EXTRACTION AUTOMATIQUE DE DESCRIPTEURS RYTHMIQUES DANS DES EXTRAITS DE MUSIQUES POPULAIRES POLYPHONIQUES

PAR

FABIEN GOUYON

Sony CSL Paris

Rapport présenté pour l'obtention du DEA ATIAM : Diplôme d'Etudes Approfondies en Acoustique, Traitement du signal et Informatique Appliqués à la Musique

Université de la Méditerranée, Université Paris VI, IRCAM, Télécom-Paris, Université du Maine, Ecole Normale Supérieure, ACROE-IMAG

Juillet 2000

INFORMATIONS GENERALES

Ce rapport donne un aperçu du travail de recherche que j'ai effectué entre Mars et Juillet 2000 à :

Sony CSL Paris
6, rue Amyot
75005 Paris
http://www.csl.sony.fr/

Ce travail correspond au stage pratique du Diplôme d'Etudes Approfondies en Acoustique, Traitement du signal et Informatique Appliqués à la Musique (ATIAM), enseignement conjoint à l'Université de la Méditerranée, l'Université Paris VI, l'IRCAM, Télécom-Paris, l'Université du Maine, l'Ecole Normale Supérieure et l'ACROE-IMAG.

Mon responsable de stage est M. FRANÇOIS PACHET.

Le travail effectué lors de ce stage consiste en la participation active aux recherches relatives au rythme poursuivies par l'équipe "Music" de Sony CSL, notamment :

- L'amélioration d'un algorithme d'extraction d'événements rythmiques par corrélation.
- L'étude théorique concernant l'utilisation de la corrélation dans cet algorithme.
- L'implémentation d'une méthode d'extraction de paramètres de signaux percussifs dans un objectif de classification (ce qui a donné lieu à la soumission d'un article à la conférence DAFX 2000 qui doit avoir lieu en Décembre à Verone – voir Annexe V.3).
- Des expérimentations relatives à l' Analyse/Synthèse de signaux percussifs par la modélisation de Prony (en suite à un précédent travail : cf. [GOUYON]).

Pour informations supplémentaires ou commentaires :

Adresse : 243, rue St-Jacques 75005 Paris
E-mail : fgouyon@csl.sony.fr
URL : http://www-ccrma.stanford.edu/~fgouyon

Résumé

EXTRACTION AUTOMATIQUE DE DESCRIPTEURS RYTHMIQUES
DANS DES EXTRAITS DE MUSIQUES POPULAIRES POLYPHONIQUES

Dans le contexte de MPEG7, nous nous intéressons au problème de l'extraction automatique de descripteurs des signaux audios, dans le but d'alimenter des applications fondées sur les contenus sémantiques des signaux (caractéristiques de *haut niveau*) plutôt que sur leurs contenus physiques (caractéristiques de *bas niveau*). Les descripteurs de *haut niveau* d'un signal audio sont nombreux : types d'instruments, type de voix, etc. Ceux étudiés ici sont les descripteurs rythmiques. Dans les musiques populaires, objets de nos études, le rythme est en effet une caractéristique prédominante, rentrant pour une grande part dans la définition des styles musicaux.

Nous abordons les problèmes de détection, de segmentation, de classification et de synthèse des instruments percussifs dans les signaux musicaux polyphoniques. Nous détaillons des analyses effectuées sur de nombreux sons de synthétiseurs et signaux réels directement extraits de morceaux musicaux. De nombreuses techniques ont été mises en oeuvre : détection par corrélation, calcul d'enveloppes, segmentation et caractérisation de régions de signal, modélisation ARMA, modélisation de Prony, analyse factorielle discriminante, analyse multidimensionnelle et classification hiérarchique ascendante. Les recherches menées résultent en un algorithme adaptatif et récursif qui raffine graduellement la définition d'un son en parallèle de la localisation de ses instants d'apparition dans le signal musical. Les informations rythmiques d'un extrait musical se trouvent donc concentrées dans plusieurs séries temporelles d'indices d'occurrences des sons percussifs importants dans la perception du rythme. Une expérience utilisant une base de 206 titres est détaillée.
Nous introduisons une représentation générique des rythmes d'extraits musicaux, propice aux comparaisons, et fondée sur l'analyse de ces séries temporelles. Diverses applications fondées sur cette représentation sont présentées.

TABLE DES MATIÈRES

REMERCIEMENTS

Tout d'abord, je remercie sincèrement mon responsable de stage, FRANÇOIS PACHET, de m'avoir permis de travailler sous sa direction. De même, les membres de Sony CSL se sont toujours montrés généreux en conseils avisés.

Bien évidemment, tout ce travail n'aurait pas été envisageable sans les apports des intervenants du DEA ATIAM : enseignants et compositeurs. Je tiens à exprimer en particulier ma gratitude à Messieurs RISSET, KRONLAND-MARTINET, ARFIB, DEPALLE, MCADAMS, STROPPA, BABONI, MALHERBE, WARUSFEL, FABRE et ASSAYAG.

