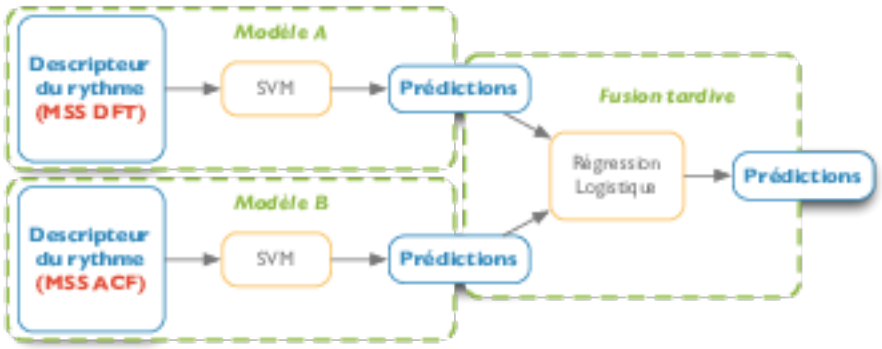

FIGURE 6.6 – *Modèle de fusion tardive entre le MSS ACF et le MSS DFT.*

La méthode la plus simple de fusion tardive est de faire une somme pondérée des différentes prédictions des classificateurs. Plusieurs possibilités pour le choix des poids des classificateurs sont alors possibles : poids uniformes, proportionnels aux performances du classificateurs, ou bien appris par un nouveau classificateur. C'est ce dernier choix que nous avons privilégié, en utilisant une Régression Logistique comme modèle d'apprentissage [5]. Nous résumons le schéma de fusion tardive que nous avons adopté sur la Figure 6.6. Cette figure présente

5. Nous utilisons une Régression Logistique car c'est une méthode habituelle et répandue dans le cadre de fusion tardive de modèles.

l'exemple particulier de la création du descripteur MSS ACF/DFT à partir des prédictions du MSS ACF et du MSS DFT.

Dans notre cadre multiclasses, nous avons binarisé les labels. Par exemple, si le jeu de données possède trois classes Samba, Tango et Valse, et que le modèle prédit Tango, nous utilisons comme vecteur d'entrée pour la fusion tardive $[0; 1; 0]$. Les résultats auraient pu être améliorés avec un modèle d'apprentissage donnant comme sortie des probabilités d'appartenance à une classe plutôt qu'un choix binaire, comme les modèles de mélanges gaussiens (GMM) ou en utilisant les affinités (distance à la marge) des SVMs. De tels modèles permettraient d'avoir en entrée de la fusion tardive des vecteurs de type $[0,1; 0,7; 0,2]$, ce qui pourrait permettre une meilleure classification finale. Nous n'avons cependant pas eu le loisir de tester cette hypothèse.

6.3 Évaluation

6.3.1 Expérience 1 : Résultats préliminaires

Nous proposons d'évaluer notre descripteur sur une expérience de reconnaissance de motifs rythmiques. L'objectif est d'estimer automatiquement la bonne classe de rythme sur les trois ensembles de données décrits dans la partie 3.4 : BALLROOM, EXTENDED BALLROOM et CRETE. Dans cette première expérience, notre descripteur MSS ACF est utilisé en entrée d'un classificateur de type SVM. Nous déterminons les meilleurs paramètres de celui-ci par une validation croisée à 10 plis.

TABLE 6.1 – *Résultats du descripteur de rythme MSS sur 3 ensembles de données.*

	BALLROOM	EXTENDED BALLROOM	CRETE
État de l'art : [Holzapfel et al., 2011]	91,7%	-	77,8%
Proposition : MSS ACF	95,0%	94.4%	77,1%

Nous présentons les résultats dans le Tableau 6.1. Notre MSS surpasse de 3.3% l'état de l'art [Holzapfel et al., 2011] sur l'ensemble de données de référence : BALLROOM. Notre descripteur donne aussi de bons résultats sur l'EXTENDED BALLROOM, surtout sachant qu'il contient six fois plus de morceaux et une nouvelle classe de rythme. Il est même plus intéressant d'avoir un rappel moyen de 94.4% sur 9 classes que de 95% sur 8 classes. Les résultats de notre MSS sont similaires à ceux de l'état de l'art sur le corpus CRETE.

Nous rappelons que les résultats de [Holzapfel et al., 2011] sont une précision alors que nous donnons un rappel moyenné sur toutes les classes. Ces résultats préliminaires montrent déjà l'intérêt de notre descripteur pour la description du rythme.

6.3.2 Expérience 2 : Comparaisons des modèles

Dans cette seconde expérience, nous explorons plus en détails notre descripteur. Nous comparons les résultats obtenus pour les deux méthodes permettant de rendre le MSS invariant aux décalages temporels (MSS ACF et MSS DFT). Nous les comparons avec la fusion tardive des modèles (MSS ACF/DFT). Enfin nous comparons deux classificateurs (KNN et SVM).

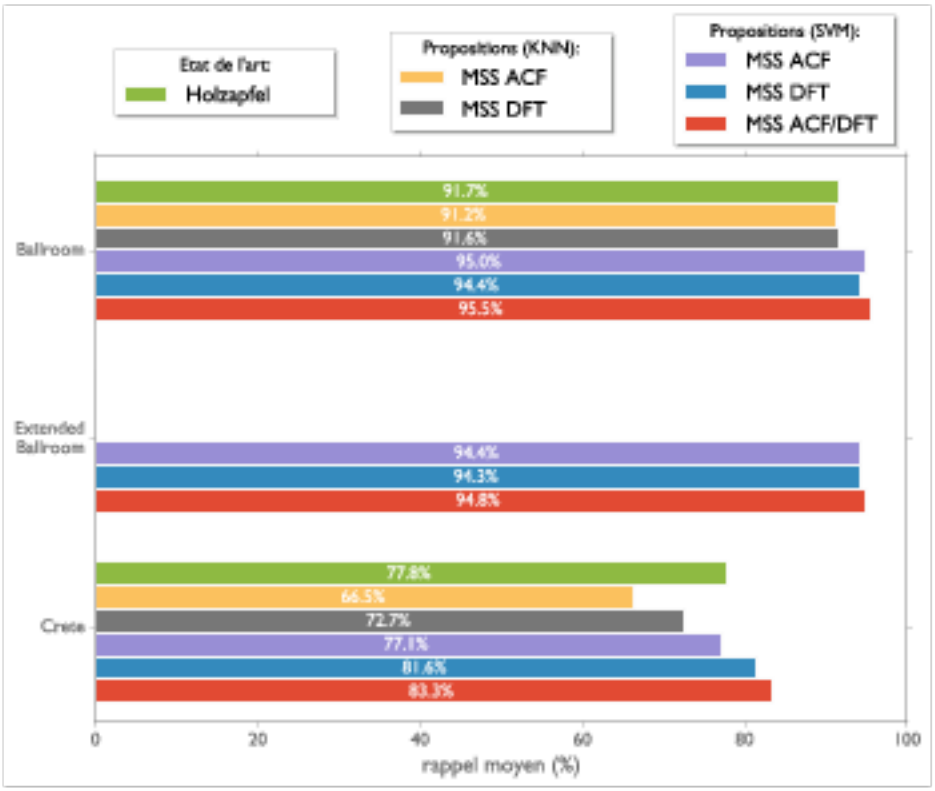

FIGURE 6.7 – *Comparaison des différentes méthodes d'invariance aux décalages temporels (ACF, DFT et fusion) et des deux méthodes d'apprentissage (KNN et SVM).*

Nous présentons dans la Figure 6.7 les résultats des trois descripteurs (ACF, DFT et ACF/DFT) appris par un modèle de type SVM, les résultats de deux descripteurs (ACF et DFT) modélisés par un classificateur de type KNN et les résultats de l'état de l'art [Holzapfel et al., 2011] pour comparaison.

Tout d'abord, nous remarquons que la méthode KNN donne des résultats équivalents à l'état de l'art sur le BALLROOM et des résultats moins bons sur le corpus CRETE. Ce dernier résultat n'est pas encourageant vu que le corpus CRETE est pour l'instant le défi principal en terme de description des motifs rythmiques. Nous n'avons donc pas essayé la méthode sur l'EXTENDED BALLROOM car les temps de calculs sont assez longs. Nous confirmons par ces observations les résultats de [Holzapfel et al., 2011] qui ont aussi montré qu'un SVM modélise mieux les spectres d'échelle qu'un KNN pour ce type d'expérience.

Deuxièmement nous comparons les méthodes d'invariance aux décalages temporels (ACF et DFT). Nous voyons que pour les corpus BALLROOM et EXTENDED BALLROOM les deux méthodes sont équivalentes. Sur le corpus CRETE, la méthode DFT fonctionne mieux (82% contre 77%). Nous voyons de plus que la fusion des deux modèles donne toujours de meilleurs résultats, c'est pourquoi dans tous les cas, il est intéressant de fusionner les deux approches.

6.3.3 Expérience 3 : Indépendance au tempo

Nous cherchons ici à montrer l'indépendance de notre descripteur par rapport au tempo. Nous proposons pour cela trois expériences. Dans la première, nous montrons que les ensembles de données que nous utilisons sont fortement biaisés du point de vue du tempo. Dans la seconde, nous contournons ce biais

en proposant une méthode d'apprentissage permettant de s'affranchir du tempo. Dans la dernière, nous montrons comment notre descripteur est complémentaire à l'information de tempo.

Tempo seul. Comme nous l'avons montré dans la partie 3.4.4, les trois corpus que nous utilisons n'ont pas une répartition de tempo homogène à travers les classes de rythmes. Le phénomène inverse est même présent : chaque classe de rythme possède un tempo prédominant. [Gouyon et al., 2004] ont montré que l'on pouvait estimer les différentes classes du corpus BALLROOM avec la seule valeur du tempo annoté et un algorithme de type KNN avec une précision de 82.3%.

Nous proposons l'expérience de classification suivante : le tempo (annoté ou estimé par la méthode de [Peeters et al., 2011]) est mis en entrée d'un algorithme de classification supervisée de type SVM pour estimer les classes de rythmes sur nos trois corpus. Nous présentons les résultats dans la Figure 6.8 et nous les comparons à notre MSS ACF.

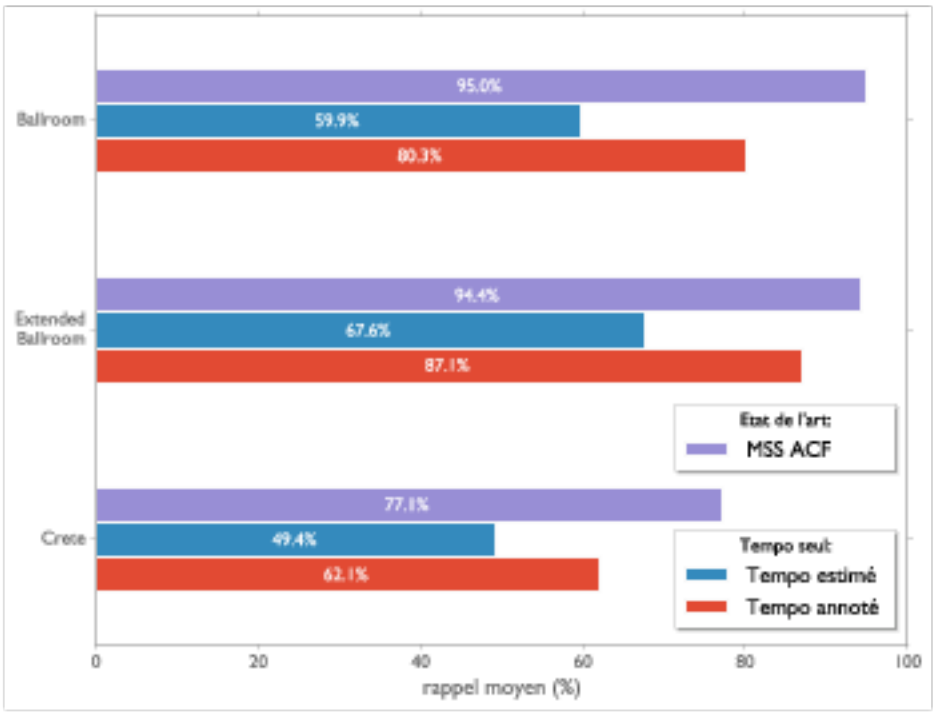

FIGURE 6.8 – *Estimation des classes de motifs rythmique à partir du tempo seul, comparée à l'état de l'art (notre proposition de MSS).*

Les résultats en utilisant le tempo seul sont plutôt bons dans l'ensemble. Le tempo annoté seul obtient 80% sur le BALLROOM (soit 15% de moins que l'état de l'art), 62% sur CRETE (15% de moins) et 87% sur l'EXTENDED BALLROOM (soit seulement 7% de moins que le MSS). Ces résultats, bien qu'inférieurs au MSS, restent très élevés pour une classification n'utilisant qu'un descripteur à une dimension et pour ce type de jeu de données. En utilisant le tempo estimé seul, les rappels moyens sont inférieurs de 10 à 20% aux résultats avec le tempo annoté.

Ces résultats montrent que l'ensemble de données CRETE est effectivement plus compliqué du point de vue du tempo que le BALLROOM et que l'EXTENDED BALLROOM, même si cela n'était pas totalement clair au vu des répartitions de

tempo des Figures 3.6, 3.8 et 3.7. Ces résultats montrent aussi que, bien que le tempo soit un descripteur de choix pour ces corpus, il n'est pas suffisant pour atteindre de très bons résultats. Ceci est rassurant, car cela signifie que même si notre MSS modélisait en partie l'information de tempo, cela n'expliquerait pas ses résultats aussi hauts.

