Université Pierre et Marie Curie
Mémoire de Stage de Master 2^{eme} année
Sciences de l'ingénieur mention ATIAM
2006-2007

Sylvain Cadars

Modélisation temporelle et synthèse concatenative de boucles rythmiques

Mars - Septembre 2007

Laboratoire d'accueil : IRCAM - Equipe Interaction Musicale Temps Réel
Responsables : Norbert Schnell et Diemo Schwarz

Remerciements

Mes remerciements s'adressent à Diemo Schwarz sans qui ce travail n'aurait pu aboutir. Je le remercie aussi vivement pour sa patience et sa disponibilité.

Je tiens aussi à exprimer ma reconnaissance à Guillaume Lemaître, Antoine Minard, Karine Aura, Olivier Houix, Nicolas Misdariis, Aymeric Devergie pour les nombreux moments passés ensemble tout au long de ce stage, leur bonne humeur, leur sens du partage et de la communication.
Mais également, Patrick Susini de m'avoir fait une place dans son équipe durant ces 6 mois.

J'adresse aussi un grand merci à Anthony Sypniewsky, Remy Muller, Tommaso Bianco pour leurs conseils et surtout l'aide précieuse qu'ils ont su me donner au long de ce stage.
Et enfin, je remercie Norbert Schnell pour m'avoir accueilli dans son équipe.

Table des matières

Introduction

L'utilisation de bases de données composées de boucles rythmiques pour des projets musicaux nécessite généralement une large bibliothèque d'échantillons. Cependant, il s'avère souvent plus satisfaisant de construire nos propres boucles rythmiques de quelque étendue que puisse être la base de données. En effet, une répétition excessive d'un même motif rythmique est souvent source d'ennui. En partant de ce constat, nous avons pensé qu'un outil capable de générer, transformer et créer de nouveaux rythmes en temps réel à partir d'une base de sons serait une aide précieuse pour la création d'oeuvres musicales. Ainsi, la tâche à laquelle nous avons décidé de nous consacrer est la génération automatique de motifs rythmiques pour un système informatique. Elle est ardue. En effet, un tel système — définit par Russel comme dynamique, non déterministe, continu et non épisodique — fait appel à l'intelligence artificielle. Par ailleurs, l'ordinateur doit accomplir une tâche créative par l'intégration de contraintes ou de restrictions musicales, lors de la modélisation. Cette tâche créative s'apparente à la phase d'improvisation chez le musicien. Pour y parvenir, le musicien fait appel à des éléments de son histoire et de sa culture. Ce sont précisément ces éléments qui permettent de faire évoluer le rythme au cours du temps par la modification des événements successifs qui le compose tout en conservant sa structure. En effet, la perception du rythme est faite toute à la fois dans la perception de structures et de leurs répétitions. Mais pour éviter l'ennui, ces répétitions doivent intégrer de subtiles variations, rendant ainsi, le rythme "vivant". De plus, lorsque le musicien complète la fin d'une figure rythmique, toutes les poursuites sont acceptables dans la mesure où celles-ci s'intègrent dans le contexte de départ. Ainsi, un outil informatique visant à reproduire ces phénomènes se doit d'intégrer des contraintes lui permettant de prendre en compte ces aspects.

Pour nous aider dans cette tâche, notre travail entre dans le cadre des recherches exploratoires, visant à l'amélioration des outils temps réels utilisant le principe de la synthèse concaténative par corpus, en réponse aux difficultés représentées par la modélisation d'un motif rythmique. En effet, il existe deux grandes approches au problème de la génération automatique de motifs rythmiques pour un système informatique. La première consiste à développer un modèle capable de générer des séquences à partir de rien et la seconde résulte de l'ajout de fragments de musique récupérés d'une librairie ou d'une base de son pour la création de nouveaux rythmes.
L'utilisation de la seconde approche, propre à la synthèse par concaténation, possède les avantages suivant :

- chaque fragment sonore contient, en lui même, un certain savoir musical.
- le modèle peut être transposé sur d'autres instruments en utilisant d'autres corpus.

Ainsi, chaque fragment sonore permet de restituer les performances et l'expressivité du

musicien et c'est l'un des atouts majeur de ce mode de synthèse. En effet, la synthèse concaténative est issue des techniques de composition par collage manuel d'extraits sonores vers le milieu du 20^{e} siècle, période musicale en référence à la musique concrète initié par Pierre Schaeffer [Pie], et Oswald [Joh]. Elle rejoint aussi une période de recherche artistique axée sur la décomposition atomique de la matière*(Klein)*, où la synthèse granulaire [Cur88] permettait une exploration en temps réel d'un enregistrement avec libre déplacement de la tête de lecture sur des petites particules sonores. Par la suite, les recherches autour de la synthèse de la parole ont permis sa modélisation à partir d'un texte dont le principe est fondé sur la concaténation de phonèmes contenus dans de grandes bases de données. Mais cette méthode de synthèse a permis aussi de mettre au point des outils d'aide à la création artistique.
Parmi, ces outils, nous utilisons pour notre étude, le logiciel CataRT, développé par Diemo Schwarz.

CataRT utilise une grande base de données de sons segmentés, appelé corpus, en unités représentées par des descripteurs. Il permet une navigation totalement libre à travers cet espace de descripteurs et offre de vastes possibilités créatives, accrues par les choix multiples de segmentation du signal audio et par ses différentes possibilités de contrôle : tablettes graphiques, interfaces utilisant le protocole MIDI ou bien OSC.

L'état actuel de cataRT offre une grande flexibilité à l'artiste qui l'utilise soit lors de la phase compositionnelle, soit lors d'une performance live. D'ailleurs, plusieurs compositeurs ont utilisés CataRT pour des oeuvres musicales : *swarming essence pour orchestre et électronique* de Dai Fujikura, *Distance liquide* de Hans Tutschku, *Whisper Not* de Stephano Gervasoni.

En revanche, l'exploration d'un corpus rassemblant des éléments percussifs soulève encore de grandes interrogations. Lorsque l'on souhaite atteindre certaines unités du corpus, la transition est continue, autrement dit, si l'on se situe vers le bas de l'espace du descripteur choisi et que l'on souhaite entendre une unité située à l'opposée de cet espace, nous serons contraints de synthétiser toutes les unités situées sur le chemin entre ces deux points. Par cette action, la perception sonore qui en résulte est semblable à l'effondrement d'une armoire à vaisselle. Il n'est donc pas possible de sélectionner telle ou telle unité à la volée pour construire un rythme.

Ainsi, la problématique de ce stage repose dans les thématiques suivantes :
- La modélisation temporellement d'un motif rythmique à partir d'un tel corpus.
- La prise en compte de la notion de contexte à chaque instant de ce processus génératif.

Nous avons choisi de débuter l'expérimentation en utilisant des boucles rythmiques à pulsation régulière, segmentées en double croche.

Pour commencer cette étude et dans le but de mieux saisir les enjeux de la modélisation temporelle de motifs rythmiques, nous nous sommes attachés à définir les composantes essentielles du rythme. En effet, un rythme et plus précisément une unité rythmique, naît de la succession de deux moments absolument complémentaires : l'*arsis* et la *thesis* — l'élan et la retombée du mouvement. Cette notion est fortement inspirée de la danse, où le corps effectue successivement un mouvement ascensionnel (*arsis*, ou élan du corps) et un mouvement de retombée (*thesis*, ou repos du corps). Par la suite,

l'élan est apparenté à l'anacrouse et la *thesis* est dissociée en deux moments : l'accent, point de culmination de l'énergie, et la désinence — la retombée de l'énergie.
Ainsi, le rythme se définit comme la distribution ordonnée de l'énergie dans le temps et il apparaît chez Democrite comme synonyme de "forme" au sens de "l'ordre" et de la "position". Par la suite, Platon le définira comme *"l'ordre dans le mouvement"*.
Ainsi, ordre, mouvement et contrôle sont alors les points d'ancrage de toute étude visant la création de motifs rythmiques.

La première partie de ce travail dresse une étude bibliographique pour saisir les enjeux de la génération de structures rythmiques et les principes techniques généralement utilisés pour organiser les éléments d'une base de données. Le second chapitre présente les descripteurs sonores utilisés au cours de cette étude. Le troisième chapitre expérimente une méthode de classification supervisée pour obtenir une meilleure organisation de notre espace de navigation composé d'éléments percussifs. Le quatrième chapitre aborde le principe de la synthèse concaténative et apporte une nouvelle approche fondée sur la recherche par similarité de rythme et le cinquième chapitre décrit les aspects techniques pour la mise en oeuvre de cette méthode ainsi que les résultats obtenus. Enfin, le sixième chapitre nous permettra de conclure et de présenter les perspectives d'amélioration de cette approche.

Chapitre 1

Etat de l'Art

" l'émotion produite par le rythme vient de la sucéssion répétée de phases d'attente et de satisfaction.
Wundt".

Les enjeux et le contexte exposés dans l'introduction, nous a conduit vers une recherche bibliographique principalement orientée autour de deux axes :

- les travaux concernant la description des signaux audio de type percussifs et les différentes méthodes employées pour la classification de ces éléments.
- les études traitant de la génération de motifs rythmiques et les approches scientifiques permettant d'intégrer une notion de contexte lors de leur analyse.

1.1 Description et classification d'un espace sonore percussif

1.1.1 Les descripteurs audio

Notre principale motivation consiste à chercher un moyen de décrire au mieux les propriètés acoustiques d'un instrument de type percussif. Une première approche du problème sera de dresser une taxonomie des modèles généralement utilisés qui servira de point de départ à notre étude.
Regroupées sous des thèmes de recherches comme la détection de tempo [Phd][Gou04][F.G05] ou bien l'indexation de genre musicaux, ces publications font en général état de l'utilisation de descripteurs de bas niveau comme le centre de gravité spectral ainsi que les différents moments d'ordres supérieurs, l'énergie du signal, la méthode de l'autocorrélation, la décomposition cepstral[1], etc.
Cette vaste liste de descripteurs constitue le set généralement utilisé dans les travaux faisant appel à un problème de classification.