Je tiens également à exprimer toute ma sympathie à mes camarades de DEA, avec qui il a été à la fois intriguant et extrêmement enrichissant de partager l'expérience de l'intégration à l'IRCAM ; tout particulièrement EMILIA GOMEZ, JEREMY MAROZEAU et THIBAUT EHRETTE.

Finalement, je remercie ma famille (M., mes parents, ZIZOU & DANIEL, VALERIE & PF et mes grands-parents) pour m'avoir toujours apporté un soutien matériel plus que généreux, mais surtout une oreille attentive, et un encouragement constant.

I. INTRODUCTION

Ce chapitre propose une introduction du sujet étudié, des définitions préliminaires ainsi que notre positionnement par rapport aux recherches existantes sur le thème de l'extraction de rythmes.

I.1. PRECISIONS SUR LE SUJET

Une mélodie est bien plus qu'une simple succession de notes, il existe des relations précises qui structurent cette suite de hauteurs. De la même manière, l'appréhension d'un rythme va au-delà de la simple perception des durées d'évènements sonores successifs. Percevoir un rythme, c'est regrouper différents sons en des motifs structurés.
Un morceau de musique est composé de divers évènements sonores aux propriétés physiques différentes qui s'enchevêtrent et/ou se succèdent. Les notions de rythme, de tempo et de métrique résultent du fait que certains de ces éléments sonores sont accentués par notre analyse perceptive – marqués comme étant des éléments particuliers.
Alors même que le concept d'accentuation – ou de marquage – d'un évènement sonore est central dans l'analyse du rythme, aucune explication de référence en terme de causes psychologiques n'existe. On ne peut pas, en l'état actuel de nos connaissances, énoncer sans équivoque ce qui détermine l'importance d'un évènement sonore dans la perception du rythme. Des facteurs tels que les hauteurs, les intensités, les timbres, les harmonies et les durées jouent évidemment un rôle dans l'impression d'accentuation : la répétition dans le temps d'un de ces facteurs conduit certainement à l'accentuation des évènements sonores qu'il caractérise, puis au repérage de leurs divers instants d'occurrences, au phénomène de regroupement et finalement à la perception du motif rythmique. Cependant, il est difficile de certifier qu'un facteur particulier possède une suprématie sur les autres.

Il semble raisonnable de penser que l'association des processus perceptifs et cognitifs de *marquage* et d'*organisation en structures* des évènements sonores nous permette de définir mentalement des schémas rythmiques.
La méthode d'extraction du rythme développée dans ce rapport adopte une méthodologie similaire. Tout d'abord nous basons nos analyses sur le signal lui-même, par opposition à une étude d'une représentation abstraite de ce signal, comme la partition par exemple. Mais surtout nous distribuons les tâches d'analyses sur deux modules :

- Un premier module ayant comme objectif de générer automatiquement à partir du signal musical des listes d'évènements, sous forme de séries d'indices temporels. Chaque série correspond à un élément sonore supposé important dans la perception du rythme de ce morceau. Ceci fait l'objet du paragraphe II.
- Le deuxième module consiste – par l'utilisation de ces séries – en la génération d'informations représentatives du rythme du morceau et facilement manipulable (par exemple lors de comparaisons), comme par exemple son tempo, ou son positionnement dans un "espace des rythmes" multidimensionnel. Ceci fait l'objet du paragraphe III.

Dans un soucis d'applicabilité des algorithmes mis en oeuvre, nous focalisons notre intérêt sur les musiques populaires[1], dans lesquelles le rythme est une dimension prédominante.

[1] Précisons que nous entendons par "*musiques populaires*" les musiques qui bénéficient de moyens de diffusion importants, qui sont au centre du commerce de la musique, et qui, par conséquent, correspondent plus trivialement aux musiques que les gens chantent dans la rue (ou sous la douche) de nos jours en occident.

I.2. TERMINOLOGIE

Afin de clarifier les sens que nous donnons à des termes tels que "rythme", "accentuation", "tempo", etc., nous proposons de dédier ce paragraphe aux définitions de termes importants dans les études sur le rythme. Ces définitions sont inspirées de l'ouvrage [COOP/MEY].

Pulsation[2] : Une pulsation est une occurrence d'un stimulus extrait d'une série de stimuli tous identiques. Comme les "tic" d'un métronome, les pulsations définissent des unités dans le continuum temporel.

Accentuation : Marque apposée à certaines pulsations par notre système perceptif ; l'accentuation peut être dynamique, mélodique, harmonique, ou bien due à un facteur de durée, ou encore de timbre. Une pulsation accentuée constituera un point de repère dans la construction mentale des aspects rythmiques de la musique.
Nous l'avons déjà abordé dans l'introduction, la perception d'un rythme peut exister lorsque certaines des pulsations sont accentuées (les *fortes*), alors que d'autres ne le sont pas (les *faibles*).