Suppression tempos similaires. Nous proposons une deuxième expérience qui permet de mesurer à quel point notre MSS s'affranchit du tempo. L'expérience est la suivante : à considérer que notre classificateur est un KNN, nous allons ignorer dans les k plus proches voisins les morceaux dont le tempo est proche de celui de la cible (c'est-à-dire les morceaux dont le rapport de tempo avec celui de la cible est inférieur à 4%) [6].

Cette expérience est particulièrement intéressante vu la composition de nos ensembles de données. Nous comparons nos résultats à ceux de [Jensen et al., 2009] qui a proposé l'expérience, et à ceux de [Holzapfel et al., 2011] dont nous avons ré-implémenté la méthode. Nous présentons tous les résultats dans la Figure 6.9.

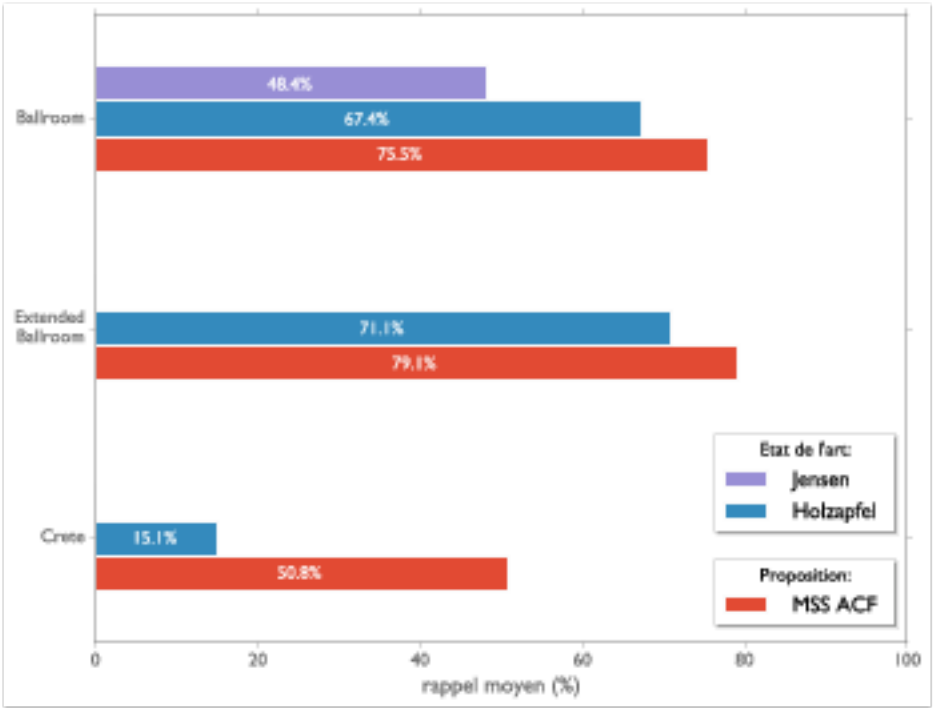

FIGURE 6.9 – *Résultats de l'expérience 'notempo'.*

Notre proposition (MSS ACF) est meilleure que l'état de l'art sur les trois ensembles de données. Le rappel moyen est de 75,5% sur le BALLROOM, ce qui est supérieur de 27% à Jensen [Jensen et al., 2009] qui a proposé l'expérience et supérieur de 8% à notre ré-implémentation de la méthode d'Holzapfel [Holzapfel et al., 2011]. Sur l'ensemble EXTENDED BALLROOM, notre méthode a un score de 79.1%, supérieur de 8% à la ré-implémentation d'Holzapfel. Sur l'ensemble CRETE, notre méthode chute à 51%. On peut cependant noter que la ré-implémentation d'Holzapfel chute elle à 15%.

6. Il est à noter que nous avons ignoré la classe Valse Viennoise de l'ensemble BALLROOM car tous ses morceaux ont un tempo égal à 5% près.

Ces résultats confirment que le corpus CRETE est plus complexe que les deux autres.

Notre proposition (MSS ACF) est bien meilleure que l'état de l'art sur cette expérience, quelque soit le corpus d'évaluation. Cependant, nous notons que ces résultats n'atteignent pas du tout les scores du MSS ACF sans la suppression des morceaux proches en tempo (Tableau 6.1). Les faibles résultats sur le corpus CRETE peuvent s'expliquer par le faible nombre de morceaux disponibles par classe. En effet chaque classe ne possède que 30 exemples, le KNN a donc du mal à trouver des voisins surtout quand on supprime une bonne partie des exemples de la classe.

Fusion tardive MSS et tempo. Nous allons enfin montrer à quel point notre MSS est complémentaire de cette information de tempo.

Nous utilisons les trois modèles MSS ACF, MSS DFT et MSS ACF/DFT décrits précédemment et deux modèles de tempo t_a, t_e. Le modèle t_a utilise comme descripteur la valeur du tempo annoté. Le modèle t_e utilise comme descripteur la valeur du tempo estimé par l'algorithme de [Peeters et al., 2011].

Nous utilisons toujours la même méthode de fusion tardive décrite précédemment pour agréger nos différents modèles. Dans le cas présent le modèle A de la Figure 6.6 est un des trois modèles (MSS ACF, MSS DFT ou MSS ACF/DFT) et le modèle B est un classificateur SVM appliqué soit sur la valeur de tempo annoté, soit sur la valeur de tempo estimé. Nous présentons les résultats dans la Figure 6.10 pour le tempo estimé et 6.11 pour le tempo annoté.

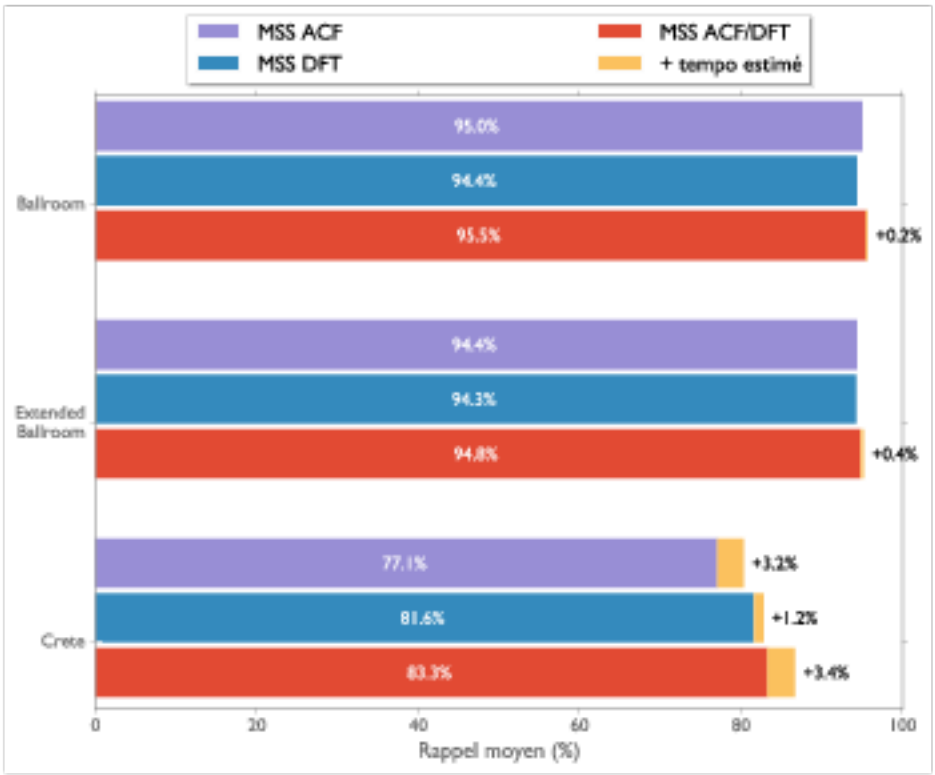

FIGURE 6.10 – *Graphique montrant le gain d'ajouter l'information de tempo estimé à chacune de nos méthodes.*

Nous observons dans la Figure 6.10 qu'ajouter l'information de tempo estimé n'améliore pas nos résultats, à part sur le corpus CRETE. Sur ce corpus, les gains

sont seulement de quelques pourcents. Nous pouvons avancer deux hypothèses pour ces résultats : soit l'information de tempo est déjà présente d'une manière ou d'une autre dans notre MSS, soit le tempo estimé n'est pas un descripteur assez fort de ces ensembles (il ne permet de classifier que les exemples faciles, qui sont déjà bien classifiés par le MSS). Nous avons donc renouvelé l'expérience en prenant le tempo annoté, afin d'exclure la première hypothèse.

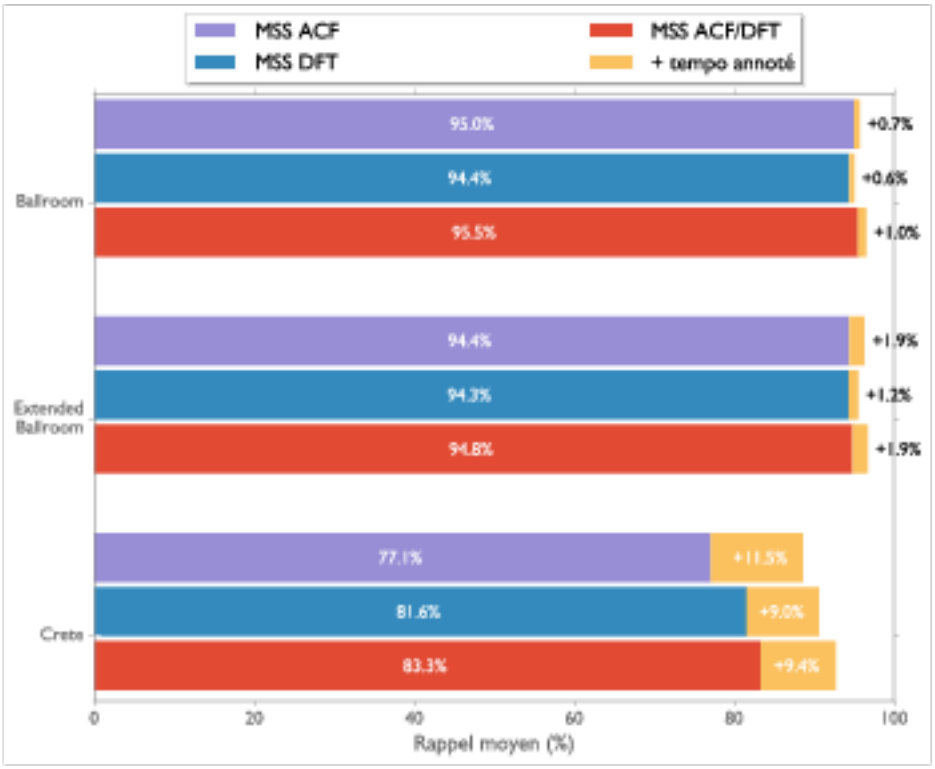

FIGURE 6.11 – *Graphique montrant le gain d'ajouter l'information de tempo annoté à chacune de nos méthodes.*

Nous présentons dans la Figure 6.11 les résultats de la fusion tardive du tempo annoté avec notre MSS. Dans l'ensemble les résultats sont très bons. Utiliser le tempo annoté par une fusion tardive permet d'améliorer d'environ 1% nos résultats sur le BALLROOM, d'environ 2% sur l'EXTENDED BALLROOM et d'environ 10% sur CRETE. Ces résultats montrent que, sur le jeu de données CRETE, l'information de tempo est bien complémentaire à notre MSS. Ce phénomène est visible moins nettement sur les ensembles BALLROOM et EXTENDED BALLROOM où les résultats du MSS sont déjà très bons de base : il ne reste donc plus beaucoup d'information à capturer pour améliorer le descripteur. Ces résultats montrant une très faibles amélioration sur BALLROOM et EXTENDED BALLROOM montrent donc que soit le MSS capture déjà l'information de tempo, soit que le MSS comme représentation des motifs rythmiques est déjà suffisant pour décrire ces ensembles. Nous penchons plutôt pour la seconde hypothèse vu les résultats sur CRETE. Les résultats sur le corpus CRETE sont très prometteurs. Ils montrent clairement que le tempo est une information complémentaire à notre MSS. En effet, on peut rappeler que le tempo annoté seul avait un rappel moyen de seulement 62% sur ce corpus, et le MSS seul a un rappel moyen d'environ 80%. Le fait de fusionner les deux fait gagner 10% de bonne classification pour passer d'environ 80% à environ 90%. Nous allons regarder en détails œs derniers résultats.

Nous montrons dans la Figure 6.12 la matrice de confusion pour le modèle MSS ACF/DFT (à gauche) que nous comparons à la matrice de confusion pour le modèle MSS ACF/DFT avec le tempo annoté sur le corpus CRETE (à droite).