1.1.2 La classification

Modéliser des séquences rythmiques à partir d'une grande base de données segmentée peut être envisagé comme une tache de classification qui consiste à assigner une catégorie selon la nature de l'instrument (caisse claire, claves, toms, cymbales,...). Cette

[1] MFCC : Mel-Frequency Cepstral Coefficients

classification peut s'effectuer par des algorithmes d'apprentissage exercés sur des bases de données. Ainsi, [P.H01], [E.R] utilise des réseaux de neurones, des machines à vecteurs de support ou l'algorithme des K plus proches voisins.

Pour un exemple plus détaillé on peut citer,[eGR04] qui utilise des GMM[2] et d'autres classificateurs que l'on peut présenter d'une manière générale comme étant, soit un apprentissage supervisé, soit un apprentissage non supervisé.

En ce qui concerne la classification de boucles rythmiques, les publications de Richard et Gillet [GO05], utilisent aussi des modèles de GMM et SVM sur un système complet d'indexation et de recherche par onomatopée au sein d'une large base de données.

L'indexation de boucles rythmiques fait aussi référence aux travaux de recherches effectuées dans le contexte de la bio-informatique [eBP07]. Utilisés dans le cadre de la recherche d'alignement de séquences ADN, il existe des algorithmes, de score local, *Smith et Waterman algorithm*, i.e la meilleure similitude entre deux régions, ou de score global *Needleman-Wunsch-Sellers algorithm*. Ce dernier autorise l'intertion d'un *gap* à l'intérieur de la séquence pour l'obtention du meilleur score final. Testé par Ravelli sur des boucles rythmiques transcrites en séquences symboliques [Rav], ces algorithmes permettent la modification automatique de motifs. L'idée repose ici sur la classification des éléments percussifs d'une base de donnée par GMM ou SVM et ensuite de transcrire les motifs rythmiques sous la forme de séquences de caractères comme le montre les figures 1.1 et 1.2.

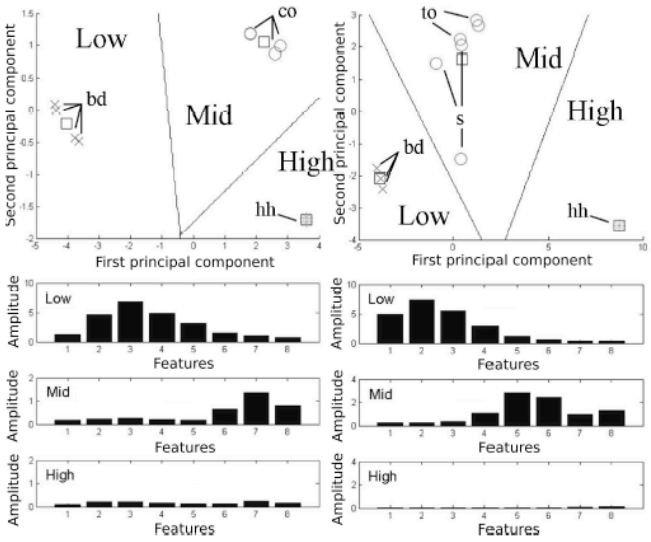

FIG. 1.1 – Classification de la base de son selon trois régions

[2]Gaussian Mixture Model

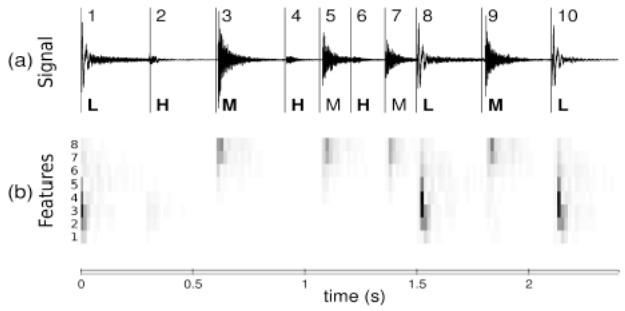

FIG. 1.2 – Transcription de chaque boucle selon une séquence

1.2 La génération de motifs rythmiques

L'utilisation des règles d'harmonie avec l'ordinateur à suscité une très large littérature. Par contre, il existe peu de travaux qui relatent des choix rythmiques. Ces travaux sont généralement appliqués à la batterie et les algorithmes de prédictions travaillent en général sur des séquences symboliques (séquences utilisant le protocole MIDI) plutôt que sur le signal audio en lui même.

Les aticles de J.Paulus [Pau04] [PK03] concernant les N-GRAMs dans la transcription de séquences rythmiques montrent qu'il est possible de construire un "modèle de langage" pour la musique fondée sur la nature répétitive de séquences rythmiques en utilisant des N-Grams suivant une approche probabiliste.
Un N-Gram utilise $N-1$ mots précédents pour prédire quel sera le mot suivant. Cette prédiction peut être formulée comme :

$$P(w_1^k) = \prod_{k=1}^{K} P(w_k | w_{k-N+1}^{k-1})$$

La probabilité des N-grams $P(w_k|w_{k-N+1}^{k-1})$ est estimée en divisant la fréquence d'apparition d'un mot, suivant un certain préfix, par la fréquence d'apparition de ce préfix. Ensuite, Paulus exploite le coté cyclique de chaque boucle rythmique en définissant des N-grams périodiques comme le montre la figure 1.3 où, la prédiction du mot W_K est prise à des intervalles multiples de L qui représente la longueur de la mesure.
Le modèle est ensuite étendu à des ordres supérieurs, bi-grams, tri-grams,...,deca-grams.

L'apparente facilité d'extraction des informations musicales telles que la pulsation et le rythme d'un flux sonore cache l'extrême complexité du système perceptif mis en jeu lors de la réalisation de motifs rythmiques chez l'homme. Cette complexité explique les difficultés rencontrées lors de l'élaboration d'un modèle informatique capable de comprendre et de reproduire un rythme.

D'une manière générale, la génération de motifs rythmique semble plus intuitive que justifiable formellement.

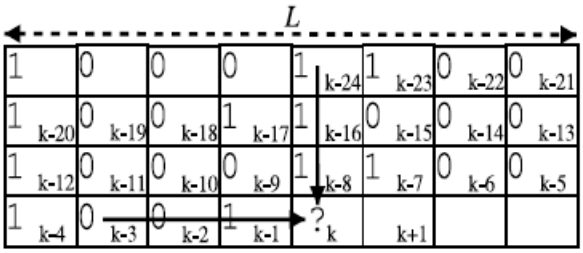

FIG. 1.3 – Périodique N-Grams

1.3 Conclusion et perspectives

Ce parcours bibliographique a permis de déterminer les différentes tâches à suivre pour notre étude.

Au tout début de ce stage, nous avons abordé la problématique en cherchant parmi les travaux relatifs à la synthèse de la parole, dans l'idée que la génération de motifs rythmiques s'apparente à la construction de phrases (une ou plusieurs mesures), constituées de mots (les unités de notre corpus). De plus, la synthèse par concaténation d'unités est une des approches utilisées dans ce domaine. Une approche possible pour la constitution d'un dictionnaires de mots est de classifier les unités les plus similaires. Les classes ainsi formées pourront servir à l'élaboration de probabilités d'apparitions d'un certain élément appartenant à une classe connaissant le contexte précédent.

Notre objectif était d'inscrire cette étude dans le cadre de la synthèse concaténative par corpus, avec l'intention d'utiliser le logiciel CataRT, pour son moteur de synthèse par concaténation. Nous avons donc décidé d'intégrer dans CataRT des descripteurs supplémentaires mieux adaptés aux propriétés acoustiques des signaux que nous utilisons, de type percussif.
Cette étape vise ainsi, l'amélioration des performances d'une navigation dans notre corpus pour obtenir un espace où les différents éléments de nos boucles rythmiques seront mieux organisés i.e en fonction de chaque type d'instrument (caisse claire, tambour, cymbale).

Le chapitre suivant présente les descripteurs ajoutés au logiciel CataRT.

Chapitre 2

Decripteurs sonores

"L'énergie (du grec energeia, force en action) est la résultante synthétique des forces des divers paramètres sonores et figures du discours. Les variations de l'énergie produisent le rythme. Est composante de l'énergie tout paramètre sonore, tout jeu de l'écriture, toute particularité formelle interférant sensiblement sur l'évolution rythmique d'un mouvement."

Ce chapitre présente l'ensemble des descripteurs utilisés au cours de cette étude. Ils sont présentés de façon non-exhaustive et selon les circonstances nous les utilisons sous différents sets.

Au cours de ce stage nous avons cherché des descripteurs sonores mieux adaptés aux signaux acoustiques percussifs.
On note ainsi, que le choix des descripteurs conditionne fortement les résultats finaux, d'où l'importance cette étape.

La version de CataRT de départ effectue une analyse des caractéristiques du signal audio par le bias des descripteurs de bas niveau suivant :

- Centroide spectral
- Pitch (Version de l'objet Gabor Yin)
- Periodicity
- High frequency content
- Mid Frequency content
- Loudness (Version de l'objet Gabor Yin)
- Auto-corrélation d'ordre 1
- l'énergie par trame d'analyse FFT,
- platitude spectrale *(Spectral Flatness)*.

Pour le calcul de ces descripteurs, l'analyse spectrale est calculée sur un signal fenêtré par une fenêtre de hamming et par le biais d'une FFT de 1024 points.

Cette liste correspond à des descripteurs issu de la norme MPEG-7 qui servent actuellement de référence en matière de description musicale et dont les caractéristiques sont présentées dans les travaux de Peeters et Herrera [G.P99][Pee02] [Pee04] [Phi00], [eGP99] .

Un descripteur de bas niveau exprime des propriétés sonores significatives directement observables ou calculables à partir du signal acoustique comme les caractéristiques du spectre du signal, l'énergie de son enveloppe spectrale ou bien son temps d'attaque. Cette approche fournit la base pour la classification de contenu audio par mesure de similarité entre fichiers.