On définit deux modes d'organisations, la mesure et le rythme :

Métrique (ou *mesure*) : C'est la mesure du nombre de pulsations faibles entre des pulsations fortes récurrentes. Dans les partitions, elle est donnée par la signature et matérialisée par les lignes de mesure.
Par son numérateur, la signature nous indique le motif général d'accentuation rythmique. Le dénominateur indique l'unité temporelle (i.e. les pulsations).
Par exemple, une signature 4/4 indique que les pulsations sont comptées au niveau des noires (dénominateur=4). Elle indique de plus que quatre pulsations forment une mesure: le motif rythmique de base consiste en un regroupement de quatre pulsations (numérateur=4), une accentuée et trois non-accentuées (comme dans "*un*-deux-trois-quatre-*un*-deux-trois-quatre").
Une signature 6/8 indique que les pulsations sont comptées au niveau des croches (dénominateur=8). De plus, six pulsations forment une mesure : le motif rythmique de base tient dans le regroupement de six pulsations (numérateur=6), une accentuée et cinq non-accentuées (comme dans "*un*-deux-trois-quatre-cinq-six-*un*-deux-trois-quatre...").

Rythme : Selon [COOP/MEY], c'est la manière dont une ou plusieurs pulsations faibles sont regroupées en relation avec une pulsation forte ; les auteurs font ici le parallèle avec les cinq motifs basiques de la prosodie (*iamb*, *anapest*, *trochee*, *dactyl* et *amphibrach*). Sans abonder complètement dans le sens d'un tel parallèle qui peut paraître rapidement inopportun, on peut cependant en extraire une indication sur la différence entre rythme et mesure : la mesure prescrit un schéma général d'accentuation des pulsations, le rythme décrit plusieurs schémas particuliers, qui en général s'affranchissent des lignes de mesure. Il est donc un peu réducteur de parler d'*un* rythme pour un morceau, il faudrait plutôt envisager

[2] Les pulsations sont appelées "pulses" ou "beats" dans la littérature anglophone.

plusieurs motifs rythmiques, on parlera du rythme d'un groupe de note, le groupe pouvant être plus ou moins grand (cf. page 10), ou encore du rythme d'un extrait d'un morceau.

Finalement, la dernière définition est relative à ce qui est confondu chez la plupart des gens avec le rythme : le tempo.

Tempo : Directement lié à la mesure, c'est le nombre de pulsations par minute. On spécifie souvent l'unité temporelle correspondant à une pulsation (e.g. "*60 à la noire*" signifie que la pulsation est comptée au niveau des noires, et qu'elles sont au nombre de soixante dans une minute).

I.3. ETAT DE L'ART

ETUDES MUSICOLOGIQUES

Comme nous l'avons introduit plus haut, percevoir un rythme, c'est regrouper différents sons en des motifs structurés.
De la même manière que les lettres se combinent en mots, les mots en phrases, les phrases en paragraphes, etc. ; en musique, les notes sont regroupées en motifs, les motifs en phrases, etc. Les analyses musicales utilisant la notion de structuration à des échelles différentes sont classiques dans les domaines de l'harmonie et du contrepoint. On parle de différents *niveaux architectoniques* de composition. Les analyses rythmiques des musicologues se basent également le plus souvent sur ces notions (voir [COOP/MEY]).
De petits motifs rythmiques sont donc perçus comme ayant leurs propres formes et structures, mais aussi comme participant à une organisation rythmique de plus large échelle.
Faire une analyse du rythme à un certain niveau architectonique, c'est définir un type d'évènements sonores (note ou groupe de notes) et un critère d'accentuation d'un évènement par rapport à un autre (durée, intensité, timbre, etc.). Les définitions des *niveaux rythmiques primaire*, *inférieurs* et *supérieurs* sont les piliers du cadre théorique permettant l'introduction rigoureuse des notions de mesure, de pulsation et de rythme citées en page 8.

ETUDES SYMBOLIQUES

Une grande partie des recherches sur le rythme se basent sur les définitions des musicologues et se fixent comme objectif leurs généralisations à des analyses automatiques[3] de musiques de styles différents. La détermination à partir du signal audio des listes d'évènements rythmiques importants étant en soi un problème difficile[4], ces courants de recherche opèrent sur des données symboliques plutôt que sur des signaux audios. Certains focalisent sur l'étude de signaux MIDI (e.g. [ALLEN/DAN.]), d'autres sur l'études de partitions (e.g. [BROWN1]). Ces travaux consistent le plus souvent en la détermination du rythme de phrases musicales jouées par un seul instrument.

A PARTIR D'ANALYSES DE SIGNAUX

Très éloignés des analyses théoriques, les travaux concernant l'extraction automatique du tempo ou de la mesure (cf. [SCHEIRER]) sont en général peu soucieuses des définitions et des recherches des musicologues. Les analyses ancrées dans la réalité des signaux audios correspondent en général uniquement à des motivations de suivi de tempo en temps-réel (cf. [SCHEIRER]). Elles répondent au nom générique de "*beat-tracking*" ou "*foot-tapping*". Les méthodes employées impliquent souvent le filtrage du signal, et la recherche de maximums (i.e. "*onsets*") supposés correspondre aux instants d'attaques des sons percussifs (cf. [GOTO/MUR.1]).