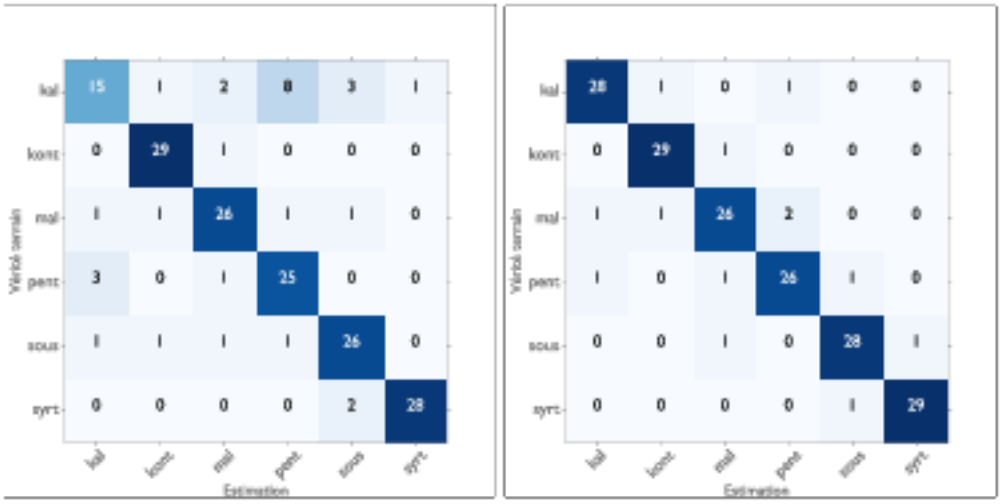

FIGURE 6.12 – *Figure de gauche : matrice de confusion pour le modèle ACF/DFT sur le corpus* CRETE. *Figure de droite : matrice de confusion pour le modèle ACF/DFT fusionné avec l'information de tempo annoté sur le corpus* CRETE.

Sans information de tempo (Figure 6.12, à gauche), nous observons que la principale source d'erreurs provient de la confusion entre *kal* et *pent* et de la mauvaise reconnaissance de la classe *kal* (seulement 50% des 30 morceaux sont reconnus en tant que KAL).

Quand nous ajoutons l'information de tempo (Figure 6.12, à droite), la plupart des problèmes sont résolus, car la classe *kal* possède un tempo très différents des autres (cf Figure 3.8). Nous notons un résultats assez intéressant : le MSS seul sépare très bien les quatre classes très proches en tempo *mal*, *pent*, *sous* et *syrt* et a plus de mal à séparer des classes dont les tempos sont très différents.

6.4 Prise en compte des inter-relations entre les bandes de fréquence

6.4.1 Limitations du MSS

Dans la partie précédente, nous avons proposé le MSS, qui permet de modéliser les informations rythmiques sur plusieurs bandes de fréquences. Il permet une meilleure modélisation du rythme que la méthode de l'état de l'art [Holzapfel et al., 2011]. En particulier, il est capable de différencier les deux rythmes de la Figure 6.13, alors que [Holzapfel et al., 2011] ne le permettait pas.

Cependant, nous avons remarqué une limitation à ce descripteur, et nous en proposons une amélioration dans cette partie. La limitation observée est illustrée par les exemples de la Figure 6.14. Les deux rythmes représentés sont différents : dans le premier la grosse caisse et la caisse claire jouent simultanément, alors que dans le second elles alternent. Malgré ses très bons résultats, le MSS n'est pas capable de distinguer ces deux rythmes car toutes les bandes de fréquences sont traitées indépendamment. Les relations entre les différents évènements musicaux ne sont donc pas toutes conservées.

FIGURE 6.13 – *Apport du MSS pour la description du rythme.*

(A) *Rythme 1* (B) *Rythme 2*

FIGURE 6.14 – *Illustration des limitations du MSS pour la description du rythme.*

Nous nous sommes intéressés à deux types de méthodes permettant de conserver les relations entre les différentes bandes de fréquences. La première est basée sur une représentation utilisée pour le traitement de l'image : la Transformée de Fourier-Mellin. Cette représentation permet de retrouver des objets dans une image, quelque soit leur plan (profondeur) et leur position : cette représentation est invariante aux décalages de positions et à l'échelle. Elle est présenté dans la partie 6.4.2. La seconde utilise les statistiques auditives proposées par [McDermott et al., 2013]. Elle est présentée dans la partie 6.4.3. Ces deux méthodes ont fait l'objet d'une publication [Marchand et al., 2016a].

6.4.2 Méthode 2DMSS

La méthode 2DMSS est basée sur la transformée de Fourier-Mellin qui correspond à une Transformée de Fourier 2D suivi d'une transformée de Mellin 2D. Nous les présentons avant de proposer notre descripteur 2DMSS.

La Transformée de Fourier 2D. Pour un signal à deux dimensions $X(t,\omega)$ (temps t, fréquences ω), la Transformée de Fourier 2D aux fréquences Ω_t et Ω_ω est définie par :

$$F(\Omega_t,\Omega_\omega) = \int_t \int_\omega X(t,\omega) e^{-j\Omega_\omega \omega - j\Omega_t t} d\omega dt \tag{6.15}$$

Cette transformée permet de modéliser les relations entre les bandes de fréquences ω par une TF, tout en modélisant aussi les relations temporelles.

La Transformée d'Échelle 2D. Comme nous l'avons vu dans la partie 6.2.1, la Transformée d'Échelle (à une dimension) est définie par :

$$S(c_t) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_0^\infty x(t) t^{-jc_t - \frac{1}{2}} dt \tag{6.16}$$

Une de ses propriétés les plus importantes est l'invariance à l'échelle.

Comme pour la Transformée de Fourier, Il est possible, de définir une version multi-dimensionnelle de la Transformée d'Échelle. Pour un signal à deux dimensions $X(t,\omega)$ sur le temps t et les fréquences ω, la Transformée d'Échelle 2D aux échelles c_t et c_ω est définie par :

$$S(c_t,c_\omega) \stackrel{\text{déf}}{=} \int_0^\infty \int_0^\infty X(t,\omega)\omega^{-jc_\omega-\frac{1}{2}}t^{-jc_t-\frac{1}{2}}d\omega dt \tag{6.17}$$

Comme pour le cas à une dimension, la TE 2D peut-être vue comme la Transformée de Fourier 2D d'un signal re-échantilloné exponentiellement sur les deux dimensions $X(e^t,e^\omega)e^{\omega/2}e^{t/2}$.

$$S(c_t,c_\omega) = \int_{-\infty}^\infty \int_{-\infty}^\infty \left(X(e^t,e^\omega)e^{\omega/2}e^{t/2}\right)e^{-jc_\omega\omega-jc_t t}d\omega dt \tag{6.18}$$

La transformée Fourier Mellin. Il est à noter que ni la transformée d'Échelle 1D, ni la transformée d'Échelle 2D ne sont invariantes aux petits décalages de leur axes : (dans le cas 1D : $|T_E[f(t)]| \neq |T_E[f(t+\alpha)]|$).

C'est pourquoi, dans la partie précédente, nous avons appliqué la TE 1D a une représentation invariante aux décalages temporels (ACF ou DFT de la fonction de détection d'onset). Dans le cadre de la TE 2D, nous allons donc l'appliquer à une représentation elle aussi invariante aux décalages : nous allons prendre le module de la transformée de Fourier 2D $F_u(\omega_t,\omega_\gamma)$).

Dans la littérature liée au traitement de l'image, une TF 2D suivie d'une TE 2D est appelée une transformée de Fourier-Mellin (Fourier-Mellin Transform). Cette transformée a été introduite dans les années 70 par la communauté de recherche en optique [Casasent et al., 1976 ; Yatagai et al., 1981] et elle a été utilisée dans de nombreux champs depuis : analyse de signaux de radar 2D [Gui-Rong et al., 1990 ; Inggs et al., 1995] ou reconnaissance de motifs dans les images [Grace et al., 1991 ; Sheng et al., 1986]. Elle n'a par contre jamais été utilisée dans le cadre de la description du signal audio.

Application à la description du rythme. Nous proposons ici une implémentation spécifique de la Transformée de Fourier-Mellin qui permet de décrire le contenu rythmique d'un signal. Les étapes successives sont résumées dans la Figure 6.15.

Tout d'abord nous créons une représentation temps-fréquence de la même manière que pour le MSS. (A) et (B) de la Figure 6.15 sont identiques à œux de la Figure 6.4.

Sur chacune des fonctions d'onset $O_u(t,\gamma_i)$ nous appliquons une transformée de Fourier Mellin, afin de modéliser conjointement les relations temporelles et fréquentielles. Nous appliquons d'abord une transformée de Fourier 2D (C') afin de rendre la représentation invariante aux décalages. Puis nous appliquons une transformée d'Échelle 2D (D') sur le module de la TF 2D pour rendre la représentation invariante à l'échelle.

6.4.3 Méthode MASSS

Notre deuxième méthode notée MASSS modélise les inter-relations entre les bandes de fréquences à l'aide de coefficients de corrélation inter-bandes.

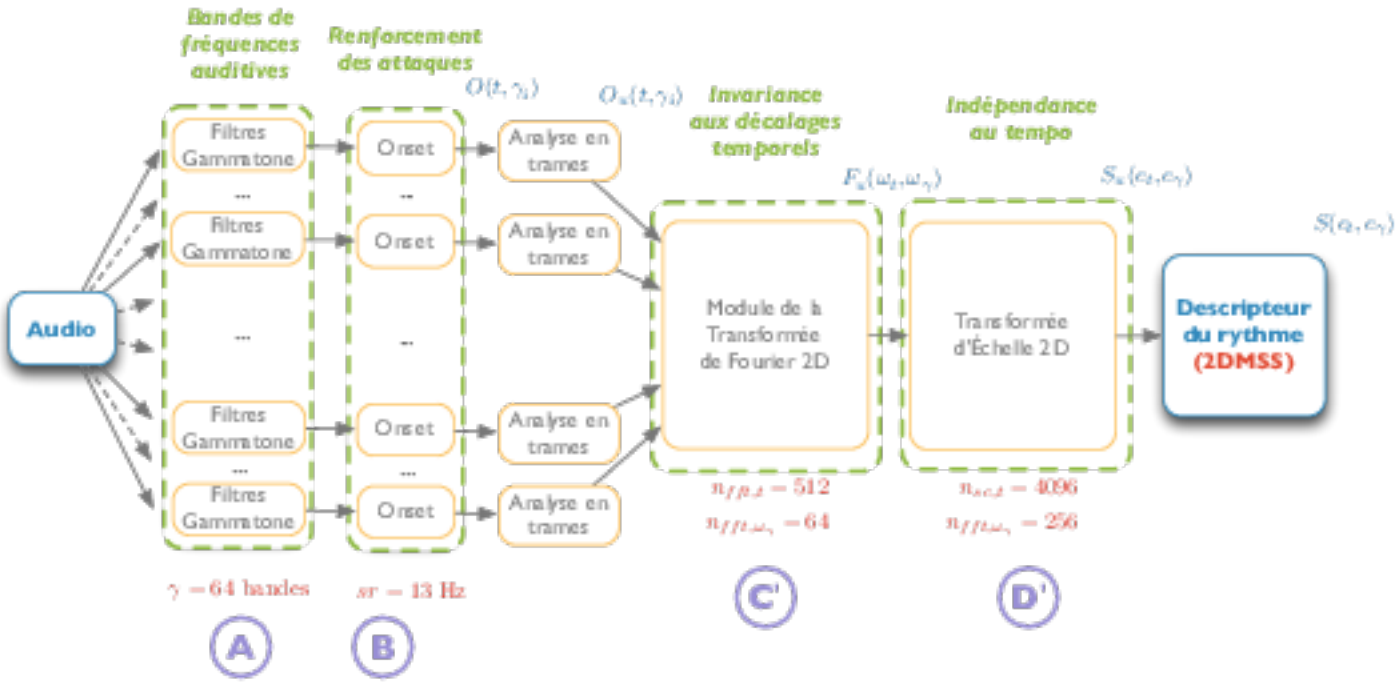

FIGURE 6.15 – *Le 2DMSS pour l'extraction des motifs rythmiques.*

Présentation de la méthode. Nous nous sommes inspirés des travaux de [McDermott et al., 2013, 2011] pour notre seconde méthode. Les auteurs montrent que le système auditif résume les informations temporelles des sons par des statistiques moyennées dans le temps (« the auditory system summarizes temporal details of sounds using time-averaged statistics »). Ils s'intéressent à la modélisation de textures sonores comme les applaudissements, le bruit d'un ruisseau, du feu ou d'insectes, … Ils proposent de modéliser les textures sonores par un ensemble de statistiques, en se basant sur un modèle auditif humain. Ce modèle est présenté sur la Figure 6.16[7]. Le système auditif est fondamentalement séparé en bandes de fréquences par les filtres de la cochlée. Les auteurs utilisent les moyennes ainsi que les corrélations entre les enveloppes du signal séparé par un banc de filtres de fréquences auditives (ce sont les premiers M et C en rouge). Après une seconde étape de filtres de modulation, ils extraient à nouveau les moyennes et les corrélations entre tous les signaux obtenus. Ils appliquent leur idée à la modélisation de textures sonores. Ils montrent que les statistiques sur les bandes de fréquences seules (c'est-à-dire sans les corrélations inter-bandes) ne suffisent pas à modéliser et re-générer des textures sonores. En revanche, l'addition des inter-corrélations entre les bandes de fréquences le permet. C'est cette idée simplifiée (modéliser les relations entre les différentes bandes de fréquences perceptives) que nous allons utiliser par la suite.

7. Ce modèle a aussi inspiré les travaux de [Ellis et al., 2011] pour la classification en genre musical. Les auteurs utilisent un banc de 18 filtres de mel, sur lequels ils calculent les corrélations croisées de leurs enveloppes. Ils comparent les résultats de ces statistiques seules, de MFCC seuls et des deux combinés. Ils proposent une expérience de classification de sons en 9 concepts (extérieur rural, extérieur urbain, intérieur calme, intérieur bruyant, piste audio, langage compréhensible, musique, acclamations et applaudissements) Ils montrent que ces statistiques seules ne sont pas meilleurs que les MFCCs, mais qu'elles permettent une petite amélioration si les deux représentations sont combinées.