De l'utilisation des descripteurs.... Lorsque l'on y a recours, il est coutume de définir la granularité des descripteurs mis en jeu.

Dans notre cas, nous avons principalement utilisé des descripteurs "locaux" et "semi-locaux". Respectivement, cela implique que les premiers sont calculés à partir de durées qui sont de l'ordre de quelques dixièmes de secondes *(généralement proportionnelle à la taille de l'analyse FFT utilisée)*, et dans le second cas, sur des portions plus importantes du signal.
Nous invitons le lecteur à consulter l'ouvrage de Moreau [Sik05] pour une définition plus approfondie de cette granularité ainsi que les aspects bas-niveau et haut niveau dans la définition d'un descripteur, ainsi que la norme MPEG-7.

Le choix s'est donc porté sur des descripteurs temporels comme le *Zéro Crossing Rate*, et le *le log attack time* ainsi que des descripteurs spectraux comme la *skewness* , la *kurtosis*, *le spectral spread* et une analyse du spectre par bandes fréquentielles selon l'échelle de *Bark*.

On distingue deux classes de descripteurs, ceux calculés sur la partie temporelle du signal et ceux issus de l'analyse fréquentielle du signal.

Pour le domaine temporel

Taux de passages par zéro
C'est le nombre de passage du signal temporel par l'amplitude zéro, rapporté sur le nombre d'échantillons de cette fenêtre, il est plus connu dans la littérature anglophone comme Zero-Crossing Rate (ZCR). Notre algorithme de calcul du ZCR se focalise sur le nombre de changements de signe du signal.

$$Zcr = 0.5 \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|$$

Log Attack Time
Ce descripteur défini le temps que met l'enveloppe du signal pour atteindre son amplitude maximale à partir d'un seuil.*(McAdams,1999)* La formule est :

$$LAT = \log_{10}(T_{stop} - T_{start})$$

Pour le domaine fréquentiel

Le centroïde spectral
Le centroïde spectral donne le centre de gravité de l'énergie du spectre d'un signal en fréquences discrètes, linéaires ou logarithmiques.
Pour un segment audio donné, il est noté μ et défini par :

$$\mu = \frac{\sum_{k=0}^{N_{fft}/2} f(k)P_l(k)}{\sum_{k=0}^{N_{fft}/2} P_l(k)}$$

où, $P_l(k)$ est le spectre de puissance de la n-iéme trame extraite,
$f(k)$ l'indexe des fréquences discrètes.

Il est spécialement désigné pour la distinction de timbres musicaux et la mesure de "brillance d'un son". L'approche visée par l'utilisation de ce descripteur est d'assurer un bonne distinction entre les sons percussifs comme la grosse caisse ou les toms avec les cymbales, cf figure 2.1.

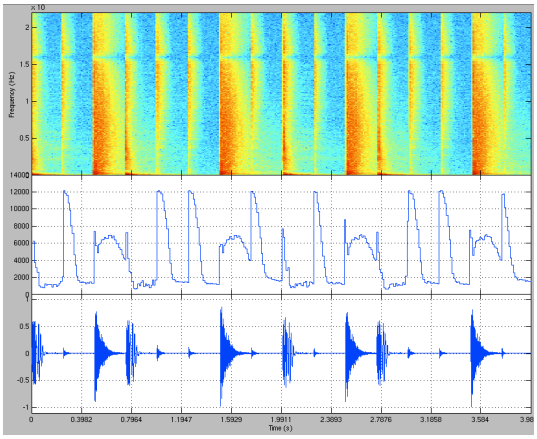

FIG. 2.1 – *Centre de gravité du spectre calculé sur une boucle de batterie*

A partir du centroïde *(moment d'ordre 1)*, nous avons fait le choix de calculer aussi les moments d'ordre supérieurs qui sont respectivement : le spectrum spread, la skewness, la kurtosis.

L'étendue spectrale
Le "spectrum spread" (SS) appelé aussi *"Instantaneous Bandwidth"*. Nous le calculons ainsi :

$$SS^2 = \frac{\sum_{k=0}^{N_{fft}/2} (f(k) - \mu)^2 . P_l(k)}{\sum_{k=0}^{N_{fft}/2} P_l(k)}$$

La skewness
La skewness donne une mesure de la symétrie du spectre autour de sa moyenne. Elle est calculée à partir du moment d'ordre 3. Une skewness nulle correspond à une distribution symétrique, une skewness négative indique une plus grande énergie à droite et une skewness positive une plus grande énergie à gauche.

$$Skewness = \frac{\sum_{k=0}^{N_{fft}/2} (f(k) - \mu)^3 . P_l(k)}{\sum_{k=0}^{N_{fft}/2} P_l(k)}$$

la kurtosis
La kurtosis donne une mesure de l'aplatissement du spectre autour de sa moyenne.

$$Kurtosis = \frac{\sum_{k=0}^{N_{fft}/2} (f(k) - \mu)^4 . P_l(k)}{\sum_{k=0}^{N_{fft}/2} P_l(k)}$$

Fréquence de coupure du spectre : *RollOff*
Fréquence sous laquelle, on trouve la majorité de l'énergie spectrale *(typiquement 95 %)* ; c'est une approximation de la fréquence de coupure entre les parties sinusoïdale et le bruit du signal.*(Tzanetakis and Cook, 2002)*

$$\sum_{k=0}^{k_{roll}} |S(k)| = O.95 \sum_{k=0}^{N_{fft}/2} |S(k)|$$

Echelle de Bark

Pour améliorer la distinction entre les temps accentués, où l'énergie est maximum, et les silences, nous avons choisi de calculer l'énergie en 12 bandes de fréquences en utilisant l'échelle de Bark [E.Z57].

L'échelle de Bark est fondé sur les bandes critiques telles qu'elles sont perçues par l'oreille. La largeur des bandes est donnée par les tableau 2.1 et 2.2 .

	1	2	3	4	5	6
fréquences Hz	20-200	200-400	400-630	630-920	920-1270	1270-1720

TAB. 2.1 – *Bandes de fréquences de 1 à 6*

	7	8	9	10	11	12
fréquences Hz	1720-2320	2320-3150	3150-4400	4400-6400	6400-9500	9500-15500

TAB. 2.2 – *Bandes de fréquences de 7 à 12*

Chapitre 3

Classification par modèle Gaussien

Cette étape vise à mettre en oeuvre une procédure d'identification des éléments percussifs constitutifs d'une boucle rythmique selon une technique de classification s'appuyant sur une modélisation gaussienne de la distribution statistique des données.

3.1 Introduction

Notre première approche était de mieux organiser notre corpus de sons segmentés pour construire des motifs rythmiques par l'assemblage d'unités.

A cela s'ajoute que de nombreux classifieurs ont été testés sur des bases de données comportant des éléments de batterie isolés mais les boucles rythmiques – résultat d'une performance humaine – sont généralement constituées d'événements temporels résultant de plusieurs frappes simultanées sur plusieurs instruments. Par la suite, nous désignons ces instants comme des **éléments mixtes**.

En effet, avec CataRT, la navigation à travers notre espace de descripteurs composé d'éléments percussifs, s'effectue dans un espace à deux dimensions où chaque dimension est un descripteur choisi parmi la liste disponible. En utilisant seulement deux descripteurs nous n'avons pas la possibilité de nous déplacer dans un espace suffisamment bien organisé pour obtenir des régions capables de représenter les différents instruments de nos boucles rythmiques et d'effectuer une bonne distinction entre des unités composées d'un seul instrument et des unités composées d'éléments mixtes.

Pour y parvenir, nous avons opté pour une méthode de classification supervisée, où les données sont dites "complète" car elle contiennent les valeurs $x_1, ..., x_n$ et leur appartenance aux k classes choisies.

Dans la quasi totalité de notre étude et par souci de simplification, nous nous sommes attachés à travailler sur des boucles rythmiques uniquement composées de sons de batterie acoustique. En effet, la grande diversité de sons percussifs électroniques dans les boucles rythmiques issues de courant musicaux comme la techno, l'acid, le rap, aurait rendu

cette étape de classification beaucoup trop laborieuse pour la définition des classes et l'annotation de notre base d'apprentissage.

L'ensemble de cette procédure de classification à été réalisé dans `Matlab` avec l'idée qu'une fois notre corpus classifié nous pourrions importer ce dernier dans le logiciel CataRT et travailler directement sur les classes obtenues pour créer des motifs rythmiques.

3.2 Construction de la base

La première étape à été de constituer une base de données annotées manuellement suivant la taxonomie suivante :

- 4 classes d'éléments de batterie isolés regroupant les grosses caisses *(BD)*, les caisses claires *(SD)*, les cymbales *(CY)* et les charlestons *(HH)*.
- 4 classes d'éléments mixtes représentatifs des ensembles [BD-CY], [BD-HH], [SD-CY], [SD-HH].

Nous avons un total de 8 classes assez représentatives des évènements rythmiques que l'on rencontre généralement dans le rock, le blues et la pop.

Dans son ensemble la base d'apprentissage est constituée de 128 éléments répartis en classe de 16 échantillons.
Ensuite, nous effectuons une étape de standardisation de notre jeu de données.

Standardisation

Pour handicaper le moins possible le systême de classification, il est nécéssaire de normaliser les distributions des descripteurs. Cette étape permet de rendre le modèle moins sensible aux points isolés de notre espace de descripteurs. Le jeu de données est alors standardisé selon la méthode $\mu\sigma$ qui consiste à centrer chaque dimension la base de données et de soustraire la moyenne pour chaque composante. Chaque composante est alors divisée par son écart-type et l'on obtient des dimensions indépendantes de l'échelle de mesure. Dans ce cas, la classification n'est plus perturbée par un descripteur dont la variance serait trop importante.
Par la suite nous considérons que cette transformation est appliquée.