---

[3] i.e. utilisant l'ordinateur

[4] Il faut en effet définir le type d'évènement sonore important, le phénomène d'accentuation des pulsations, et implémenter des algorithmes automatiques qui leurs soient relatifs

I.4. APPROCHE SUIVIE

Un certain nombre de travaux se basent sur l'hypothèse que la *durée* d'un évènement sonore est le facteur principal définissant l'accentuation rythmique de celui-ci (e.g. [BROWN1]). D'autres se basent sur l'hypothèse que les accentuations nous permettant d'appréhender le rythme sont des accentuations *harmoniques* (e.g. [GOTO/MUR.2]).
En admettant que la focalisation sur *un seul* facteur soit pertinente, la direction que nous choisissons d'emprunter ici est celle de la reconnaissance de *timbres*. Nous pensons que cela nous permettra d'extraire avec réussite des *séries temporelles* d'évènements sonores cohérents et, de plus, que ces éléments correspondent à ceux auxquels l'oreille humaine est naturellement la plus sensible dans sa perception du rythme.

L'originalité de notre approche tient dans le fait que nous tentons d'extraire des listes d'indices rythmiques à partir du signal audio lui-même, c'est à dire de l'enchevêtrement des différents sons constitutifs de l'extrait musical. Nous essayons également de rapprocher notre travail d'un cadre plus théorique, en prenant en compte les définitions des notions introduites page 8 :

Ici, l'*accentuation* que nous attribuons au facteur *timbre* correspond à notre capacité à percevoir un instrument percussif possédant un timbre particulier et à le "détacher" d'un flot sonore, et la notion de *pulsation* (unité de division du continuum temporel) est donnée par les occurrences des sons percussifs importants dans la perception du rythme.
Dans ce cadre, on peut redéfinir la notion de *métrique* à partir d'une liste d'indices temporels – ce qui revient à faire une analyse du même type que celle de BROWN (cf. [BROWN1]), mais plus ancrée dans la réalité du signal audio musical, et donc plus proche des applications de grandes échelles (e.g. sur des base de données importantes).
On s'intéresse également à des informations contenues dans les relations temporelles qu'entretiennent les différents instruments percussifs *entre eux*. Les instruments respectant tous la même mesure, une redondance est vraisemblablement contenue dans ces informations, la détermination de la *métrique* doit en être confortée. D'un autre côté, ces informations "inter-instrumentales" nous permettent d'avoir accès à une notion de *rythme* en accord avec la définition de la page 8 : Chaque instrument suit un schéma rythmique particulier – en se conformant à une mesure commune à tous –, c'est l'association de différents schémas qui définit un rythme. La représentation exhaustive des informations temporelles "inter-instrumentales" doit donc être caractéristique des motifs rythmiques du morceau.
Le *tempo* est déterminé par la connaissance de la pulsation et de la métrique.

Notre ambition est que notre méthode, basée sur l'étude des corrélations de séries temporelles – elles-mêmes issues d'analyses de signaux audios musicaux –, nous permette de déterminer la *pulsation* d'un morceau musical, son *tempo* et sa *métrique* (les trois étant liés comme nous l'avons vu en page 8), mais qu'elle nous permette aussi d'accéder plus généralement à une représentation du *rythme* de l'extrait considéré (i.e. des niveaux rythmiques supérieurs et inférieurs). Le paragraphe III.1.1 donne le détail de ce que notre méthode est actuellement capable d'effectuer.

I.5. AVANT DE COMMENCER...

L'assertion théorique principale qui est faite ici est qu'il est pertinent de focaliser les analyses sur *un seul* facteur d'accentuation des évènements sonores, de plus nous choisissons d'emprunter l'hypothèse de la suprématie des caractéristiques de *timbres* sur les autres (durées, harmonies, hauteurs et intensités).
Faire une telle assertion entraîne évidemment une perte de généralité. Nous pouvons d'ores et déjà prévoir que certaines musiques ne répondront pas à un tel critère. Cela étant, il nous semble qu'à ce point de la réflexion, il est justifiable de prendre position et d'effectuer des tests en pratique. La mise en place d'expérimentations et l'étude de leurs résultats conduira certainement à des conclusions quant à la validité des hypothèses de départ.
De plus, les musiques sur lesquelles nous effectuons nos tests correspondent à des styles particuliers. Nous n'envisageons pas d'atteindre un degré de généralité similaire à celui des études musicologiques sur le rythme.

Nous tenons à prévenir le lecteur que la majeure partie de notre travail a porté sur l'extraction automatique d'évènements rythmiques, et l'amélioration d'un algorithme développé en Janvier-Février 2000 à Sony CSL. Cet algorithme consistait en :

1. Effectuer la corrélation entre un signal musical et un son de référence.
2. Utiliser le critère de qualité #1 explicité en page 21.
3. Générer un nouveau son de référence à partir des indices d'occurrences.
4. Itérer

Des améliorations ont été apportées aux diverses étapes de cet algorithme, elles sont développées tout au long de ce rapport. L'état actuel de l'algorithme est explicité en page 44.

Nous ne proposons au chapitre III que le début d'une recherche qui devrait résulter en une méthode de représentation complète du rythme d'un morceau, mais qui pour l'instant ne nous permet de déterminer efficacement que le tempo d'un extrait musical (voir exemples au paragraphe III.1.2).