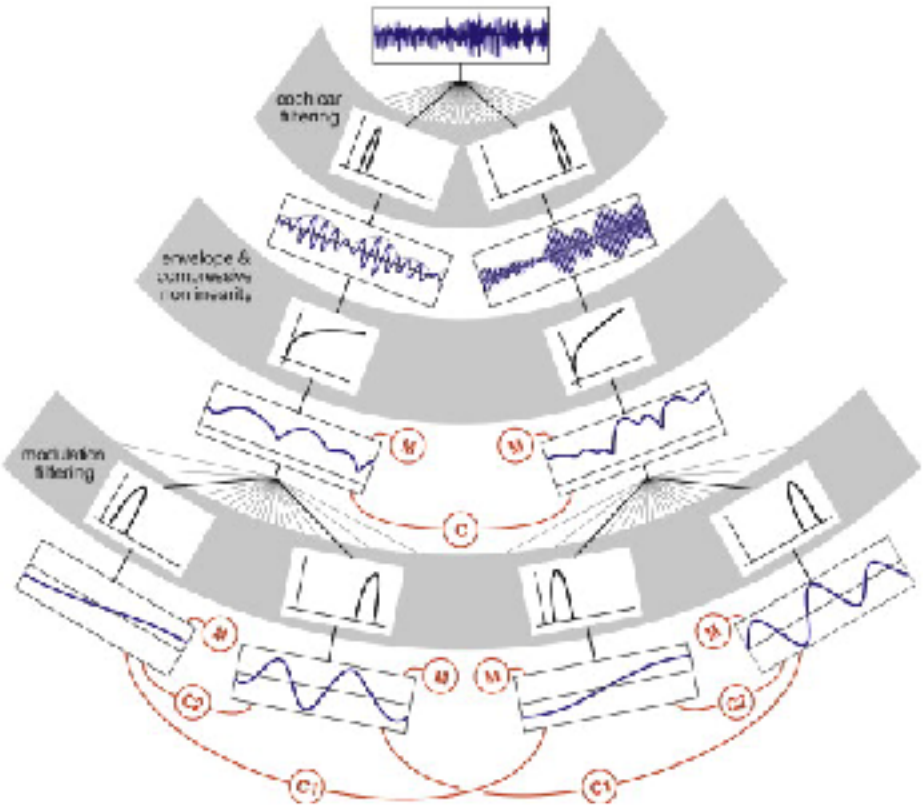

FIGURE 6.16 – *Modèle de texture auditif (« Auditory texture model »). Figure tirée de [McDermott et al., 2011].*

Application à la description du rythme. Nous proposons donc d'ajouter à notre MSS certaines corrélations entre bandes de fréquences afin de modéliser les interactions entre les différents rythmes. Les étapes successives sont présentées dans la Figure 6.17.

Tout le calcul du MSS (Ⓐ, Ⓑ, Ⓒ et Ⓓ) est identique à celui proposé dans la partie précédente (Figure 6.4 page 92). Les coefficients de corrélations croisées ccc sont en revanche nouveaux. Leur calcul est relativement simple. Le coefficient $ccc(b_i, b_j)$ mesure la corrélation qu'il y a entre les bandes de fréquences b_i et b_j. Il est calculé, entre les fonctions d'onset de ces deux bandes de fréquences, comme suit :

$$ccc(b_i, b_j) = \sum_k O(t_k, b_i) \cdot O(t_k, b_j) \tag{6.19}$$

Nous fusionnons enfin les prédictions des deux modèles MSS et ccc par la méthode de fusion tardive déjà présentée.

Différences avec la méthode de McDermott. [McDermott et al., 2013] utilisent deux étages successifs de filtrage et utilisent les moyennes et les corrélations qui en sont issues. Nous n'utilisons que le premier étage. Le filtrage cochléaire est réalisé grâce au banc de filtres gammatone. L'extraction de l'enveloppe est, quant à elle, remplacée par une fonction de détection d'onset afin de mieux s'adapter à la description du rythme.

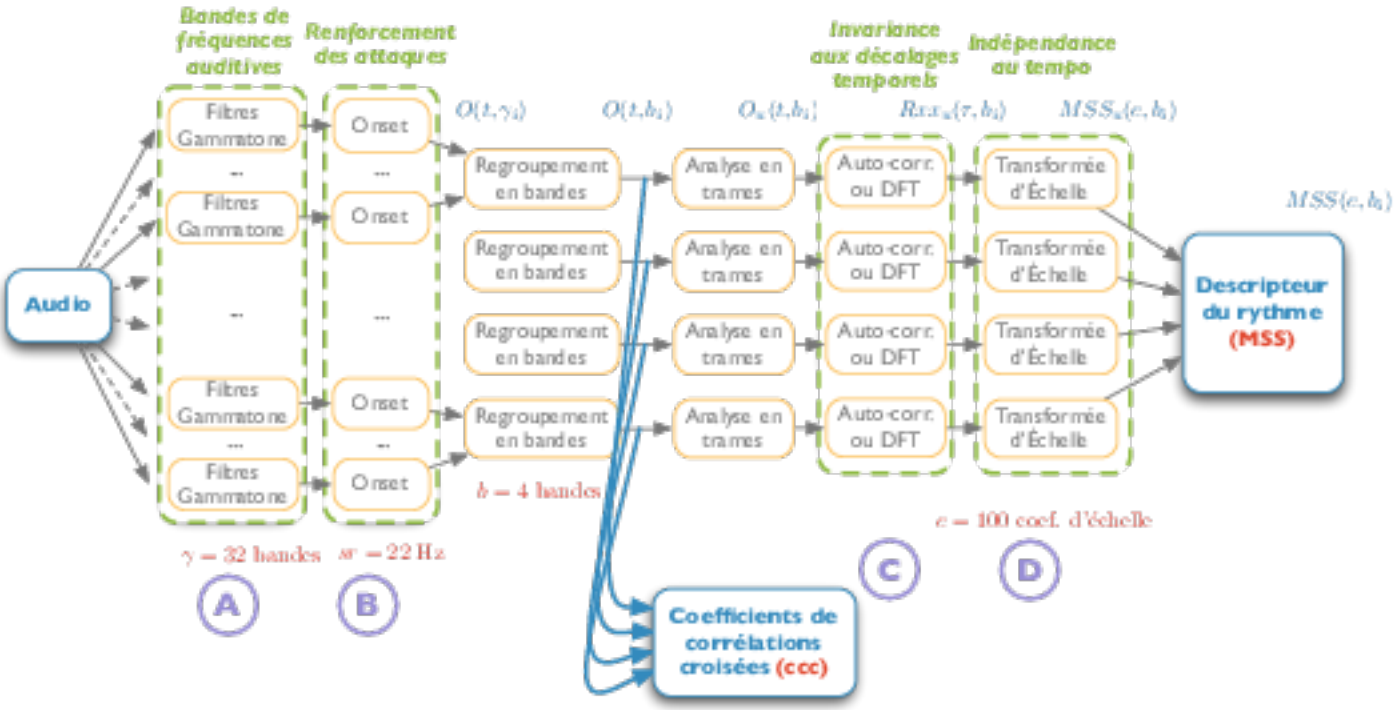

FIGURE 6.17 – *Le MASSS pour l'extraction des motifs rythmiques.*

6.4.4 Évaluation

Nous évaluons le 2DMSS et le MASSS sur la même expérience de classification en motifs rythmiques que précédemment.

2DMSS. Nous présentons les résultats de la méthode 2DMSS dans la Figure 6.18. Nous remarquons que la méthode 2DMSS donne de moins bon résultats que notre méthode précédente (MSS). Elle n'obtient que 91% sur le BALLROOM et 63% sur CRETE[8]. Nous pouvons noter néanmoins que 91% sur le BALLROOM correspond à l'état de l'art avant l'introduction de notre MSS.

Nous proposons une explication à ces résultats. Le 2DMSS permet la modélisation conjointe des axes temporels et fréquentiels. Du point de vue du temps, cette modélisation est identique au MSS, elle est invariante aux décalages temporels et aux changements d'échelle. Comme nous avons pu le voir, ces deux propriétés sont indispensables pour modéliser des motifs rythmiques. Le problème provient de notre façon de modéliser l'information fréquentielle. Du point de vue fréquentiel, le 2DMSS modélise les fréquences aussi de façon invariante aux décalages et aux changements d'échelles. L'invariance aux décalages introduit une invariance à une permutation circulaire de l'axe des fréquences. Concrètement, cela signifie que le 2DMSS ne peut pas faire la différence entre les deux rythmes présentés dans la Figure 6.19 : il produit la même représentation pour les deux rythmes de cette figure.

8. Nous n'avons pas présenté les résultats pour l'EXTENDED BALLROOM car ceux-ci n'auraient pas été bons au vu des résultats sur les autres ensembles, et que les temps de calculs sont assez longs.

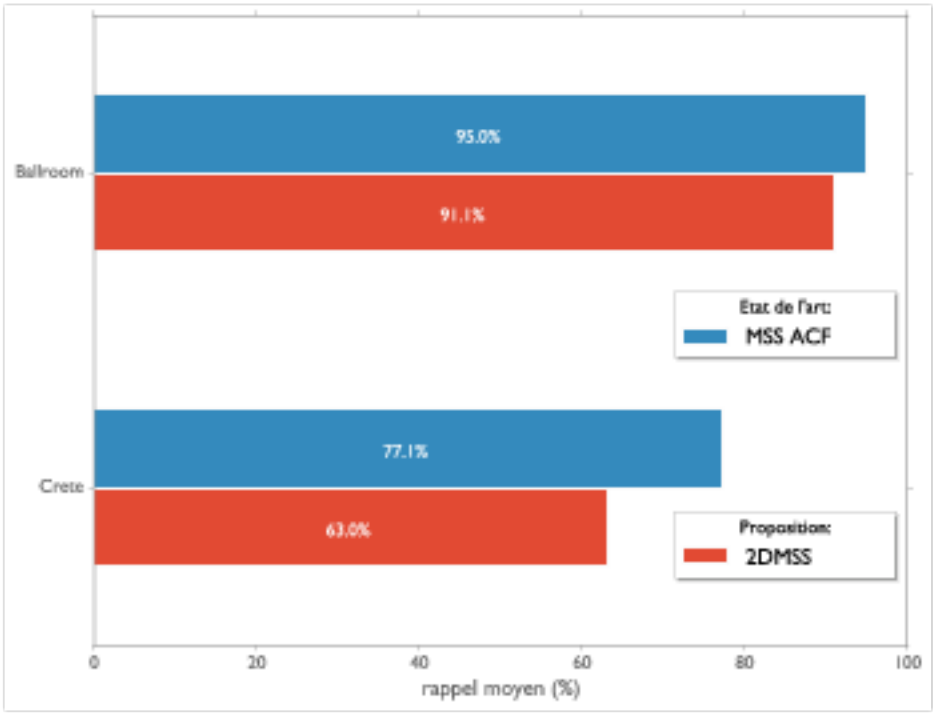

FIGURE 6.18 – *Résultats du 2DMSS comparé à l'état de l'art (notre proposition : MSS ACF).*

FIGURE 6.19 – *Limitations du 2DMSS pour la description du rythme. La méthode 2DMSS produit la même représentation pour les deux rythmes de cette figure.*

MASSS Nous présentons les résultats de la méthode MASSS dans la Figure 6.20.

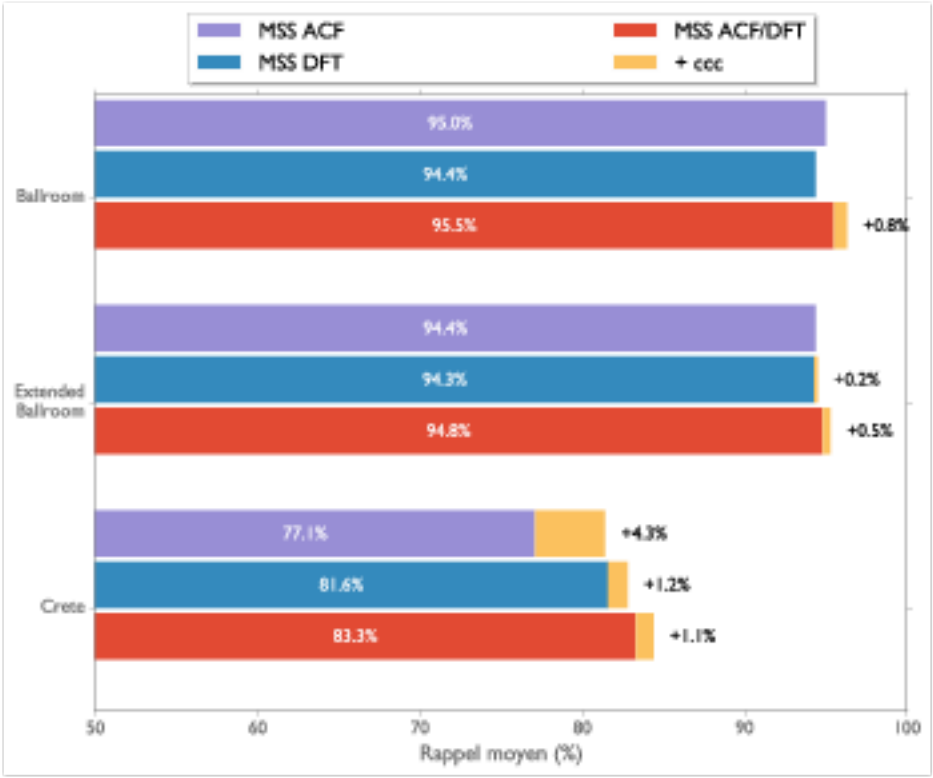

FIGURE 6.20 – *Résultats du MASSS comparé à l'état de l'art (notre proposition : MSS).*

La méthode MASSS n'améliore pas vraiment les résultats sur BALLROOM et EXTENDED BALLROOM. Elle donne par contre des résultats prometteurs sur CRETE où elle permet d'améliorer de 1 à 4% les résultats existants.