3.3 Le modèle Gaussien

L'étape de classification mis en oeuvre est une méthode d'estimation due à Fisher. Elle permet d'organiser l'ensemble E des objets à classer, en classes homogènes où la classe $C_1, ..., C_k$ est un sous ensemble de E.

Pour définir le cadre mathématique dans lequel chaque nouvel objet est placé, nous avons choisi l'approche probabiliste (approche Bayésienne), où les descripteurs audio issus de l'analyse du signal sont les réalisations d'un vecteur dont la distribution dans $\mathbb{R}^p$ permet d'évaluer les probabilités.
Pour rendre compte de l'information nécessaire à la bonne séparation des différentes classes considérées, nous avons utilisé les vecteurs obtenus par le calcul des descripteurs présentés au cours du chapitre 2.

On se place dans le cas où, chaque vecteur X est constitué de p variables quantitatives et sa distribution suit une loi normale centré μ_k et de matrice de variance covariance σ_k $N(\mu_k, \sigma_k)$.

Notre choix s'est également porté sur l'approche générative, visant à définir la frontière entre les classes, et reposant sur les probabilités *a posteriori* de $P(Y|X)$, où Y représente la variable à valeurs dans $\{1, ..., k\}$, l'ensemble des classes.

La règle de décision choisie associe au vecteur $X \in \mathbb{R}^p$ un vecteur $Y \in \{1, ..., k\}$ qui minimise le risque conditionnel pour chaque observation x par la **règle de Bayes** ou règle du **maximum a posteriori**.

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

où les quantités $P(X|Y)$ et $P(Y)$ ne sont connues que lors de la phase d'apprentissage sur la base de données annotée selon la taxonomie choisie.

La règle consiste à affecter l'observation x à la classe la plus probable *a posteriori* où la classe candidate est celle pour laquelle la nouvelle donnée fournit la densité de probabilité maximale selon :

$$(x) = argmax P(Y = i|X = x)$$

Les paramètres dans le cas du modèle gaussien utilisé sont la moyenne μ_k et la variance σ_k où chacune des classes est modélisée par une densité de probabilité de type gaussien :

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

3.4 Estimation des performances du classifieur

Pour une estimation réaliste du taux d'erreur de notre classifieur, nous nous sommes ramenés à la situation classique de divisés les échantillons d'apprentissage par les échantillons de validation, en ré-echantillonnant le jeu d'apprentissage initial.
La technique utilisée est celle du ***Bootstrap*** qui reprend l'idée de la **validation croisée** mais en introduisant de l'aléa dans la construction du jeu d'apprentissage et de validation.
Elle consiste à faire pour chaque itération $q = 1, ...Q$, la construction d'un jeu de validation par tirage aléatoire de n observations dans le jeu d'apprentissage initial et d'utiliser le reste pour apprendre le classifieur. Une fois le classifieur appris, les n observations du jeu de validation sont classées dans les k classes connues.
In fine, le taux d'erreur est déterminé par le nombre moyen de points mal classés sur les Q itérations de la méthode.

3.5 Resultats

En premier lieu, nous avons expérimenté notre modèle sur une base de données uniquement constituée d'éléments percussifs indépendants i.e les quatre premières classes présentés précédemment. Cela permet d'avoir une idée des performances de ce modèle dans un cas assez classique.
Les résultats sont présentés dans le tableau 3.1 .

D'autre part, pour chaque expérimentation nous avons constitué trois sets de descripteurs :
Pour le set A :

Le centroïde spectral et les 3 moments supérieurs, le ZCR, le RollOff et les 11 premiers coéfficients MFCC.
Pour le set B : Le centroïde spectral et l'énergie dans 12 bandes de fréquences.
Pour le set C : Le centroïde spectral et les 11 premiers coéfficients MFCC.

Set de descripteur	% de classification réussie
Set A	81.66 %
Set B	83.33 %
Set C	79.18 %

TAB. 3.1 – *Résultats pour la classification d'éléments simples selon 4 classes*

A la suite de ces premiers résultats, nous avons appliqué le modèle de classification sur les 8 classes définies. On cherche alors à regrouper des éléments percussifs mixtes en utilisant les mêmes sets de descripteurs que précédemment. Les résultats sont présentés dans le tableau 3.2.

Set de descripteur	% de classification réussie
Set A	59.16 %
Set B	61.66 %
Set C	58.33 %

TAB. 3.2 – *Résultats pour la classification d'éléments mixtes selon 8 classes*

D'une manière générale, le set B fournit les meilleurs résultats. Le set C était expérimental dans le sens où nous voulions comparer le cas avec les bandes de fréquences et les coefficients MFCC. Etant donné que les objets permettant de calculés les coefficients

MFCC ne sont pas disponibles pour Max/Msp, nous conservons par la suite les 12 bandes de fréquences.

3.6 Conclusion

A l'issue de ce travail, nous pouvons affirmer que le modèle est satisfaisant dans le cas d'une classification d'éléments indépendants. Néanmoins, pour la discrimination des sons regroupant plusieurs frappes simultanées, les résultats sont assez décevants.

En effet, il est à noter que la petite taille de notre base d'apprentissage (128 éléments répartis sur 8 classes) est à prendre en compte dans la validité de ces résultats ainsi que les caractéristiques intrinsèques du système de classification utilisé.
En effet, le modèle gaussien adopte l'approche générative et est sensible au nombre d'exemples limités. Une hypothèse d'amélioration serait d'augmenter la quantité de données ou d'utiliser des méthodes moins sensibles à ce problème.

De plus, les modèles probabilistes sont d'autant meilleurs qu'ils s'appuient sur un nombre réduit de variables — c'est à dire le nombre de dimensions ("fléau de dimension"). Une méthode de réduction de dimension comme l'analyse par composante principale est envisageable pour améliorer ces résultats.

Néanmoins, nous avons fait le choix de ne pas continuer sur cette voie. La mise en oeuvre de cette méthode de classification avait pour objet de distinguer les éléments mixtes des éléments indépendants. De plus, l'idée d'effectuer une transcription symbolique de notre corpus sous la forme de classes, implique une perte d'informations supplémentaire par rapport à l'espace continu qu'offre les outils tels que CataRT.

C'est pour toutes ces raisons, que nous avons envisagé une autre approche ; Celle-ci fait l'objet du chapitre suivant.

Chapitre 4

Modélisation par mesure de similarité

" Ce qui fonde l'unité de l'unité rythmique, c'est le fait que les divers sons qui la composent s'agrègent et s'ordonnent continûment de part et d'autre d'un son unique, la note accentuée."
Christian Accaoui.

4.1 Introduction

Les résultats obtenus précédemment nous ont conduit à envisager une autre approche du problème, plus proche de la philosophie de la synthèse concaténative et faisant appel à des méthodes facilement intégrables dans un système temps réel.

Fondée sur l'idée que le corpus représente toujours notre dictionnaire de mots possibles, segmentés à l'état de doubles-croches, et puisque la génération de motifs rythmiques nécessite des structures et des règles pour qu'il soit compréhensible, nous introduisons une séquence cible. Cette séquence est choisie parmi les boucles rythmiques de notre base de sons. Vu comme la "grammaire" d'un langage, ou comme les contraintes logiciels d'un modèle, cette cible permet de guider le choix des unités à synthétiser lors de la création de motifs rythmiques.

Puisque nous cherchons à conserver les caractéristiques d'un style musical défini, nous nous sommes orientés vers les méthodes de recherche par similarité.

4.2 La sélection d'unités dans la synthèse concaténative

Pour cette section, nous décrivons les détails concernant la méthode de sélection des unités utilisé dans un système de synthèse concaténative par corpus. Cette approche fut proposée pour la première fois par [Hunt and Black] et utilisée dans un système de synthèse de la parole comme CHATR et ses dérivés [Sch06],[Die03], [Sch03], [Die00] [Sch04].

Dans le cadre de la synthèse concaténative par corpus, l'algorithme qui réalise la sélection optimum, entre des unités u_i de la base de données composée de N unités et une séquence composée d'unités t_τ s'éffectue par le calcul de deux fonctions de distance :

- distance de la séquence cible correspondant à la similarité des unités de la base u_i avec les unités de la cible. Celle-ci est donnée comme la somme des distances, entre le calcul des descripteurs de la base, pondérée par un poids.

$$C^t(u_i, t_\tau) = \sum_{k=1}^{p} w_k^t C_k^t(u_i, t_\tau)$$

- distance de concaténation, ou de continuité, prédisant la discontinuité introduite par la concaténation de l'unité u_i avec le candidat précédent u_{i-1}, par la recherche du meilleur chemin à travers la base de données.

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{p} w_k^c C_k^c(u_{i-1}, u_i)$$

Le choix de l'unité à synthétiser minimise le coût de ces deux fonctions de distance.

Dans le cas particulier de CataRT, la distance de concaténation $C^c(u_{i-1}, u_i)$ n'est pas réalisée lors de la re-synthèse et la jonction entre deux unités consécutives est assurée par le biais d'un crossfade.

Le principe de la mesure de distances est une méthode de recherche par similarités, ou mesure de dissemblance. Parmi les travaux sur la modélisation de la parole, ce principe est utilisé dans le cadre de la ***quantification vectorielle***[1] [tDJHHL00]. Cette approche permet de substituer, un vecteur X_k à k composantes, par un vecteur voisin Y_i appartenant à un ensemble fini i de M vecteurs défini comme prototypes. Cette étape de substitution provoque une distorsion souvent appelée distance, où la distance la plus utilisée est la distance euclidienne *(erreur quadratique)*.

Ici, nous ne cherchons plus à classifier les unités comme pour la quantification vectorielle, nous effectuons une recherche par similarité entre une unité de notre séquence cible avec toutes les unités de notre corpus à un instant donné du temps sans se soucier d'une quelconque appartenance à une classe.
Pour y parvenir, notre choix de métrique est la distance de *Minkowski*[2] prise à l'ordre 2, c'est à dire la distance euclidienne.