II. L'EXTRACTION AUTOMATIQUE D'EVENEMENTS RYTHMIQUES

Il s'agit de "résumer" au mieux le signal à son information rythmique : pour chaque évènement sonore important dans la perception du rythme, on cherche une série temporelle d'indices d'occurrences.

Posons le problème : nous disposons d'un extrait musical polyphonique donné ; plus ou moins d'instruments sont présents selon le morceau, mais nous supposons que certains instruments participent plus que d'autres à notre appréhension du rythme de ce morceau. L'hypothèse est faite que ce sont les timbres percussifs et récurrents qui nous font percevoir l'aspect rythmique. L'objectif est donc de déterminer quels sont les sons présents dans le signal qui correspondent à une telle définition, et également de déterminer les indices temporels auxquels ils sont présents.

A cette fin, nous proposons un algorithme itératif d'extraction d'évènements rythmiques, centré sur la reconnaissance des timbres, et qui précise la définition des timbres recherchés en même temps qu'il améliore la détection de leurs occurrences.

II.1. EN BREF

On fait tout d'abord l'hypothèse que les timbres percussifs et récurrents sont les évènements sonores rentrant en compte dans notre appréhension du rythme.
L'algorithme d'extraction des évènements percussifs récurrents présents dans un signal donné est itératif. Il est basé sur l'échange d'informations entre deux modules distincts :

- Un module de Détection
- Un module de Synthèse

Plus précisément :
On définit un son de référence général (voir le paragraphe II.6.1.a pour une discussion sur le choix du son de référence initial).
On détecte dans le signal les évènements sonores proches de ce son de référence (la détection est effectuée par un algorithme basé sur le calcul de fonctions de corrélations, cet algorithme est explicité au paragraphe II.2).
Ensuite, un tri est effectué parmi ces évènements sonores afin de déterminer avec précision lesquels correspondent à un instrument percussif particulier, le plus proche du son de référence, et lesquels correspondent à d'autres instruments, plus éloignés (voir paragraphe II.4).
On synthétise (voir paragraphe II.5) un son percussif qui devient le nouveau son de référence à rechercher dans le signal.
C'est après cette phase de synthèse que l'itération peut s'effectuer.

Nous focalisons notre recherche sur deux sons de références uniquement. Ceci dans un soucis de simplicité et afin de pouvoir obtenir rapidement des résultats, mais également car il semble vraisemblable qu'en ce qui concerne la musique populaire, il y a peu souvent plus de deux séries temporelles différentes et d'importance similaires qui participent à l'élaboration du rythme.

Résumons par un schéma :

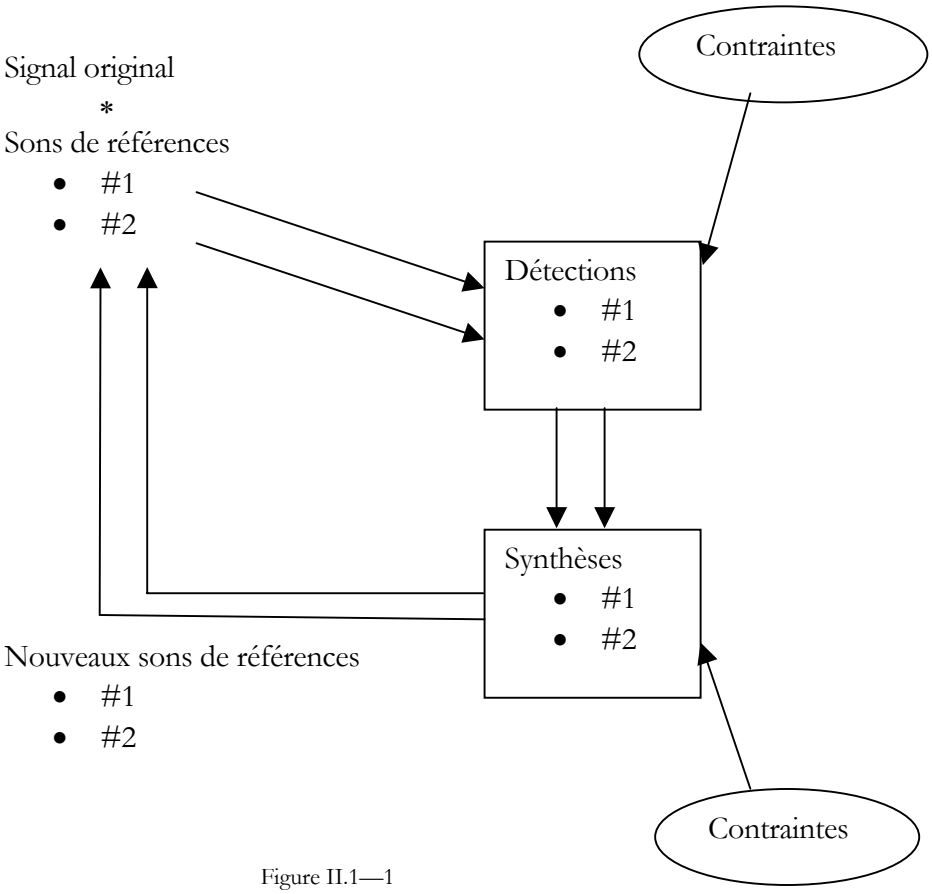


Figure II.1—1

Les contraintes appliquées sur la phase de détection sont pour une part laissées à notre choix (on veut les définir afin de séparer les bons indices des mauvais), d'autres nous sont imposées (comme nous le verrons au paragraphe II.2.1.d page 23).
Au niveau de la synthèse, on choisit d'appliquer des contraintes nous garantissant que les deux sons seront effectivement différents, en nous garantissant ainsi du problème de recoupement des séries temporelles (voir page 43).