6.5 Conclusions

Dans ce chapitre, nous avons étudié la représentation des motifs rythmiques. Une telle représentation se doit de satisfaire deux contraintes : être invariante aux décalages temporels et être invariante aux changements de tempo. Pour cela, nous nous sommes appuyés sur l'invariance aux décalages temporels fournie par l'auto-corrélation ou le module de la DFT et la quasi-invariance aux changements de tempo fournie par la transformée d'échelle. Nous avons proposé le Modulation Scale Spectrum (MSS) comme l'application de la transformée d'échelle sur différentes bandes de fréquence. Nous avons montré que notre MSS donne de meilleurs résultats que l'état de l'art [Holzapfel et al., 2011] sur les deux corpus de référence BALLROOM et CRETE. Ceci démontre que la prise en compte de la localisation fréquentielle des évènements est importante pour la description des motifs rythmiques. Lors d'une expérience nous avons montré que le MSS est effectivement invariant aux changements de tempos, en particulier sur le corpus CRETE.

Le MSS modélise cependant les différentes bandes de fréquences de manière indépendante. De ce fait, il ne permet pas de distinguer les deux motifs rythmiques représentés sur la partie gauche de la Figure 6.21. Nous avons donc proposé deux méthodes 2DMSS et MASSS afin de compenser cette limitation.

Le 2DMSSS modélise les inter-relations entre les bandes de fréquence par utilisation d'une transformée de Fourier 2D suivie d'une transformée d'échelle 2D. Le 2DMSSS permet de distinguer les deux motifs rythmiques représentés sur la partie gauche de la Figure 6.21. Cependant nous avons montré que cette modélisation était insuffisante car elle est insensible à une permutation circulaire de l'axe des fréquences.

Nous avons ensuite proposé le MASSS qui est une fusion tardive du MSS et de coefficients de corrélations croisées entre les bandes de fréquences. Ces coefficients sont inspirés des expériences perceptives de [McDermott et al., 2013]. La méthode MASSS permet de faire la distinction entre tous les motifs rythmiques présentés de la Figure 6.21. Nous avons montré que notre MASSS fourni les meilleurs résultats à l'heure actuelle et ce quel que soit le corpus. Ceci démontre que la prise en compte non seulement de la localisation fréquentielle des évènements mais également de leurs inter-relations est importante pour la description des motifs rythmiques.

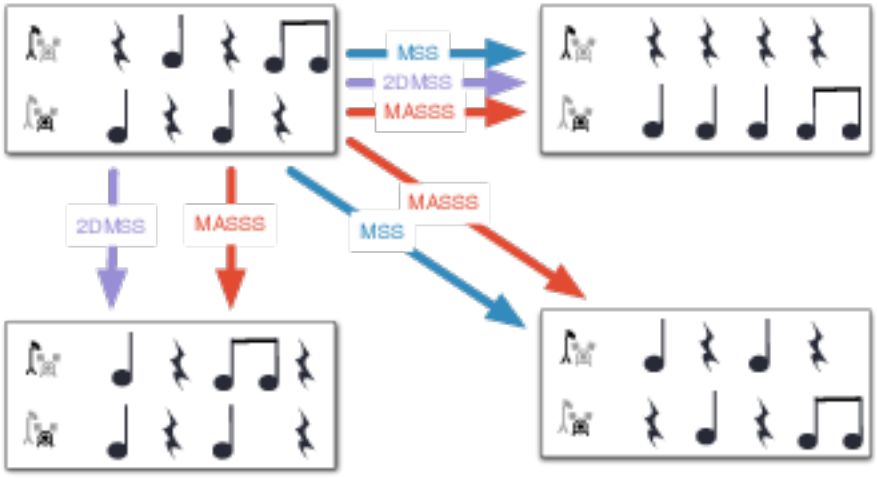

FIGURE 6.21 – *Distinctions des motifs rythmiques selon les descripteurs proposés. Les flèches symbolisent que la distinction entre les motifs rythmiques reliés est possible. Aucune flèche signifie que les motifs rythmiques ne peuvent pas distingués par la méthode.*

Publications associées

Marchand, U. et Peeters, G. (2014). « The Modulation Scale Spectrum and its Application to Rhythm-Content description ». *Proceedings of the 17th International Conference on Digital Audio Effects (Dafx)*.

Marchand, U. et Peeters, G. (2016a). « Scale and shift invariant time/frequency representation using auditory statistics : application to rhythm description. » *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

Chapitre 7

Conclusion

L'objectif de ce travail était d'étudier l'estimation automatique des paramètres relatifs au rythme à partir de l'analyse du signal musical. Nous avons cherché à montrer l'importance des études perceptives pour l'analyse automatique du rythme. Après un état de l'art approfondi sur les études perceptives liées au rythme (chapitre 2), nous avons choisi d'étudier les trois aspects principaux du rythme : tempo (chapitre 4), déviations (chapitre 5) et motifs rythmiques (chapitre 6). Nous résumons ici nos principales contributions puis nous proposons des directions pour des travaux futurs.

Corpus. Dans chacun des chapitres, nous avons porté un intérêt particulier aux corpus d'évaluation. Pour l'estimation du tempo perceptif, nous avons vérifié à la main toutes les annotations de tempo du corpus LEVY car celles-ci n'étaient pas forcement fiables étant donnée la façon dont le corpus a été constitué. Pour l'estimation du swing, nous avons nous même créé un corpus GTZAN-RHYTHM dans lequel nous avons annoté manuellement la position des battements, des 1ers temps des mesures et du swing sur un ensemble de 1000 morceaux. Enfin nous avons réalisé l'EXTENDED BALLROOM qui est une extension du corpus de référence BALLROOM pour l'estimation des motifs rythmiques.

Nous pensons qu'il est très important pour l'analyse automatique de la musique d'avoir des corpus ayant leur concepts annotés clairement définis afin d'avoir conscience des limites de ce à quoi le corpus peut servir ou ne pas servir. Nous avons donc veillé dans le chapitre 3 à décrire clairement ces trois corpus et les concepts qu'ils représentent.

Échelle logarithmique. Dans chacun de nos trois chapitres, une partie ou tous les descripteurs que nous utilisons sont basés sur une fonction d'auto-corrélation en échelle logarithmique de décalages. L'intérêt d'une telle représentation est qu'elle transforme des changements de tempo dans le signal en simple décalage du retard.

Une telle représentation est très intéressante dans le cadre de l'estimation du tempo : nous avons montré dans la partie 5 la supériorité de cette échelle logarithmique (présente dans la méthode LLACF) sur une échelle linéaire (méthode ACF) pour l'estimation de tempo.

Le ré-échantillonnage de la fonction d'auto-corrélation en échelle logarithmique est de plus une étape du calcul de notre descripteur de motifs rythmiques (MSS) qui est réalisée par une transformée d'échelle. Nous avons montré dans le chapitre 6 que la quasi-invariance au tempo obtenue grâce à cette transformée d'échelle est indispensable pour représenter des motifs rythmiques.

Lien avec la perception. Dans chacun des chapitres, nous avons pris le soin de faire des liens entre nos travaux et les études perceptives liées au rythme.

Dans le chapitre 4, nous travaillons sur le tempo perceptif (c'est-à-dire tel qu'annoté par des auditeurs) et plus particulièrement sur le problème d'ambiguïté de tempo mis en évidence par certaines études perceptives [McKinney et al., 2004, 2006 ; Moelants et al., 2004], mais qui n'a pas été étudié dans le cadre de l'analyse automatique de la musique. Grâce à notre modèle Tempo-GMM, nous validons notre hypothèse, à savoir que si l'information de tempo est partagée entre différents descripteurs du signal audio, les utilisateurs seront d'accords sur la perception du tempo. Nous montrons donc qu'il existe un lien fort entre le signal et la perception du tempo.

Dans le chapitre 5, nous faisons le lien entre les annotations de notre corpus GTZAN-Rhythm et les études perceptives montrant qu'il existe une relation linéaire entre tempo et ratio de swing. Nous montrons qu'il n'existe pas, pour notre corpus, de relation linéaire claire entre d'une part tempo et swing et d'autre part tempo et artiste. Ceci est contraire aux expériences de [Dittmar et al., 2015 ; Friberg et al., 2002].

Dans le chapitre 6, nous proposons la méthode MASSS qui utilise des coefficients de corrélation représentant les relations inter-bandes des fonctions d'onset. Ces coefficients sont inspirés des études perceptives de [McDermott et al., 2013]. Elle fournit les meilleurs résultats à l'heure actuelle en description des motifs rythmiques et ce quel que soit le corpus. Elle permet aussi de démontrer que la prise en compte non seulement de la localisation fréquentielle des évènements mais également de leurs inter-relations est importante pour la description des motifs rythmiques.

Perspectives et travaux futurs

Accents perceptifs

Nous avons étudié le tempo, les déviations et les motifs rythmiques. Pour être complet dans la description du rythme, il aurait pu être intéressant d'étudier les accents. Comme nous l'avons vu dans l'état de l'art, les accents sont à la base du rythme mais leur origine n'est pas encore totalement définie. Les accents peuvent provenir des durées, des intensités, du contour mélodique, des régularités, des hauteurs, du timbre... Pour l'instant, les fonctions d'onset se contente généralement des variations d'énergie du signal. Il pourrait être intéressant de développer de nouvelles fonctions d'onset perceptives, prenant en compte certains phénomènes complexes comme l'accentuation subjective (partie 2.2.1). Ces fonctions d'onset perceptives pourraient être utiles à un très grand nombre d'applications dans le cadre de l'analyse automatique de la musique.

Estimation du tempo perceptif

En ce qui concerne l'estimation du tempo perceptif et la prédiction *Accord*/*Désaccord*, deux axes d'améliorations sont à envisager. Le premier serait d'utiliser d'autres représentations du signal (descripteurs). Nous avons vu par exemple que les fonctions d'onset, de similarité à court-terme et de balance spectral fonctionnaient bien, contrairement à la fonction d'harmonicité. Nous n'avons pas du tout travaillé sur le choix des descripteurs et il serait intéressant

de proposer de nouvelles fonctions de périodicité représentant de nouvelles caractéristiques du signal. On pourrait évidemment se servir des fonctions d'onset perceptives décrites ci-dessus. Le deuxième axe d'amélioration serait d'étendre le modèle pour estimer le tempo en fonction des auditeurs. Le principal frein actuellement est le faible nombre d'exemples (15 auditeurs ayant plus de annoté 10 extraits) du corpus LEVY. Il serait nécessaire d'étendre l'expérience perceptive de [Levy, 2011] pour élargir le corpus.

Estimation du ratio de swing

La principale limitation de nos méthodes d'estimation du ratio de swing est qu'elles nécessitent de connaître le tempo préalablement ou au moins d'en avoir une estimation fiable. Les travaux futurs devraient s'attacher à produire un descripteur n'ayant pas besoin de connaître quel est le niveau métrique du tempo.

Motifs rythmiques

Le MSS permet une très bonne description des motifs rythmiques. Il serait intéressant maintenant de voir à quel point il est utile pour d'autres tâches en MIR comme la classification en genre musical, l'estimation de tempo ou de structure musicale. En effet, nous pensons qu'une bonne description du rythme peut vraiment bénéficier à toutes les tâches de l'analyse automatique de la musique.

Annexe A

Re-échantillonnage exponentiel

Notre implémentation du ré-échantillonnage exponentiel est tirée de [De Sena et al., 2007]. Nous utiliserons cette même méthode dans la partie 5.2.3 pour calculer la fonction log-lag-auto-correlation et dans la partie 6 pour prendre des échantillons espacés exponentiellement ($x(e^t)$) pour le calcul de la transformée d'échelle.

Méthode. La première étape consiste à créer un axe temporel avec des échantillons espacé de façon logarithmique. La deuxième étape consiste à interpoler le signal $x(t)$ sur œ nouvel axe temporel.