Par sa simplicité et les résultats qu'elle fournit, elle se positionne comme le candidat idéal pour des systèmes temps réel.

4.3 Appariement avec la cible

L'appariement concerne la recherche de séquences, communes, ou les plus similaires possibles, entre notre corpus et le motif rythmique de référence, la *cible*, par le biais d'une mesure de distance.
Ici, la cible, dans sa représentation discrète, détermine le monde "idéal", la matrice finale,

[1] Les méthodes de référence pour la quantification vectorielle en traitement de la parole sont : La méthode de Lloyd *(K-means)*, l'algorithme d'apprentissage à seuil et l'algorithme par éclatement binaire

[2] La distance de Minkowski d'ordre 1 est la distance de Manhattan, ou *(City-block)*
La distance de Minkowski d'ordre 2 est la distance euclidienne.

vers laquelle nous devons évoluer.
Mais un rythme ne peut se dire *"vivant"* que s'il intègre de subtiles variations au sein même d'une structure répétitive [Acc01].

Pour intégrer ces variations, on utilise une mesure de distance euclidienne pondérée dans un espace E composé d'une matrice U, représentative de l'ensemble des $k - units$ du corpus, d'une matrice T représentative de l'ensemble des $n - units$ de la cible et d'une matrice de pondération W aux dimensions de l'espace euclidien que nous utilisons. Cet espace est déterminé par le nombre de descripteurs utilisés pour l'analyse de nos boucles rythmiques.

$$U_{corpus} = \begin{bmatrix} u_{01} & u_{02} & ... & u_{0d} \\ u_{11} & u_{12} & ... & u_{1d} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ u_{k1} & . & ... & u_{kd} \end{bmatrix} \times \begin{bmatrix} W_1 & W_2 & ... & W_d \end{bmatrix}$$

où $d \in \{1, 2, ...D\}$ représente la dimension donnée par les descripteurs et $k \in \{1, 2, ..., K\}$ les unités constituant le corpus.

$$T_{target} = \begin{bmatrix} t_{01} & t_{02} & ... & t_{0d} \\ . & . & ... & . \\ . & . & ... & . \\ t_{n1} & . & ... & t_{nd} \end{bmatrix}$$

Ainsi, l'appariement "simple", que nous appelons dans la suite de ce document, ***la distance n0***, ou contexte nul, s'obtient par le calcul de la fonction de distance entre chaque unité du corpus et chaque unité de la cible à un instant donné.

$$dist^2(u_k, t_n) = \sum_{d=1}^{D} W_d |u_k^d - t_n^d|^2$$

Ensuite, la synthèse par concaténation réalise la recherche de l'index de l'unité qui minimise cette distance.

$$Indexunit_{n0} = min \begin{bmatrix} C_0 = dist^2(u_0, t_{n0}) \\ : \\ : \\ C_k = dist^2(u_k, t_{n0}) \end{bmatrix}$$

où la matrice $C(k)$ représente ici la somme des distances entre chaque unité du corpus à l'instant t_{n0}, la première unité de la cible.

Cette première approche nous permet de générer des motifs rythmiques proches de notre cible de départ mais non parfaitement identiques puisque :

- Les unités issues de l'analyse de la cible ne sont pas dans le corpus,
- La sélection de l'unité est donnée par une mesure de distance pondérée par un poids sur les descripteurs du corpus.

Lors d'une utilisation en temps réel, la matrice W_d sera un des paramètres de contrôle.

4.4 Modélisation avec contexte

Pour ajouter une notion de contexte dans le processus de génération de motifs, nous rajoutons une contrainte supplémentaire dans le choix de l'unité à synthétiser.
Cette approche utilise toujours une mesure de distance et nous définissons dans la suite de ce document la notion de contexte comme étant ***la distance de continuation*** — notée C^x pour contexte.

Ce choix terminologique est destiné à éviter toute confusion avec la distance de continuité C^c présentée et commentée dans la section 4.2.

4.4.1 La distance de continuation

La distance de continuation envisagée est fondé sur l'appariement de l'unité à synthétiser avec les unités de la cible en cours, c'est à dire à un instant donné, comme nous l'avons vu précédemment, mais à laquelle nous avons rajouté, pour trouver l'index de l'unité à synthétiser, la mesure de distance des unités de la cible aux instants précédent, $(n-1), (n-2), ..., (n-N)$. N étant le nombre d'unités de la cible.

La matrice de n unités variables, ainsi obtenue, définit notre *matrice cible avec contexte.*

Avec cette approche, nous cherchons la fonction qui minimise la distance pondérée par un poids selon :

$$C^x(k) = \sum_{n=0}^{N} W_n dist^2(t_n, u_{k-n})$$

où la matrice W_n avec $n \in \{1, 2, ..., N\}$ permet d'accorder un poids plus ou moins important entre la distance de l'unité cible $n0$ (vu précédemment) et la contribution de ses n unités précédentes.

Pour que le calcul de distance soit effectué sur toutes les composantes de U, les unités de la matrice cible avec contexte effectuent une rotation cyclique à travers tout le corpus. La recherche de l'index à synthétiser est alors :

$$Idx_unit = min \begin{bmatrix} C_1^x = dist^2(u_1, t_{n0} + t_{n-1} + ... + t_{n-N}) \\ \vdots \\ \vdots \\ \vdots \\ C_k^x = dist^2(u_k, t_{n0} + t_{n-1} + ... + t_{n-N}) \end{bmatrix}$$

4.5 Fonctionnements de la génération de motifs

Cette approche par mesure de similarité permet de mettre en évidence trois modes de fonctionnement lors du processus de génération de motifs rythmiques. Ils sont respectivement :

- L'approche par similarité de rythme
- La distance de continuation
- Le mode auto-contexte

Nous présentons ici, d'un point de vue fonctionnel le processus de sélection des unités dans ces trois cas. Ces derniers utilisent les modes de synthèse *fence* et *beat* de CataRT.

Avec le ***mode fence*** : l'unité sélectionnée est synthétisé en entier et sans répétition.

Avec le ***mode beat*** : l'unité sélectionnée est répétée entièrement à un tempo défini par l'utilisateur.

Ultérieurement, nous verrons plus en détails les paramètres permettant de naviguer à travers ces trois modes.

4.5.1 L'approche par similarité de rythme

Ce mode de fonctionnement implique :

- Seule la pondération W_d des descripteurs est prise en compte dans la modélisation.
- Le contexte des unités de la cible est nul.
- La rotation de la cible à travers tout le corpus pour le calcul des distances.
- Mode de synthèse : fence.

Nous verrons dans la partie concernant les résultats que ce mode permet de conserver les temps accentués, *(downbeat)*, contenus dans la cible tout en changeant la "couleur" des éléments percussifs ou en rajoutant, entre les temps accentués, de subtiles variations rythmiques en fonction du caractère spécifique de chaque descripteur.
La figure 4.1 présente un schéma de principe résumant le processus de sélection d'unités dans ce mode.

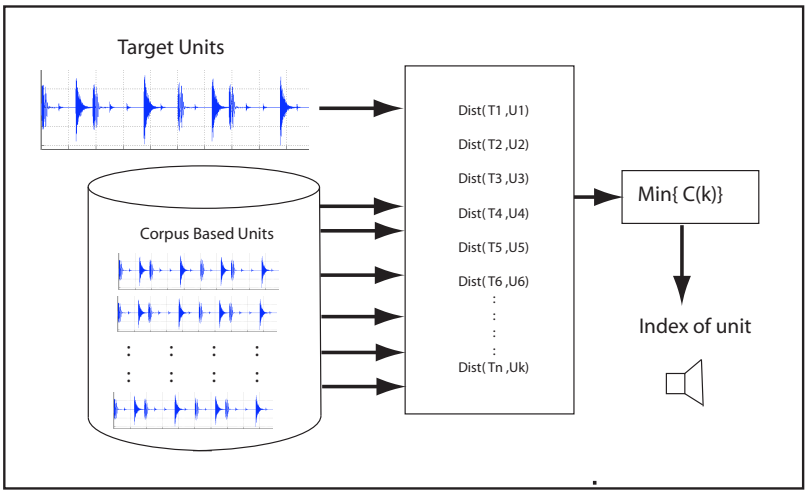

FIG. 4.1 – Principe du model 1

4.5.2 La distance de continuation

Ce mode de fonctionnement implique :

- Une prise en considération des unitées précedentes de la cible pour chaque calcul de distance, de telle sorte que la somme des distances entre $(BassDrum + Hi - Hat + Hi - Hat) \neq (Snare + Hi - Hat + BassDrum)$,

- La rotation de la cible avec un contexte à travers tout le corpus pour le calcul de distance.
- Mode de synthèse : fence.

On obtient dans ce mode, soit la répétition d'une région particulière des boucles de la base.

4.5.3 Le mode auto-contexte

Ce mode est un peu particulier par rapport aux précedents. En effet, la matrice cible n'est plus une boucle rythmique donnée. Aussi introduisons-nous le terme de ***matrice de contexte***. Elle est explicitement fondée sur la notion *d'évolution* au travers des différents choix successifs obtenus par le processus de sélection.

Ce mode nécessite, soit une interaction humaine, soit le mode radius de CataRT.

Le Radius

Le radius rajoute un aléa dans le choix de l'unité. Ainsi, l'index de l'unité à synthétiser n'est pas systématiquement la distance minimale entre les unités cibles de la matrice de contexte. De plus, la sélection est réalisée dans une région particulière — celle définie par le radius — et non plus sur l'ensemble des dimensions du corpus.

La figure 4.2 présente l'utilisation du radius représenté par le cercle.