II.2. DETECTION D'EVENEMENTS PERCUSSIFS RECURRENTS

L'idée principale est d'utiliser la corrélation entre l'extrait musical et un son de référence afin de détecter les occurrences d'évènements percussifs importants dans cet extrait.
Nous décrivons en premier lieu une suite d'expérimentations effectuées, visant à justifier l'utilisation de la corrélation dans un algorithme de détection, nous proposons également une méthode permettant de palier aux faiblesses de la détection par simple corrélation.
Ensuite, nous nous penchons sur l'intérêt qu'il existe à utiliser un algorithme itératif.

II.2.1. DE L'UTILISATION DE LA CORRELATION

La fonction de corrélation entre le signal x(n) et le son de référence y(n) est donnée par l'équation suivante :

$$R_{xy}(n)=\sum_{i=1}^{N} x(i)\times y(i-n)$$

Équation II.2—1

La fonction de corrélation à un instant n donne le degré de proximité entre un signal x(i) et un deuxième signal y(i) décalé de n échantillons. L'autocorrélation d'un signal à son maximum pour n égal à zéro. Intuitivement, il apparaît donc justifiable d'utiliser la fonction de corrélation pour détecter si un signal y(n) est présent dans un signal x(n).
Dans le cas réel – qui nous occupera finalement –, nous ne possédons pas a priori un modèle des évènements rythmiques de l'extrait étudié. De plus, l'évènement rythmique dont nous essayons de déterminer toutes les occurrences n'est évidemment pas exactement le même à chacune de ces occurrences. Le fait qu'il puisse par exemple être d'amplitude forte ou faible montre bien que les propriétés physiques du signal aux occurrences des évènements qualifiés comme similaires par notre perception sont différentes.

II.2.1.a. EXPERIENCES

Pour tester la validité de l'utilisation de la fonction de corrélation, nous avons effectué la progression d'expériences suivante :

1. Corrélation d'une séquence de son monophoniques et non-bruités issus d'un synthétiseur (banque General MIDI du Korg05RW) avec les sons de cette séquence
2. Corrélation d'une séquence, transformée par ajout d'effets, de son de synthétiseur monophoniques et non-bruités avec les sons de la séquence non-transformée
3. Corrélation d'une séquence de sons réels (donc bruités par un environnement polyphonique) avec ces mêmes sons
4. Corrélation d'une séquence de sons réels transformée (donc bruités par un environnement polyphonique) avec les sons non-transformés

EXPERIENCE 1

On génère une séquence de 44 sons avec des sons percussifs monophoniques et non-bruités.

NB : Les sons 1 à 5 correspondent à des grosses caisses, les sons 6 à 13 sont des caisses claires, 14 à 18 sont des charlestons, il y a ensuite des toms, des cymbales, des cloches et diverses percussions.

On calcule la valeur absolue de la corrélation de cette séquence avec chacun des sons la composant, et on mesure les hauteurs des pics de corrélation.
On note $c_k(l)=R_{xy_k}(n_l)$, $c_k(l)$ est la hauteur de la fonction de corrélation entre les signal $x(n)$ et le signal $y_k(n)$, à l'indice temporel correspondant au son numéro l.
NB : n est un indice temporel ($\in[1,N]$) ; et k, l, i et j sont des indices de numéro de sons ($\in[1,44]$).
On définit une mesure normalisée (entre 0 et 1) de proximité entre sons à partir des hauteurs des pics de corrélation, cette mesure est la suivante :

$$d_k(l)=\frac{c_k(l)}{\sqrt{c_k(k)}\times\sqrt{c_l(l)}}$$

Équation II.2—2

Elle a été construite afin de vérifier les propriétés suivantes :

- $d_k(k)>d_k(j) \quad \forall j,k\in[1,44]$
- $d_k(j)=d_j(k) \quad \forall j,k\in[1,44]$
- $d_k(k)>d_i(j)\geq 0 \quad \forall k,(i,j \ tel \ \text{que} \ i\neq j)\in[1,44]$
- $d_k(k)=d_j(j)=1 \quad \forall j,k\in[1,44]$

On dispose donc d'une mesure nous permettant de comparer les proximités des sons entre eux.
Les mesures peuvent se résumer en une matrice carrée symétrique, se prêtant bien à la représentation, comme le montre le graphique suivant.