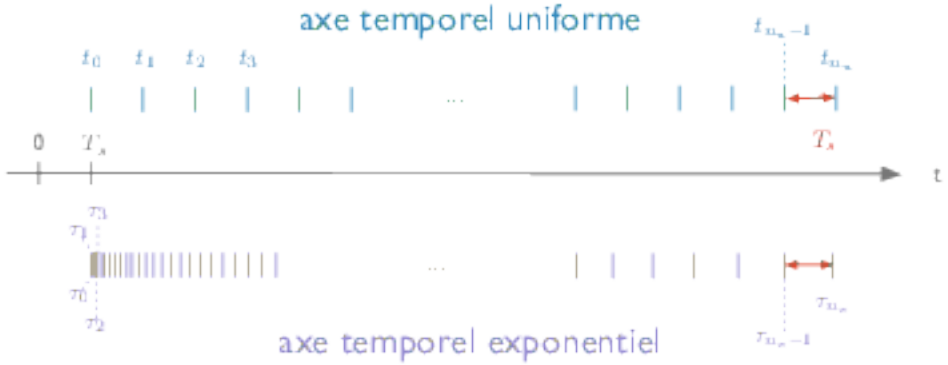

FIGURE A.1 – *Axes uniformes et exponentiels.*

Notations. Nous notons :

- t_n le n-ième échantillon de l'axe temporel.
- τ_k le k-ième échantillon de l'axe exponentiel.
- $\tau_0 = t_0$ le premier point temporel qui est commun aux deux axes mais différent de 0.
- T_s le pas d'échantillonnage uniforme.
- τ_s le pas d'échantillonnage exponentiel.
- n_u le nombre d'échantillons de l'axe uniforme.
- n_e le nombre d'échantillons de l'axe exponentiel.

Choix du pas exponentiel. Concrètement, le choix du pas exponentiel τ_s doit se faire en gardant à l'esprit deux conditions : la période T_s est le pas maximal entre deux échantillons (Nyquist-Shannon), il faut couvrir tout le signal (cette condition n'est pas triviale, puisque pour couvrir le signal à partir de 0, il faudrait une infinité d'échantillons exponentiels).

La première contrainte implique que les deux derniers échantillons de chaque axe sont confondus, $t_{n_u} = \tau_{n_e}$ et $t_{n_u-1} = \tau_{n_e-1}$ Nous pouvons ensuite définir le pas exponentiel τ_s comme le rapport entre deux échantillons exponentiels consécutifs. En particulier, τ_s est donc le rapport entre le dernier et l'avant-dernier échantillon de l'axe exponentiel.

$$\tau_s = \frac{\tau_{n_e}}{\tau_{n_e-1}} = \frac{t_{n_u}}{t_{n_u-1}} = \frac{t_0 + n_u.T_s}{t_0 + (n_u - 1)T_s} \tag{A.1}$$

Nombre d'échantillons exponentiels Comme nous ne pouvons pas à la fois fixer la première contrainte et à la fois fixer $\tau_0 = t_0$, nous utilisons l'idée de [De Sena et al., 2007] : nous allons d'abord calculer le nombre d'échantillons de l'axe exponentiel grâce à τ_s, puis nous allons en déduire un nouveau τ_s à partir du nombre d'échantillons. Ce nouveau τ_s sera quasiment identique au précédent mais il permettra de satisfaire :

$$\begin{aligned} \tau_0 &= t_0 \\ \tau_{n_e} &= t_{n_u} \\ \tau_{n_e-1} &\approx t_{n_u-1} \end{aligned} \tag{A.2}$$

La deuxième contrainte nous permet de calculer le nombre d'échantillons de l'axe exponentiel n_e en fonction du nombre d'échantillons de l'axe uniforme n_u. En pratique, nous prenons $t_0 \overset{\text{déf}}{=} \tau_0 \overset{\text{déf}}{=} T_s$ afin de réduire le nombre d'échantillons exponentiels.

Cette condition donne, après quelques calculs le nombre d'échantillons de l'axe exponentiel :

$$n_e = \frac{\ln(n_u + 1)}{\ln\left(\frac{n_u+1}{n_u}\right)} + 1 \tag{A.3}$$

Quand n_u est grand, cette équation s'approxime par :

$$n_e = n_u \ln(n_u) \tag{A.4}$$

Ré-échantillonage Le nouveau pas exponentiel τ_s est donc :

$$\tau_s = \log \frac{\tau_{n_e}}{\tau_0} \tag{A.5}$$

Nous pouvons calculer les positions des instants temporels τ_k :

$$\begin{aligned} \tau_0 &= T_s \\ \tau_k &= \tau_0.(\tau_s)^k \end{aligned} \tag{A.6}$$

Nous interpolons finalement les valeurs de x sur l'axe exponentiel des τ_k. Cette interpolation est basée sur les algorithmes de [Dierckx, 1993].

Annexe B

Apprentissage

Il existe beaucoup de méthodes d'apprentissage statistique permettant de modéliser un concept à partir d'un ensemble de descripteurs. L'objectif de ce manuscrit n'est pas de trouver les couples méthodes d'apprentissage et paramètres qui fonctionnent le mieux pour une tâche donnée, mais de trouver des descripteurs ayant le plus de sens possible, afin de simplifier au maximum l'étape de modélisation. Nous n'essayerons donc pas toutes les méthodes possibles pour chacun de nos problèmes, comme il a pu être fait par exemple par [Peeters, 2011]. Cela aura l'intérêt de grandement simplifier la présentation des résultats, même si c'est peut-être au détriment de quelques pourcents de classification.

Dans ce manuscrit, nous allons être confrontés à deux types de problèmes, des problèmes de classification (en classe de motifs rythmiques) et des problèmes de régression (en valeur du ratio de swing ou du tempo perceptif). Nous décrivons ici tous les outils que nous allons utiliser par la suite.

B.1 Classification

Pour les problèmes de classification, nous utilisons les méthodes de type KNN, SVM et de régression logistique.

KNN. La classification par la méthode des plus proches voisins est une méthode qui ne cherche pas à construire un modèle généralisant les données d'apprentissage mais qui stocke simplement celles-ci. La classification se fait par un vote majoritaire au sein des plus proches voisins du point demandé : la classe choisie est celle qui est la plus représentée parmi les k plus proches voisins. C'est une des méthodes de classification les plus intuitive.

SVM. La classification par SVM est une méthode d'apprentissage supervisée. Elle repose sur deux idées clé.

La première est la notion de marge maximale. La marge est la distance entre la frontière de séparation et les exemples les plus proches (qui sont appelés vecteurs supports). L'idée est de choisir la frontière de séparation qui maximise la marge. Intuitivement, une bonne séparation est obtenue avec la plus grande distance aux exemples d'entraînement : plus cette marge est grande, moins le classificateur fera d'erreurs de généralisation. Le problème est de trouver cette frontière séparatrice optimale, à partir d'un ensemble d'apprentissage. Ceci est fait en formulant le problème comme un problème d'optimisation quadratique, pour lequel il existe des algorithmes connus.

Afin de pouvoir traiter des cas où les données ne sont pas linéairement séparables, la deuxième idée clé des SVM est de transformer l'espace de représentation des données d'entrée en un espace de plus grande dimension (voire de

dimension infinie), dans lequel il est probable qu'il existe une séparation linéaire de ces données. Ceci est réalisé grâce à une fonction noyau. Les fonctions noyau qui respectent certaines conditions (théorème de Mercer) permettent de transformer un produit scalaire dans un espace de grande dimension qui est coûteux, en une simple évaluation ponctuelle d'une fonction (cette technique est connue sous le nom de kernel trick).

L'avantage des SVM est qu'ils sont capables de s'adapter à une grande variété de données d'apprentissage moyennant une recherche de paramètres appropriés.

Régression logistique. La régression logistique, contrairement à son nom, n'est pas une méthode de régression, mais bien une méthode de classification binaire. Elle est très utilisée dans le cadre de fusion tardive de modèles car elle est très adaptée à la modélisation de données binaires. Le modèle peut être résumé par l'équation suivante :

$$\ln \frac{p(X|1)}{p(X|0)} = a_0 + a_1 x_1 + \ldots + a_J x_J \tag{B.1}$$

Il s'agit bien d'une « régression » car on veut montrer une relation de dépendance entre une variable à expliquer p et une série de variables explicatives x_i. Il s'agit d'une régression « logistique » car la loi de probabilité est modélisée à partir d'une loi logistique.

B.2 Régression

Pour les problèmes de régression, nous utilisons les méthodes de type SVM et de type GMM.

SVR. La méthode SVM peut être étendue pour résoudre des problèmes de régression, que l'on appelle SVR ou « Support Vector Reggression ». Elle a été introduite par [Drucker et al., 1996].

GMM. Un GMM est un modèle probabiliste qui suppose que tous les points de données à modéliser sont générés à partir d'un nombre fini de distributions gaussiennes à paramètres inconnus. Une régression GMM utilise les différents paramètres des distributions gaussiennes estimées pour prédire une valeur continue. Dans le cas simple d'une régression avec un GMM à une composante, la fonction de prédiction est simplement le vecteur moyenne de cette composante. Dans le cas à plusieurs composantes, la fonction de prédiction dépend aussi des probabilités que le vecteur de données à estimer appartienne aux différentes composantes du mélange.

Liste des publications

Marchand, U., Fresnel, Q. et Peeters, G. (2015a). « GTZAN-Rhythm : Extending the GTZAN Test-Set with Beat, Downbeat and Swing Annotations ». *Late-Breaking-Demo Session of the 16th International Society for Music Information Retrieval Conference (ISMIR)*.

Marchand, U. et Peeters, G. (2014). « The Modulation Scale Spectrum and its Application to Rhythm-Content description ». *Proceedings of the 17th International Conference on Digital Audio Effects (Dafx)*.

— (2015b). « Swing ratio estimation ». *Proceedings of the 18th International Conference on Digital Audio Effects (Dafx)*. Trondheim, Norway.

— (2016a). « Scale and shift invariant time/frequency representation using auditory statistics : application to rhythm description. » *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

— (2016b). « The Extended Ballroom Dataset ». *Late-Breaking-Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR)*.

Peeters, G. et Marchand, U. (2013). « Predicting agreement and disagreement in the perception of tempo ». *Proceedings of the 10th International Symposium on Computer Music Modeling and Retrieval*, p. 253–266.

— (2014b). « Predicting agreement and disagreement in the perception of tempo ». *Sound, Music, and Motion*. Springer, p. 313–329.

Bibliographie

Abecasis, D., Brochard, R., Granot, R. et Drake, C. (2005). « Differential brain response to metrical accents in isochronous auditory sequences ». *Music Perception* 22.3, p. 549–562.

Antonopoulos, I., Pikrakis, A., Theodoridis, S., Cornelis, O., Moelants, D. et Leman, M. (2007). « Music retrieval by rhythmic similarity applied on greek and african traditional music ». *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*.

Atlas, L. et Shamma, S. A. (2003). « Joint acoustic and modulation frequency ». *EURASIP Journal on Applied Signal Processing* 2003, p. 668–675.

Bartsch, M. A. et Wakefield, G. H. (2005). « Audio thumbnailing of popular music using chroma-based representations ». *IEEE Transactions on Multimedia* 7.1, p. 96–104.

Bello, J. P., Duxbury, C., Davies, M. et Sandler, M. (2004). « On the use of phase and energy for musical onset detection in the complex domain ». *IEEE Signal Processing Letters* 11.6, p. 553–556.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. et Sandler, M. B. (2005). « A tutorial on onset detection in music signals ». *IEEE Transactions on Speech and Audio Processing* 13.5, p. 1035–1047.

Benetos, E. et Dixon, S. (2011). « Polyphonic music transcription using note onset and offset detection ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 37–40.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. et Klapuri, A. (2013). « Automatic music transcription : challenges and future directions ». *Journal of Intelligent Information Systems* 41.3, p. 407–434.

Bennetzen, J. et Maegaard, K. (1982). « Reggae ». *Fontes Artis Musicae*, p. 182–186.

Bilmes, J. A. (1993). « Timing is of the essence : Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm ». Mém.de mast. Massachusetts Institute of Technology.

Böck, S., Arzt, A., Krebs, F. et Schedl, M. (2012a). « Online real-time onset detection with recurrent neural networks ». *Proceedings of the 15th International Conference on Digital Audio Effects (Dafx)*, p. 1–4.

Böck, S., Krebs, F. et Schedl, M. (2012b). « Evaluating the Online Capabilities of Onset Detection Methods. » *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, p. 49–54.

Brochard, R., Abecasis, D., Potter, D., Ragot, R. et Drake, C. (2003). « The "Ticktock" of Our Internal Clock Direct Brain Evidence of Subjective Accents in Isochronous Sequences ». *Psychological Science* 14.4, p. 362–366.

Brown, J. C. (1991). « Calculation of a constant Q spectral transform ». *The Journal of the Acoustical Society of America* 89.1, p. 425–434.

Brown, J. C. et Puckette, M. S. (1989). « Calculation of a "narrowed" autocorrelation function ». *The Journal of the Acoustical Society of America* 85.4, p. 1595–1601.

Calinon, S. (2009). *Robot Programming by Demonstration : A Probabilistic Approach*. EPFL Press ISBN 978-2-940222-31-5, CRC Press ISBN 978-1-4398-0867-2. EPFL Press.

Cano, P. et al. (2006). « ISMIR 2004 audio description contest ». *Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep.*

Casasent, D. et Psaltis, D. (1976). « Scale invariant optical transform ». *Optical Engineering* 15.3, p. 153258–153258.

Cemgil, A. T., Kappen, B., Desain, P. et Honing, H. (2000). « On tempo tracking : Tempogram representation and Kalman filtering ». *Journal of New Music Research* 29.4, p. 259–273.

Chen, C.-W., Cremer, M., Lee, K., DiMaria, P. et Wu, H.-H. (2009). « Improving Perceived Tempo Estimation by Statistical Modeling of Higher-Level Musical Descriptors ». *Audio Engineering Society Convention 126*. Audio Engineering Society.

Chen, J. L., Penhune, V. B. et Zatorre, R. J. (2008). « Moving on time : brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training ». *Journal of Cognitive Neuroscience* 20.2, p. 226–239.

Clarke, E. F. (1999). « Rhythm and timing in music ». *The psychology of music* 2, p. 473–500.