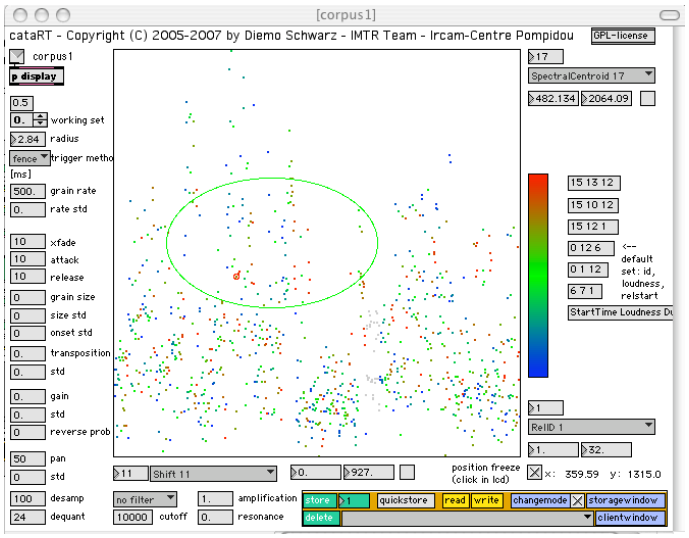

FIG. 4.2 – Ecran contrôle de CataRT permettant de naviguer dans le corpus

Ce dernier mode de fonctionnement implique :

– Le mode de synthèse : beat qui permet de séquencer la sélection à un tempo précis.
– Une interaction humaine ou l'intervention du mode radius.

La figure 4.3 présente un schéma de principe résumant le processus de sélection des unités dans ce cas.

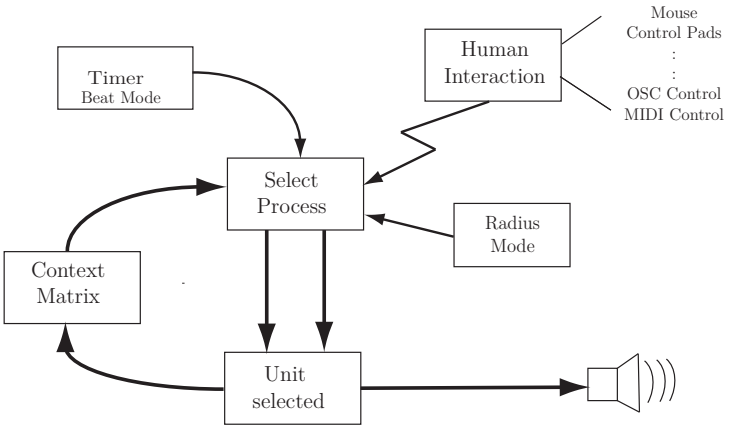

FIG. 4.3 – Principe du mode auto-context

Chapitre 5

Mise en oeuvre

Cette partie décrit chronologiquement l'ensemble des étapes nécessaires à l'élaboration du modèle final.

5.1 Réalisation

Pour l'élaboration de notre corpus de test *(segmentation, calculs des descripteurs)*, nous avons utilisé des séquences de batteries issues du commerce. La base de données pour les tests est composée de 44 boucles rythmiques, soit un dictionnaire de 1408 unités. Les boucles rythmiques sont de longueurs équivalentes et contiennent deux mesures segmentées à la double croche.

Pour garantir une analyse assez fine des motifs rythmiques générés, nous avons choisi un ensemble de boucles ayant le même style musical — en l'occurrence pop/rock — et composés uniquement de sons de batterie acoustique.

Pour faciliter l'analyse des résultats, le rythme de notre cible est simple et typique du style musical de la base *(Caisse clair sur le 2ième et 4ième temps, Grosse Caisse 1er et 3ième temps)*.

5.1.1 Implémentation dans Matlab

La première implémentation est effectuée dans l'environnement de `Matlab`. Elle permet d'obtenir des résultats rapides et d'utiliser un set assez large[1] de descripteurs audio.

Lors des essais, nous avons utilisé les sets de descripteurs présentés au chapitre 3 mais nous avons utilisé aussi ces descripteurs de façon isolée pour analyser leurs influences respectives.

Une fois l'appariement effectuée entre le corpus et la cible, la synthèse audio est réalisée dans `Max/MSP` en utilisant le protocole OSC[2] entre `Matlab` et CataRT.

[1] En effet, certain descripteurs ne sont pas encore disponibles dans Max/Msp.
[2] librairie matlab-OSC liblo 0.2. http ://opensoundcontrol.org

Les premiers résultats montrent que la structure du rythme de la cible — caractéristique du style — est assez bien conservée dans le cas d'une mesure de distance simple ie la distance n0.
De plus, l'écoute des fichiers sons obtenus en jouant sur le nombre de descripteurs utilisés, figure , démontre que l'énergie par bandes de fréquences permet de conserver le *downbeat* alors que le centre de gravité du spectre, pris comme seul descripteur pour l'appariement, s'avère insuffisant.

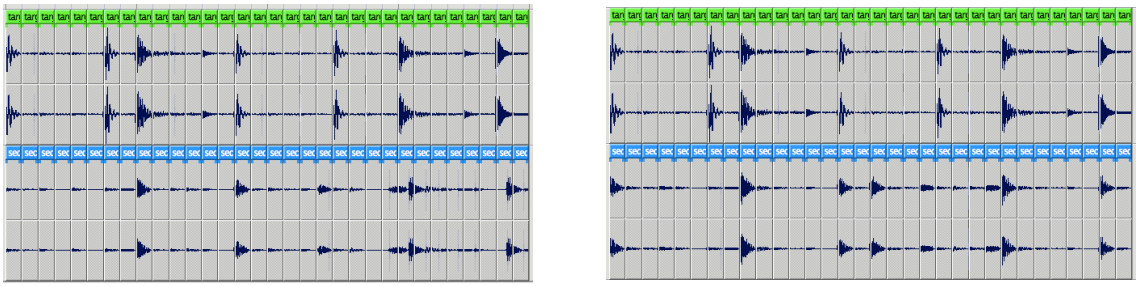

FIG. 5.1 – *Appariement des unités en utilisant : le centroide (à gauche) et l'énergie par bandes de fréquences (à droite). La cible est en haut et le motif reconstitué se trouve en bas.*

Dans le cas de la figure de droite, les sons de caisse claire et de grosse caisse sont exécutés aux mêmes instants que notre cible. Ce résultat garantit la proximité avec le style musical pop/rock.

Pour évaluer l'influence des poids sur les descripteurs et du poids sur le contexte dans un environnement temps réel, nous avons travaillé directement dans CataRT.

Néanmoins, cette étape a permis de conclure que :
- Le choix de la métrique est bonne,
- seule l'action cumulée de plusieurs descripteurs permet d'obtenir une large palette de variations rythmiques,
- les coefficients MFCC n'apportent pas d'informations supplémentaires par rapport aux bandes de fréquences.

5.1.2 Implémentation dans CataRT

L'implémentation de la distance de continuité dans CataRT a permis d'incrémenter la version de l'objet **MnM.Mahalanobis** disponible dans la librairie MnM[3] [NR05][ND05] [FBS05]. Son code source est réalisé en C^{++} [Gri04].
L'objet est doté de méthodes supplémentaires accessibles par des messages FTM. Le message *set-context-size* permet de définir le nombre d'unités de la cible pris en compte dans le calcul de distance et le message *set-context-weight* correspond au poids que l'on accorde au contexte.

[3] Music is not Mapping

La figure regroupe les fonctions mises en oeuvres dans cataRT pour obtenir la modélisation finale :
Une figure présentée en annexe donne un aperçu de l'espace de création de l'interface CataRT.

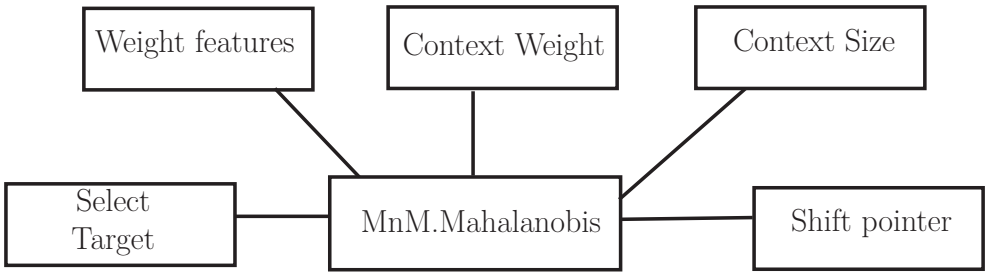

FIG. 5.2 – default

Nous détaillons ici le rôle de chaque module :

Le module *Weight features*

Comme l'objectif est d'obtenir un maximum de contrôle sur la matière synthétisée, nous avons étendu le calcul de la distance de Mahalanobis sur l'ensemble des descripteurs disponibles dans CataRT. En effet, précédemment, le choix de la distance optimum était réalisée dans un espace à deux dimensions, le descripteur projeté sur l'axe horizontal de l'écran de contrôle, ainsi que celui projeté sur l'axe vertical. Dorénavant, l'utilisateur peut sélectionner un descripteur en particulier ou un set complet parmi la liste disponible et par conséquent modifier le poids qu'il souhaite lui accorder individuellement.

Le module *Select Target*

Ce module permet de sélectionner une des boucles rythmiques parmi la base de donnée et de lui donner le rôle de cible lors de l'appariement des unités. Les unités qui composent la cible sont alors inaccessibles par le processus de sélection en leurs attribuant une distance de l'ordre de $1e10^{12}$.
De plus, ce module permet la rotation cyclique des unités de la cible à travers l'ensemble du corpus pour le calcul des distances.
Note : Un métronome est disponible pour vérifier la pertinence des motifs générés.

Calcul du *Shift*

Intégré au système CataRT parmi la liste des descripteurs standards, il est utilisé lors de l'appariement des distances avec contexte. Son rôle est d'indiquer à l'objet *MnM.Mahalanobis* la position de l'unité précédente.
Chaque fois que l'appariement de l'unité de la cible en cours avec le corpus est réalisé, il permute d'une unité.

Il est calculé dynamiquement à partir de la dernière boucle rythmique analysé.