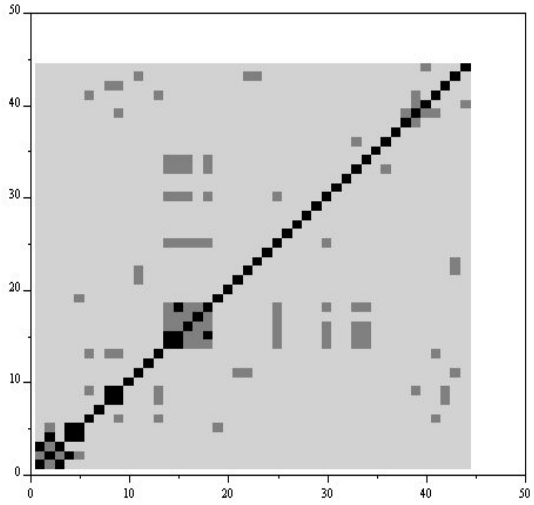


Figure II.2—1 : Mesure de proximités entre sons de la banque GM du Korg05RW

Pour visualiser cette notion de proximité, nous fixons deux seuils en amplitudes (e.g. 0.8 et 0.35), et attribuons une couleur pour les pics dépassant chaque seuil : les carrés noirs

correspondent à une proximité $d_k(j)$ entre les sons d'indices k et j supérieure à 0.8, ceux gris foncés à une proximité comprise entre 0.35 et 0.8, et enfin, les gris clairs à une proximité comprise entre 0 et 0.35.
En observant la diagonale sur ce graphique, on se rend bien compte que la corrélation entre un signal et un son de référence nous permet de retrouver ce son dans le signal. De plus, il semble que, en adaptant judicieusement le nombre et les hauteurs des seuils, apparaisse un regroupement des sons par famille de percussions (e.g. les grosses caisses correspondent à des mesures toutes comprises entre 0.8 et 1).

EXPERIENCE 2

La même séquence est transformée violemment par ajout de distorsion et de réverbération.
La même mesure de proximité entre sons est effectuée, à partir de la valeur absolue de la corrélation de la séquence transformée et des sons originaux.
Une représentation graphique du résultat est la suivante.

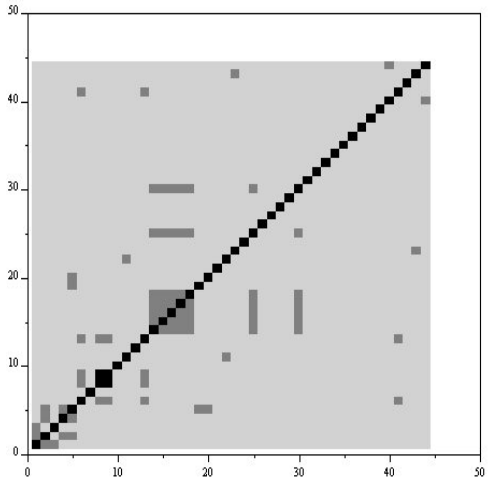


Figure II.2—2 : Mesure de proximités entre sons de la banque GM du Korg05RW originaux et transformés

Le phénomène de regroupement apparaît un peu moins que dans le cas où aucune transformation n'est appliquée, mais la détection du son de référence est toujours effectuée.

EXPERIENCE 3

On s'intéresse maintenant aux sons réels, i.e. extraits directement de morceaux de musique. Les sons de percussions peuvent toujours être considérés comme étant monophoniques, mais ils sont plongés dans un environnement polyphonique qui, de notre point de vue, peut être considéré comme du bruit.
Dans l'exemple suivant, il s'agit de la mise en séquence de 8 occurrences de caisse claire et de 10 occurrences de grosse caisse directement extraites du morceau "High times".

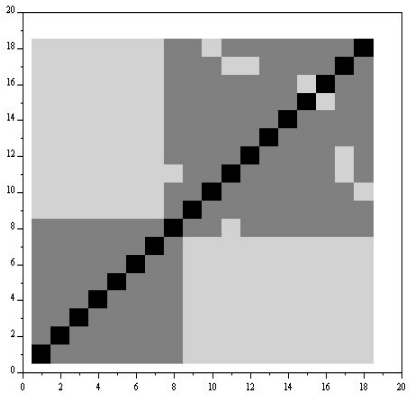


Figure II.2—3 : Mesure de proximités entre diverses occurrences de grosse caisse et caisse claire dans "High times"

On s'aperçoit que la détection de chaque son est effective – le contraire serait inquiétant car les sons de la séquence et ceux de références sont exactement les mêmes. De plus, le phénomène de regroupement en famille de son est là encore présent : les grosses caisses correspondent toutes à une proximité élevée, c'est la même chose pour les caisses claires.

EXPERIENCE 4

On s'intéresse toujours aux sons réels.
Dans l'exemple suivant, il s'agit de la mise en séquence de 2 occurrences de caisse claire et de 3 occurrences de grosse caisse directement extraites du morceau "By and Bye".
NB : les grosses caisses correspondent aux indices 2, 3 et 4.