Cohen, L. (1993). « The scale representation ». *IEEE Transactions on Signal Processing* 41.12, p. 3275–3292.

Collier, G. L. et Wright, C. E. (1995). « Temporal rescaling of simple and complex ratios in rhythmic tapping. » *Journal of Experimental Psychology : Human Perception and Performance* 21.3, p. 602.

Collins, N. (2005). « A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions ». *Audio Engineering Society Convention 118*. Audio Engineering Society.

Cooper, G. (1963). *The rhythmic structure of music*. T. 118. University of Chicago Press.

De Sena, A. et Rocchesso, D. (2007). « A fast Mellin and scale transform ». *EURASIP Journal on Advances in Signal Processing* 2007.

Degara, N., Davies, M. E., Pena, A. et Plumbley, M. D. (2011). « Onset event decoding exploiting the rhythmic structure of polyphonic music ». *IEEE Journal of Selected Topics in Signal Processing* 5.6, p. 1228–1239.

Desain, P. et Honing, H. (1994). « Does expressive timing in music performance scale proportionally with tempo ? » *Psychological Research* 56.4, p. 285–292.

Desain, P., Honing, H., Vanthienen, H. et Windsor, L. (1998). « Computational modeling of music cognition : problem or solution? » *Music Perception*, p. 151–166.

Dierckx, P. (1993). *"Curve and surface fitting with splines", Monographs on Numerical Analysis*. Oxford University Press.

Dittmar, C., Pfleiderer, M. et Müller, M. (2015). « Automated estimation of ride cymbal swing ratio in jazz recordings ». *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*.

Dittmar, C. et Uhle, C. (2004). « Further steps towards drum transcription of polyphonic music ». *Audio Engineering Society Convention 116*. Audio Engineering Society.

Dixon, S. (2001). « An interactive beat tracking and visualisation system ». *Proceedings of the International Computer Music Conference (ICMC)*, p. 215–218.

— (2006). « Onset detection revisited ». *Proceedings of the 9th International Conference on Digital Audio Effects (Dafx)*, p. 133–137.

Dixon, S., Goebl, W. et Cambouropoulos, E. (2006). « Perceptual smoothness of tempo in expressively performed music ». *Music Perception*.

Dixon, S., Gouyon, F. et Widmer, G. (2004). « Towards Characterisation of Music via Rhythmic Patterns ». *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)*.

Dixon, S., Pampalk, E. et Widmer, G. (2003). « Classification of dance music by periodicity patterns. » *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*.

Drake, C. et Bertrand, D. (2001). « The quest for universals in temporal processing in music ». *Annals of the New York Academy of Sciences* 930.1, p. 17–27.

Drake, C., Jones, M. R. et Baruch, C. (2000a). « The development of rhythmic attending in auditory sequences : attunement, referent period, focal attending ». *Cognition* 77.3, p. 251–288.

Drake, C., Penel, A. et Bigand, E. (2000b). « Rhythm perception and production ». Sous la dir. de W. L. Desain P. Swets & Zeitlinger, Lisse. Chap. Why musicians tap slower than nonmusicians, p. 245–248.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J. et Vapnik, V. (1996). « Support vector regression machines ». *Advances in neural information processing systems* 9, p. 155–161.

Duxbury, C., Bello, J. P., Davies, M., Sandler, M. et al. (2003). « Complex domain onset detection for musical signals ». *Proceedings of the 6th International Conference on Digital Audio Effects (Dafx)*. 1, p. 6–9.

Ellis, D. P. (2007). « Beat tracking by dynamic programming ». *Journal of New Music Research* 36.1, p. 51–60.

Ellis, D. P., Zeng, X. et McDermott, J. H. (2011). « Classifying soundtracks with audio texture features ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 5880–5883.

Elowsson, A. et Friberg, A. (2015). « Modeling the perception of tempo ». *The Journal of the Acoustical Society of America* 137.6, p. 3163–3177.

Eppler, A., Männchen, A., Abeßer, J., Weiß, C. et Frieler, K. (2014). « Automatic Style Classification of Jazz Records with Respect to Rhythm, Tempo, and Tonality ». *Proceeding of the Conference on Interdisciplinary Musicology (CIM)*. Berlin, Germany.

Essens, P. J. (1986). « Hierarchical organization of temporal patterns ». *Perception & Psychophysics* 40.2, p. 69–73.

Essens, P. J. et Povel, D.-J. (1985). « Metrical and nonmetrical representations of temporal patterns ». *Perception & Psychophysics* 37.1, p. 1–7.

Eyben, F., Böck, S., Schuller, B. et Graves, A. (2010). « Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks. » *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, p. 589–594.

Fitzgerald, D. (2010). « Harmonic/percussive separation using median filtering ». *Proceedings of the 16th International Conference on Digital Audio Effects (Dafx)*. Dublin Institute of Technology.

Flocon-Cholet, J. (2012). « Estimation du tempo perceptif et réduction des erreurs d'octave du tempo ». Mém.de mast. Université Pierre et Marie Curie.

Foote, J. (1999). « Visualizing music and audio using self-similarity ». *Proceedings of the 7th ACM international conference on Multimedia (Part 1)*. ACM, p. 77–80.

Foote, J. (2000). « Automatic Audio Segmentation Using a Measure of Audio Novelty ». *Proceeding of the IEEE International Conference on Multimedia and Expo (ICME)*. T. 1. IEEE, p. 452–455.

Foote, J., Cooper, M. L. et Nam, U. (2002). « Audio Retrieval by Rythmic Similarity ». *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)*.

Foote, J. et Uchihashi, S. (2001). « The Beat Spectrum : A New Approach To Rhythm Analysis. » *Proceeding of the IEEE International Conference on Multimedia and Expo (ICME)*.

Fraisse, P. (1974). *Psychologie du rythme*. Presses universitaires de France Paris.

Friberg, A. et Sundström, A. (2002). « Swing ratios and ensemble timing in jazz performance : Evidence for a common rhythmic pattern ». *Music Perception* 19.3, p. 333–349.

Fujioka, T., Trainor, L. J., Large, E. W. et Ross, B. (2012). « Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations ». *The Journal of Neuroscience* 32.5, p. 1791–1802.

Gillet, O. et Richard, G. (2005). « Extraction and remixing of drum tracks from polyphonic music signals ». *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, p. 315–318.

Gkiokas, A., Katsouros, V. et Carayannis, G. (2012). « Reducing Tempo Octave Errors by Periodicity Vector Coding And SVM Learning. » *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, p. 301–306.

Goto, M. (2001). « An audio-based real-time beat tracking system for music with or without drum-sounds ». *Journal of New Music Research* 30.2, p. 159–171.

— (2006). « AIST Annotation for the RWC Music Database. » *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, p. 359–360.

Goto, M. et Muraoka, Y. (1996). « Beat tracking based on multiple-agent architecture-a real-time beat tracking system for audio signals ». *Proceedings of the 2nd International Conference on Multiagent Systems*, p. 103–110.

Gouyon, F. (2005). « A computational approach to rhythm description-Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing ». Thèse de doct. Universitat Pompeu Fabra.

Gouyon, F., Dixon, S., Pampalk, E. et Widmer, G. (2004). « Evaluating rhythmic descriptors for musical genre classification ». *Proceedings of the25th Audio Engineering Society International Conference (AES)*, p. 196–204.

Gouyon, F., Fabig, L. et Bonada, J. (2003a). « Rhythmic expressiveness transformations of audio recordings : swing modifications ». *Proceedings of the 6th International Conference on Digital Audio Effects (Dafx)*.

Gouyon, F. et Herrera, P. (2003b). « Determination of the meter of musical audio signals : Seeking recurrences in beat segment descriptors ». *Audio Engineering Society Convention 114*. Audio Engineering Society.

Gouyon, F., Herrera, P. et Cano, P. (2002). « Pulse-dependent analyses of percussive music ». *Audio Engineering Society Conference : 22nd International Conference : Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society.

Grace, A. et Spann, M. (1991). « A comparison between Fourier-Mellin descriptors and moment based features for invariant object recognition using neural networks ». *Pattern Recognition Letters* 12.10, p. 635–643.

Grahn, J. A. et Brett, M. (2007). « Rhythm and beat perception in motor areas of the brain ». *Journal of Cognitive Neuroscience* 19.5, p. 893–906.

Gruhne, M. et Dittmar, C. (2009). « Improving rhythmic pattern features based on logarithmic preprocessing ». *Audio Engineering Society Convention 126*. Audio Engineering Society.

Gui-Rong, G., Wen-Xian, Y. et Wei, Z. (1990). « An intelligence recognition method of ship targets ». *Fuzzy Sets and Systems* 36.1, p. 27–36.

Gunes, H. et Piccardi, M. (2005). « Affect recognition from face and body : early fusion vs. late fusion ». *IEEE international conference on systems, man and cybernetics*. T. 4. IEEE, p. 3437–3443.

Helen, M. et Virtanen, T. (2005). « Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine ». *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)*. T. 2005.

Henry George, L. et Robert, S. (1999). *A Greek-English Lexicon*. http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057:entry=r(uqmo/s.

Hockman, J., Davies, M. E. et Fujinaga, I. (2012). « One in the Jungle : Downbeat Detection in Hardcore, Jungle, and Drum and Bass. » *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, p. 169–174.

Hockman, J. et Fujinaga, I. (2010). « Fast vs Slow : Learning Tempo Octaves from User Data. » *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, p. 231–236.

Hofmann-Engl, L. (2002). « Rhythmic similarity : A theoretical and empirical approach ». *Proceedings of the 7th International Conference on Music Perception and Cognition*, p. 564–567.

Holzapfel, A., Davies, M. E., Zapata, J. R., Oliveira, J. L. et Gouyon, F. (2012). « Selective sampling for beat tracking evaluation ». *IEEE Transactions on Audio, Speech, and Language Processing* 20.9, p. 2539–2548.

Holzapfel, A. et Stylianou, Y. (2009). « A scale transform based method for rhythmic similarity of music ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 317–320.

Holzapfel, A. et Stylianou, Y. (2011). « Scale transform in rhythmic similarity of music ». *IEEE Transactions on Audio, Speech, and Language Processing* 19.1, p. 176–185.

Honing, H. (2001). « From time to time : The representation of timing and tempo ». *Computer Music Journal* 25.3, p. 50–61.

Honing, H., Bouwer, F. L. et Háden, G. P. (2014). « Perceiving temporal regularity in music : The role of auditory event-related potentials (ERPs) in probing beat perception ». *Neurobiology of interval timing*. Springer, p. 305–323.

Honing, H. et De Haas, W. B. (2008). « Swing once more : Relating timing and tempo in expert jazz drumming ». *Music Perception*.

Humphrey, E. J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R. M. et Bello, J. P. (2014). « JAMS : a JSON annotated music specification for reproducible MIR research ». *Int. Society for Music Information Retrieval Conf(ISMIR 2014)*.

Inggs, M. et Robinson, A. (1995). « Neural approaches to ship target recognition ». *Proceedings of th IEEE International Radar Conference*. IEEE, p. 386–391.

Jehan, T. (2004). « Event-synchronous music analysis/synthesis ». *Proceedings of the 7th International Conference on Digital Audio Effects (Dafx)*.

Jensen, J. H., Christensen, M. G. et Jensen, S. H. (2009). « A tempo-insensitive representation of rhythmic patterns ». *Proceedings of the 17th European Signal Processing Conference (EUSIPCO)*. IEEE.

Jones, M. R. et Boltz, M. (1989). « Dynamic attending and responses to time. » *Psychological review* 96.3, p. 459.

Klapuri, A. (1999). « Sound onset detection by applying psychoacoustic knowledge ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. T. 6. IEEE, p. 3089–3092.

Klapuri, A., Eronen, A. et Astola, J. (2006). « Analysis of the meter of acoustic musical signals ». *IEEE Transactions on Audio, Speech, and Language Processing* 14.1, p. 342–355.

Krebs, F., Böck, S. et Widmer, G. (2013). « Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio. » *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, p. 227–232.

Lacoste, A. et Eck, D. (2006). « A supervised classification algorithm for note onset detection ». *EURASIP Journal on Advances in Signal Processing* 2007.

Large, E. W. et Jones, M. R. (1999). « The dynamics of attending : How people track time-varying events. » *Psychological review* 106.1, p. 119.

Large, E. W. et Palmer, C. (2002). « Perceiving temporal regularity in music ». *Cognitive Science* 26.1, p. 1–37.

Large, E. W. et Snyder, J. S. (2009). « Pulse and meter as neural resonance ». *Annals of the New York Academy of Sciences* 1169.1, p. 46–57.

Laroche, J. (2001). « Estimating tempo, swing and beat locations in audio recordings ». *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, p. 135–138.

Lerdahl, F. et Jackendoff, R. (1985). *A generative theory of tonal music*. MIT press.

Levenberg, K. (1944). « A method for the solution of certain non–linear problems in least squares ». *Quarterly of Applied Mathematics*, p. 164–168.

Levy, M. (2011). « Improving perceptual tempo estimation with crowd-sourced annotations ». *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*.

Ma, N., Green, P., Barker, J. et Coy, A. (2007). « Exploiting correlogram structure for robust speech recognition with multiple speech sources ». *Speech Communication* 49.12, p. 874–891.

Madison, G. et Merker, B. (2002). « On the limits of anisochrony in pulse attribution ». *Psychological Research* 66.3, p. 201–207.