Context Weight - Context size

Permet de spécifier à l'objet *MnM.Mahalanobis* la taille du contexte et le poids de ce dernier

5.1.3 L'algorithme pour la distance de continuation

Requis :
variable $d \in \{1, 2, ..., D\}$ relative aux dimension de l'espace.
variable $n \in \{1, 2, ..., N\}$représentative des n unités de la cible, matrice $T_{n,d}$.
variable $k \in \{1, 2, ..., K\}$ représentative des k unités du corpus, matrice $U_{k,d}$.
variable t : taille du contexte.
matrice de distance $D_{k,n}$.
vecteur W_d composé de D éléments.
vecteur W_t composé de N éléments.
vecteur S pour le shift initialisé avec $s \in \{N-1, 0, 1, 2, ..., N-2\}$.

Initialisation :
For each corpus rows $k = 0; k < K; k++$
    For n target unit $n = 0; n < t; n++$
        For d features $= 0; d < D; d++$
            $D(n,d) = +dist^2(U_{k,d}, T_{n,d}) \times W_d$
        end
        $C(k) += D_{n,S[s]} \times W_t$
        permute S de 1 position
    end
end
Index= $min\{C(k)\}$

5.2 Interprétation des résultats finaux

Cette partie dresse une analyse des différentes séquences, ou motifs, obtenus lors de navigation en temps réel en utilisant les différents modes de sélections d'unités présentés précédemment : appariement simple sans contexte, avec contexte et en mode automatique.

5.2.1 Bilan sur les descripteurs

Les premiers résultats obtenus dans `Matlab` avec un modèle simplifié permettent de conclure qu'une partie du style musical d'origine de notre cible est conservé par le recours simultané de quelques descripteurs judicieusement choisis.
En effet, l'énergie du signal par bandes de fréquences cumulée avec le centroïde spectral ou le ZCR permettent d'obtenir le matériel de base pour que la séquence générée soit assimilable à un rythme.

La figure montre que les instants de notre cible ou l'énergie est maximale se retrouvent dans notre motif aux même instants.

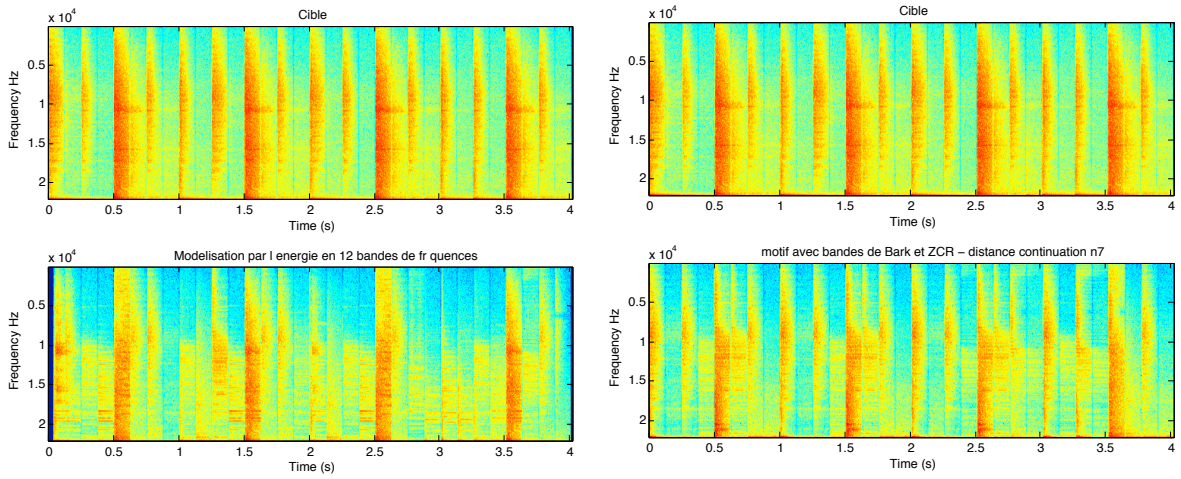

FIG. 5.3 – *Appariement des unités en utilisant l'énergie par bandes de fréquences : sans continuation (à gauche) et avec le descripteur ZCR (à droite) plus une continuation de 7 unités. La cible est en haut et le motif généré est en bas.*

L'intégration dans CataRT, permet de jouer sur les différents poids de notre modèle en temps réel.
Par conséquent, on modifie le processus de sélection de l'unité à synthétiser dans le temps. C'est par le biais de ces modifications que nous obtenons des variations rythmiques qui s'intègrent dans la structure de base.
D'un point de vue perceptif, ces variations sont le résultat de deux processus différents : La **permutation** et la **fusion** d'éléments percussifs :

La permutation a pour effet de modifier la "couleur" des différents éléments de la batterie en remplaçant une caisse claire avec un timbre par une caisse claire sans timbre ou bien une grosse caisse par des toms, etc...

La fusion, quand à elle, résulte de l'ajouts d'éléments non présents dans le motif cible, notre matériel original. On notera que cette action de fusion permet d'introduire dans certains cas un effet de ***"groove"*** qui n'était pas contenu dans la cible.

La combinaison gagnante des descripteurs parmis ceux utilisés est difficile à établir puisque chacun d'eux permet des variations plus ou moins acceptables. Dans cet esprit, nous dirons qu'il en va du goût de chacun.
Néanmoins, parmi la liste de descripteurs disponibles dans CataRT, le spectralslope, la kurtosis, la skewness et l'auto-corrélation d'ordre 1 ne jouent pas un rôle prépondérant pour la création de variations intéressantes.

5.2.2 Bilan sur la continuation

En ce qui concerne la distance de continuation, selon la taille du contexte et le poids choisi, on obtiens un nouveau motif rythmique construit par la répétition d'une région des boucles de la base. La figure montre un exemple de répétition obtenue avec une distance de continuation égale à 7 ce qui correspond à deux noires.

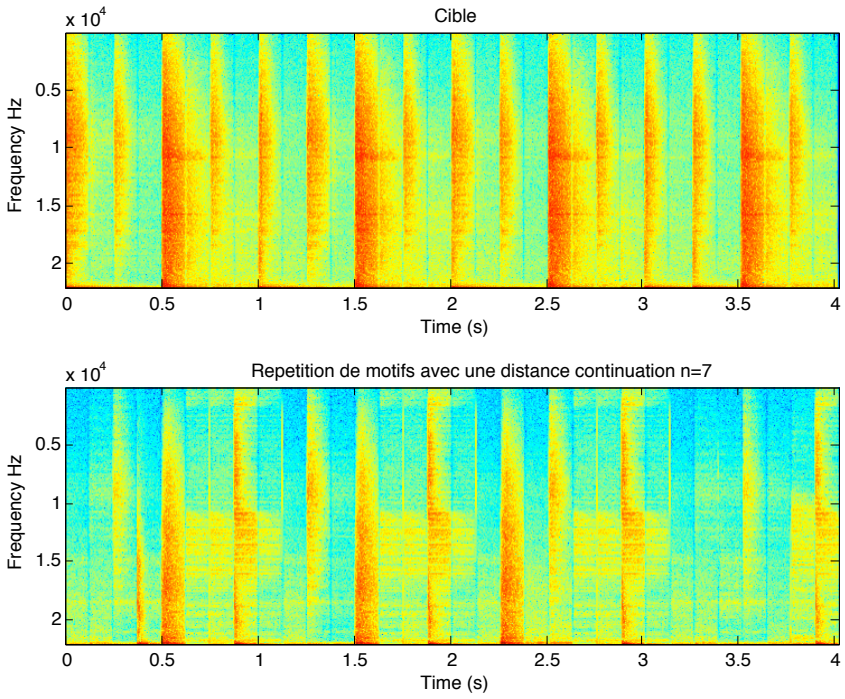

FIG. 5.4 – default

Ainsi, l'action de ces deux modes de sélection, assurent la fonction de *"motif-riff"*[4] terme souvent employé en musique lorsqu'il s'agit de répéter et de transformer un pattern.

[4]le motif-riff : est un terme venant du Jazz. initialement, il faisait référence à des motifs mélodique mais par extension son emploi s'est étendu au rythme

Néanmoins, l'écoute attentive des différents motifs générés permet de souligner les points suivants :

- Dans certain cas, les silences sont remplacés par des attaques provenant de cymbales,
- Le manque de précision dans la continuation des unités sélectionnées.

En effet, lors de l'écoute des résultats, les "supposés silences"[5] remplacés par une attaque de cymbale, donne une sensation de discontinuité dans la perception du rythme. Nous avons testé le modèle avec une boucle rythmique extrêmement simple et ne comportant que des cymbales courtes ie peu de sustain.

La figure (à gauche) montre que les silences sont détectés correctement.

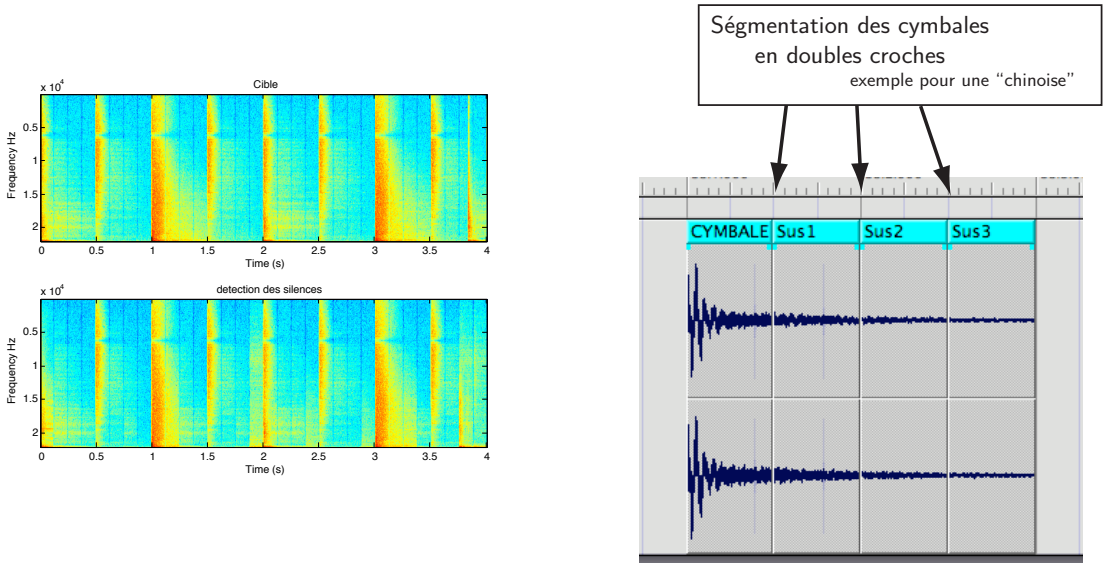

FIG. 5.5 – *Le spectrogramme à gauche montre clairement que le modèle retrouve correctement les zones de silences. A droite, exemple de segmentation sur une cymbale chinoise, cymbale qui contient généralement beaucoup de sustain.*

Dans un enregistrement continus, le sustain des cymbales assure l'homogénéité du son.

La figure 5.5 de droite montre l'effet de la segmentation à la double croche sur une cymbale ("chinoise") ayant beaucoup de sustain. La troncature de la cymbale durant sa phase décroissante donne 3 unités sonores *(sus1, sus2, sus3)*. Ces unités pourrons aboutir au "collage" d'autres unités de cymbales qui comportent l'attaque du signal d'où un effet de discontinuité.

Autrement, la distance de continuation permet d'obtenir des motifs rythmiques formés par le collage de 2 ou 3 boucles contenues dans le corpus (3 noires issues de la boucle A + 2 noires de la boucle B + 1 mesure de la boucle C). On améliore ainsi l'homogénéité de l'ensemble mais ce saut brutale entre de grandes portions de boucles rythmiques donne parfois la sensation de passer d'un morceau de musique à un autre sans y avoir été préalablement préparé. Ce cas n'exclus pas que les résultats soient inintéressants.

[5] le terme "supposé" sous entend qu'à cet instant précis, la présence d'un vrai silence ajouterait de la clarté dans la perception du rythme que l'on a.

Chapitre 6

Conclusions et perspectives

Nous avons présenté dans la première partie de ce rapport une méthode de classification supervisée utilisant le modèle gaussien sur un corpus d'unités sonores composé de boucles rythmiques de batterie segmentées à la double croche.

Notre objectif étant la génération de motifs rythmiques à partir d'un tel corpus, nous avons pensé qu'une classification de ces unités en classes d'éléments mixtes serait un bon point de départ. L'intention de cette démarche était de suivre une approche probabiliste fondée sur la fréquence d'apparition de ces classes. Mais les résultats obtenus avec notre modèle de classification n'étaient pas satisfaisants. En effet, la présence d'éléments regroupant plusieurs instruments entraînait des résultats erronés.

Aussi dans la seconde partie, avons-nous opté pour une autre méthode de recherche, par similarité de rythme. Cette nouvelle approche nous permettait d'utiliser les propriétés de la synthèse concaténative par corpus. L'utilisation d'une mesure de distance nous a permis de réaliser l'appariement, avec ou sans contexte, entre les unités du corpus et les unités d'une séquence cible, permettant par là même de créer des motifs rythmiques nouveaux.
Les résultats obtenus montrent que notre modèle est capable de restituer les composantes essentielles du rythme : répétition cyclique d'une structure, fusion d'événements supplémentaires avec la structure originale. De plus, cette approche laisse une place à la création de nouveaux rythmes en temps réel.

Néanmoins, afin d'enrichir les éléments du discours rythmique, nous pourrions envisager l'implémentation de paramêtres de contrôle supplémentaires : pondération indépendante sur les différentes unités du contexte, ajout de descripteurs. D'autre part, notre problématique de départ reposait sur une segmentation à la double croche. Ainsi, notre modèle n'envisage pas le cas où la base serait composé de rythmes ternaires et binaires. Une étude dans cette direction permettrait d'améliorer l'ensemble du système en ajoutant un mode de segmentation utilisant une détection d'attaque.

Pour obtenir des effets supplémentaires, une amélioration intéressante serait d'envisager le cas d'une translation, régulière ou irrégulière, du motif rythmique dans la mesure, à l'image du processus graduel de répétition utilisé dans la musique minimaliste.

D'autre part, l'utilisation de cette application permet de construire des corpus utilisant des boucles rythmiques avec des pulsations différentes alors que les bases de données composées de séquences rythmiques possèdent généralement un nombre limité d'exemples pour chaque tempo. Lorsque l'on souhaite alors obtenir un motif rythmique particulier au tempo recherché, il est nécessaire d'effectuer des transformations sur les fichiers audio.

Cette étape rajoute un temps supplémentaire à l'élaboration d'un projet. Dans notre cas, la seule contrainte est de prendre des fichiers sons ayant le même nombre de mesures pour assurer une segmentation correcte.

Enfin, l'élaboration de ce modèle au sein de CataRT trouve un domaine d'applications assez large puisqu'il permet de créer des séquences de sons en utilisant d'autres types de corpus.

Bibliographie

[Acc01] Christian Accaoui. Le temps musical. Desclée de Brouwer col. Philosophie, 2001.

[Cur88] Roads Curtis. Introduction to granular synthesis. Computer Music Journal, 1988.

[Die00] Schwarz Diemo. A system for data-driven concatenative sound synthesis. Digital Audio Effects (DAFx). Verona p. 97-102, Décembre 2000.

[Die03] Schwarz Diemo. New developments in data-driven concatenative sound synthesis. International Computer Music Conference (ICMC) p. 443-446., Singapore, Octobre 2003.

[eBP07] Grégory Nuel et Bernard Prum. Analyse statistique des séquences biologiques : modélisation markovienne, alignements et motifs. ed. Hermès, Paris, Mars 2007.

[eGP99] Perfecto Herrera Xavier Serra et Geoffroy Peeters. Audio descriptors and descriptors schemes in the context of mpeg-7. In Proceedings of the International Computer Music Conference (ICMC99), Bejing, China„ 1999.

[eGR04] O. Gillet et G. Richard. Automatic transcription of drum loops. In International Conference on Acoustics, Speech, and Signal Processing ICASSP04, Montréal, Québec„ May 2004.

[E.R] B.Sandler E.Ravelli. Drum sound analysis for the manipulation of rhythm in drum loops. Centre for Digital Music Queen Mary University of London.

[E.Z57] E.ZWICKER. Critical band width in loudness surmnation, February 1957.

[FBS05] Rémy Muller Frédéric Bevilacqua and Norbert Schnell. Mnm : a max/msp mapping toolbox. Proceedings of the 2005 International Conference on New Interfaces for Musical Expression (NIME05), Vancouver, BC, Canada, 2005.

[F.G05] S.Dixon F.Gouyon, A.Klapuri. An experimental comparison of audio tempo induction algorithms, 2005.

[GO05] G.Richard and O.gillet. Indexing and querying drum loop databases. GET Telecom Paris, 2005.

[Gou04] Fabien Gouyon. A review of automatic rhythm description systems, 18 Oct. 2004.

[G.P99] X.Rodet G.Peeters. Automatically selecting signal descriptors for sound classification, 1999.

[Gri04] T. Grill. A c++ programming layer for cross-platform developments of pd and max/msp externals., 2004.

[Joh] Oswald John. Plunderphonics. Webpage http ://www.plunderphonics.com.

[ND05] N.Schnell and D.Schwarz. Gabor, multi-représentation real-time analysis/synthésis. in Proc. Int. Conf. on Digital Audio Effetcts (DAFx-05), Madrid Spain, pp 122-126., 2005.

[NR05] R.Borghesi D.Schwarz F.Bevilacqua N.Schnell and R.Muller. Ftm – complex data structures for max. in Proc. Int. Comp. Music Conf. (ICMC'05), Barcelona, Spain, 2005, pp. 9–12, 2005.

[Pau04] Jouni Paulus. Acoustic modelling of drum soundswith hidden markov models for music transcription, 2004.

[Pee02] Geoffroy Peeters. Instrument sound description in the context of mpeg-7. International Computer Music Conference (ICMC), Berlin, Germany., 2002.

[Pee04] Geoffroy Peeters. A large set of audio features for sound description in the cuidado project. Ircam, 2004.

[P.H01] F.Gouyon P.Herrera, A.Yeterian. Automatic classification of drum sounds : a comparison of feature selection methods and classification techniques. Universitat Pompeu Fabra, 2001.

[Phd] Fabien Gouyon Phd. A computational approach to rhythm description.

[Phi00] P. Philippe. Low-level musical descriptors for mpeg7. Signal Processing Image Communication 16 :181–191 2000., 2000.

[Pie] Schaeffer Pierre. Traité des objets musicaux. Editions du Seuil. Paris. France.

[PK03] J. Paulus and A. Klapuri. Conventional and periodic ngrams in the transcription of drum sequences, 2003.

[Rav] Emmanuel Ravelli. Automatic rhythm modification of drum loops.

[Sch03] Schwarz. The caterpillar system for data-driven concatenative sound synthesis. Digital Audio Effects (DAFx) p. 135-140. London, Septembre 2003.

[Sch04] Schwarz. Data-driven concatenative sound synthesis, thèse de doctorat, université paris 6 - pierre et marie curie, paris, 23. 1. 2004.

[Sch06] Schwarz. Real-time corpus-based concatenative synthesis with catart. In Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx), Montreal, Canada, Septembre 2006.

[Sik05] Hyoung-Gook Kim Nicolas Moreau Thomas Sikora. Mpeg-7 audio and beyond, audio content indexing and retrieval. Wiley press, 2005.

[tDJHHL00] René Boite Hervé Boulard thierry Dutoit Joel Hancq Henri Leich. Traitement de la parole. Presses polytechniques et universitaires rommandes, 2000.