Marchand, U., Fresnel, Q. et Peeters, G. (2015a). « GTZAN-Rhythm : Extending the GTZAN Test-Set with Beat, Downbeat and Swing Annotations ». *Late-Breaking-Demo Session of the 16th International Society for Music Information Retrieval Conference (ISMIR)*.

Marchand, U. et Peeters, G. (2014). « The Modulation Scale Spectrum and its Application to Rhythm-Content description ». *Proceedings of the 17th International Conference on Digital Audio Effects (Dafx)*.

— (2015b). « Swing ratio estimation ». *Proceedings of the 18th International Conference on Digital Audio Effects (Dafx)*. Trondheim, Norway.

— (2016a). « Scale and shift invariant time/frequency representation using auditory statistics : application to rhythm description. » *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

— (2016b). « The Extended Ballroom Dataset ». *Late-Breaking-Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR)*.

Marquardt, D. W. (1963). « An algorithm for least-squares estimation of nonlinear parameters ». *Journal of the Society for Industrial & Applied Mathematics* 11.2, p. 431–441.

Mauch, M. et al. (2009). « OMRAS2 metadata project 2009 ». *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*.

McDermott, J. H., Schemitsch, M. et Simoncelli, E. P. (2013). « Summary statistics in auditory perception ». *Nature neuroscience* 16.4, p. 493–498.

McDermott, J. H. et Simoncelli, E. P. (2011). « Sound texture perception via statistics of the auditory periphery : Evidence from sound synthesis ». *Neuron* 71.5, p. 926–940.

McKay, C., Fujinaga, I. et Depalle, P. (2005). « jAudio : A feature extraction library ». *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*.

McKinney, M. F. et Breebaart, J. (2003). « Features for audio and music classification. » *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*. T. 3, p. 151–158.

McKinney, M. F. et Moelants, D. (2004). « Deviations from the resonance theory of tempo induction ». *Proceeding of the Conference on Interdisciplinary Musicology (CIM)*.

— (2006). « Ambiguity in tempo perception : What draws listeners to different metrical levels? » *Music Perception* 24.2, p. 155–166.

Merchant, H., Zarco, W., Pérez, O., Prado, L. et Bartolo, R. (2011). « Measuring time with different neural chronometers during a synchronization-continuation task ». *Proceedings of the National Academy of Sciences* 108.49, p. 19784–19789.

Meudic, B. (2004). « Détermination automatique de la pulsation, de la métrique et des motifs musicaux dans des interprétations à tempo variable d'oeuvres polyphoniques ». Thèse de doct. Paris 6.

Moelants, D. et McKinney, M. (2004). « Tempo perception and musical content : What makes a piece fast, slow or temporally ambiguous ». *Proceedings of the 8th International Conference on Music Perception and Cognition*, p. 558–562.

Noorden, L. van et Moelants, D. (1999). « Resonance in the perception of musical pulse ». *Journal of New Music Research* 28.1, p. 43–66.

Ono, N., Miyamoto, K., Kameoka, H. et Sagayama, S. (2008a). « A Real-time Equalizer of Harmonic and Percussive Components in Music Signals. » *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, p. 139–144.

Ono, N., Miyamoto, K., Le Roux, J., Kameoka, H. et Sagayama, S. (2008b). « Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram ». *Proceedings of the 16th European Signal Processing Conference (EUSIPCO)*.

Palmer, C. (1997). « Music performance ». *Annual review of psychology* 48.1, p. 115–138.

Pampalk, E., Rauber, A. et Merkl, D. (2002). « Content-based organization and visualization of music archives ». *Proceedings of the 10th ACM international conference on Multimedia*. ACM, p. 570–579.

Parncutt, R. (1994). « A perceptual model of pulse salience and metrical accent in musical rhythms ». *Music Perception*, p. 409–464.

Patel, A. D. (2014). « The evolutionary biology of musical rhythm : was Darwin wrong? » *PLoS Biol* 12.3, e1001821.

Patel, A. D., Iversen, J. R., Chen, Y. et Repp, B. H. (2005). « The influence of metricality and modality on synchronization with a beat ». *Experimental Brain Research* 163.2, p. 226–238.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. et Allerhand, M. (1992). « Complex sounds and auditory images ». *Auditory physiology and perception* 83, p. 429–446.

Paulus, J. et Klapuri, A. (2002). « Measuring the similarity of Rhythmic Patterns. » *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)*.

Peeters, G. (2005). « Rhythm Classification Using Spectral Rhythm Patterns. » *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, p. 644–647.

— (2006). « Template-based estimation of time-varying tempo ». *EURASIP Journal on Advances in Signal Processing* 2007.

— (2007). « Sequence Representation of Music Structure Using Higher-Order Similarity Matrix and Maximum-Likelihood Approach. » *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, p. 35–40.

— (2011). « Spectral and Temporal Periodicity Representation of Rhythm for the Automatic Classification of Music Audio Signal ». *IEEE Transactions on Audio, Speech, and Language Processing* 19.5, p. 1242–1252.

Peeters, G. et Bisot, V. (2014a). « Music Structure Segmentation using lag-priors ». *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*.

Peeters, G. et Flocon-Cholet, J. (2012a). « Perceptual tempo estimation using GMM-regression ». *Proceedings of the 2nd international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, p. 45–50.

Peeters, G. et Fort, K. (2012b). « Towards A (Better) Definition Of The Description Of Annotated M.I.R. Corpora ». *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*.

Peeters, G. et Marchand, U. (2013). « Predicting agreement and disagreement in the perception of tempo ». *Proceedings of the 10th International Symposium on Computer Music Modeling and Retrieval*, p. 253–266.

— (2014b). « Predicting agreement and disagreement in the perception of tempo ». *Sound, Music, and Motion*. Springer, p. 313–329.

Peeters, G. et Papadopoulos, H. (2011). « Simultaneous beat and downbeat-tracking using a probabilistic framework : theory and large-scale evaluation ». *IEEE Transactions on Audio, Speech, and Language Processing* 19.6, p. 1754–1769.

Platon (-350). *Loi, 665a*.

Potter, D., Fenwick, M., Abecasis, D. et Brochard, R. (2009). « Perceiving rhythm where none exists : Event-Related Potential (ERP) correlates of subjective accenting ». *Cortex* 1.45, p. 103–109.

Povel, D.-J. et Essens, P. (1985). « Perception of temporal patterns ». *Music perception*, p. 411–440.

Povel, D.-J. et Okkerman, H. (1981). « Accents in equitone sequences ». *Perception & Psychophysics* 30.6, p. 565–572.

Prockup, M., Ehmann, A. F., Gouyon, F., Schmidt, E. M. et Kim, Y. E. (2015). « Modeling musical rhythm at scale with the music Genome project ». *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–5.

Quinton, E., Harte, C. et Sandler, M. (2015). « Extraction of Metrical Structure from Music Recordings ». *Proceedings of the 18th International Conference on Digital Audio Effects (Dafx)*. Trondheim, Norway.

Ramona, M. et Peeters, G. (2013). « AudioPrint : An efficient audio fingerprint system based on a novel cost-less synchronization scheme ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 818–822.

Randel, D. M. (1945). *The Harvar dictionary of music*. Harvard University Press.

Rigaud, F., Lagrange, M., Robel, A. et Peeters, G. (2011). « Drum extraction from polyphonic music based on a spectro-temporal model of percussive sounds ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 381–384.

Röbel, A. (2005). « Onset detection in polyphonic signals by means of transient peak classification ». *Music Information Retrieval Evaluation eXchange (MIREX)*.

Rodet, X., Worms, L. et Peeters, G. (2003). *Method For Characterinz a Sound Signal*. WO Patent 2,003,056,455.

Sakai, K., Hikosaka, O., Miyauchi, S., Takino, R., Tamada, T., Iwata, N. K. et Nielsen, M. (1999). « Neural representation of a rhythm depends on its interval ratio ». *The Journal of Neuroscience* 19.22, p. 10074–10081.

Scheirer, E. D. (1998). « Tempo and beat analysis of acoustic musical signals ». *The Journal of the Acoustical Society of America* 103.1, p. 588–601.

— (2000). « Music-listening systems ». Thèse de doct. Massachusetts Institute of Technology.

Schloss, W. A. (1985). *On the automatic transcription of percussive music : from acoustic signal to high-level analysis*. 27. Stanford University.

Schubotz, R. I., Friederici, A. D. et Von Cramon, D. Y. (2000). « Time perœption and motor timing : a common cortical and subcortical basis revealed by fMRI ». *Neuroimage* 11.1, p. 1–12.

Sheng, Y. et Arsenault, H. H. (1986). « Experiments on pattern recognition using invariant Fourier–Mellin descriptors ». *JOSA A* 3.6, p. 771–776.

Smith, L. M. et Honing, H. (2008). « Time–frequency representation of musical rhythm by continuous wavelets ». *Journal of Mathematics and Music* 2.2, p. 81–97.

Snoek, C. G., Worring, M. et Smeulders, A. W. (2005). « Early versus late fusion in semantic video analysis ». *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, p. 399–402.

Stevens, S. S., Volkmann, J. et Newman, E. B. (1937). « A scale for the measurement of the psychological magnitude pitch ». *The Journal of the Acoustical Society of America* 8.3, p. 185–190.

Stober, S., Cameron, D. J. et Grahn, J. A. (2014a). « Classifying EEG recordings of rhythm perception ». *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, p. 649–654.

— (2014b). « Does the beat go on ? : identifying rhythms from brain waves recorded after their auditory presentation ». *Proceedings of the 9th Audio Mostly : A Conference on Interaction With Sound*. ACM, p. 23.

Stowell, D. et Plumbley, M. (2007). « Adaptive whitening for improved real-time audio onset detection ». *Proceedings of the International Computer Music Conference (ICMC)*. T. 18. Citeseer.

Sturm, B. L. (2013). « The GTZAN dataset : Its contents, faults, and their effects on music genre recognition evaluation ». *IEEE Transactions on Audio, Speech and Language Processing*.

Tolonen, T. et Karjalainen, M. (2000). « A computationally efficient multipitch analysis model ». *Speech and Audio Processing, IEEE Transactions on* 8.6, p. 708–716.

Tsunoo, E., Ono, N. et Sagayama, S. (2009a). « Rhythm map : Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals ». *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 185–188.

Tsunoo, E., Tzanetakis, G., Ono, N. et Sagayama, S. (2009b). « Audio genre classification using percussive pattern clustering combined with timbral features ». *Proceeding of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, p. 382–385.

Tzanetakis, G. et Cook, P. (2002). « Musical genre classification of audio signals ». *IEEE transactions on Speech and Audio Processing* 10.5, p. 293–302.

Uhle, C., Dittmar, C. et Sporer, T. (2003). « Extraction of drum tracks from polyphonic music using independent subspace analysis ». *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*. Citeseer, p. 843–847.

Völkel, T., Abeßer, J., Dittmar, C. et Großmann, H. (2010). « Automatic genre classification of latin american music using characteristic rhythmic patterns ». *Proceedings of the 5th Audio Mostly Conference : A Conference on Interaction with Sound*. ACM, p. 16.

Waadeland, C. H. (2000). « Rhythmic movements and moveable rhythms : syntheses of expressive timing by means of rhythmic frequency modulation ». Thèse de doct. Norwegian University of Science et Technology.

Whitman, B. et Ellis, D. P. (2004). « Automatic record reviews ». *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)*. Audiovisual Institute, Pompeu Fabra University, p. 86–93.

Williams, W. et Zalubas, E. (2000). « Helicopter transmission fault detection via time-frequency, scale and spectral methods ». *Mechanical systems and signal processing* 14.4, p. 545–559.

Winkler, I., Háden, G. P., Ladinig, O., Sziller, I. et Honing, H. (2009). « Newborn infants detect the beat in music ». *Proceedings of the National Academy of Sciences* 106.7, p. 2468–2471.

Worms, L. (1998). *Reconnaissance d'extraits sonores dans une large base de donnees*. Practical lessons.

Wright, M., Schloss, W. A. et Tzanetakis, G. (2008). « Analyzing Afro-Cuban Rhythms using Rotation-Aware Clave Template Matching with Dynamic Programming. » *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, p. 647–652.

Xiao, L., Tian, A., Li, W. et Zhou, J. (2008). « Using Statistic Model to Capture the Association between Timbre and Perceived Tempo. » *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, p. 659–662.

Yatagai, T., Choji, K. et Saito, H. (1981). « Pattern classification using optical Mellin transform and circular photodiode array ». *Optics Communications* 38.3, p. 162–165.

Yeston, M. (1976). *The stratification of musical rhythm*. Yale University Press New Haven, CT.

Yoshii, K., Goto, M. et Okuno, H. G. (2004). « Automatic Drum Sound Description for Real-World Music Using Template Adaptation and Matching Methods. »

Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR), p. 184–191.

Zapata, J. R., Holzapfel, A., Davies, M. E., Oliveira, J. L. et Gouyon, F. (2012). « Assigning a confidence threshold on automatic beat annotation in large datasets ». *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*.

Zhou, R. et Reiss, J. D. (2011). « Music Onset Detection ». *Machine Audition*, p. 297.

Zils, A., Pachet, F., Delerue, O. et Gouyon, F. (2002). « Automatic extraction of drum tracks from polyphonic music signals ». *Proceeding of the 2nd International Conference on Web Delivering of Music (WEDELMUSIC)*. IEEE, p. 179–183.

Zwicker, E. (1961). « Subdivision of the audible frequency range into critical bands (Frequenzgruppen) ». *The Journal of the Acoustical Society of America* 33.2, p. 248–248.