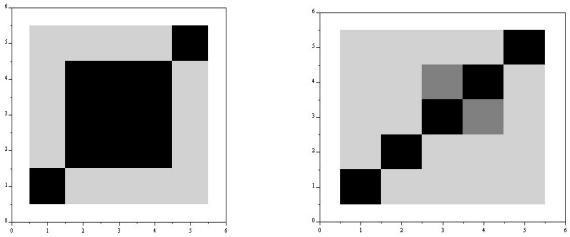


Figure II.2—4 : Mesure de proximités entre diverses occurrences de grosse caisse et caisse claire dans "By and Bye", sans puis avec transformation de la séquence

Dans l'exemple sans transformations, les grosses caisses donnent toutes des mesures de proximité proche de 1. On voit bien que la proximité diminue entre les grosses caisses lorsqu'une transformation violente est appliquée à la séquence.

II.2.1.b. PERFORMANCES D'UNE MESURE DE PROXIMITE BASEE SUR LA CORRELATION

Les expériences précédentes montrent que prendre comme critère de proximité la hauteur des pics de corrélation entre une séquence et les sons la composant est une méthode efficace de détection d'un son dans une séquence.
Plus important, la cohérence de cette mesure est robuste à des transformations (même importantes) des sons de la séquence.

Cependant, le problème important qui se pose est celui de la détermination de seuils ad hoc pour pouvoir déterminer le regroupement de divers sons.
De plus, dans le cadre réel d'application de notre algorithme de détection, nous ne disposons pas (au moins lors des premières itérations) d'un son assez proche de celui recherché dans le signal, et qui permettrait de se contenter d'utiliser simplement une mesure de hauteur de pics de corrélation. Les expériences précédentes suffisent uniquement à nous conforter dans l'idée que l'utilisation de la fonction de corrélation est sensée.
Dans la réalité de notre algorithme, la première corrélation est effectuée entre le signal musical et un son de référence, ce dernier étant étranger au signal.
L'exemple suivant montre la valeur absolue de la fonction de corrélation normalisée d'un extrait musical avec un son de référence générique.

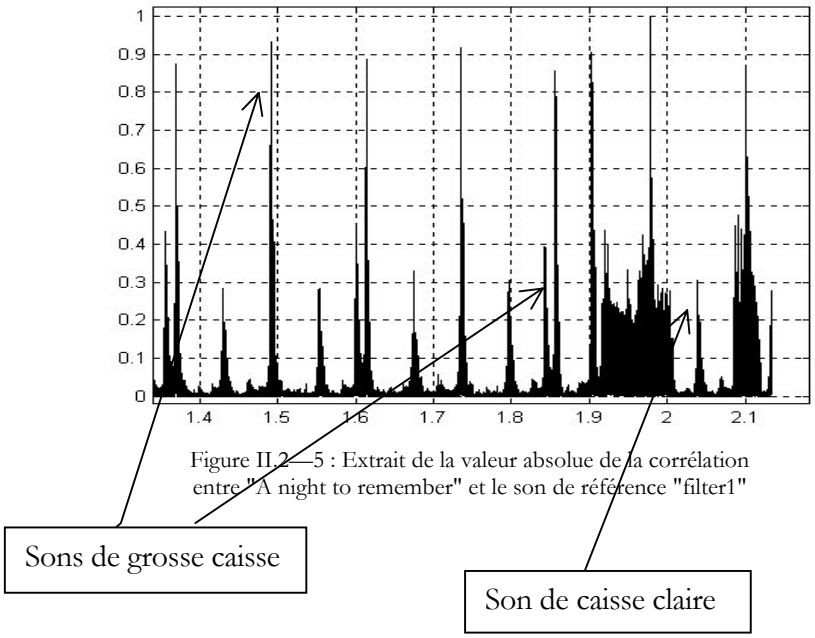


Figure II.2—5 : Extrait de la valeur absolue de la corrélation entre "A night to remember" et le son de référence "filter1"

On voit bien que dans les cas ou l'on ne regarde que la hauteur des pics de corrélation, on court le risque soit de ne pas détecter des occurrences de grosse caisse (si on fixe un seuil trop élevé), soit de commettre des erreurs de détection (si le seuil est bas) et prendre en compte des sons de caisse claire, mais aussi d'autre sons perturbateurs.

Finalement, un critère basé sur la hauteur des pics de corrélation entre le signal et un son de référence à caractère percussif nous permet de déterminer une série d'indice temporels auxquels sont présents quelques sons percussifs du signal. Il faut fixer un seuil en amplitude pour déterminer ces indices. Selon le son de référence initial, la détection sera orientée vers tel ou tel type de son. Avec une valeur de seuil fixée, il faut prendre en compte les deux problèmes suivants[5] : il existe un risque de ne pas détecter des occurrences du son recherché ; et il est probable que des occurrences de sons parasites ne soient pas "filtrées" par la corrélation. Le premier problème est abordé au paragraphe II.2.2, et le deuxième dans le prochain paragraphe.

[5] Respectivement l'un ou l'autre selon que le seuil est élevé ou faible.

II.2.1.c. DEFINITION DE CRITERES DE QUALITE

Il apparaît nécessaire de définir un critère de qualité, postérieur au calcul de la fonction de corrélation, dont l'objectif est de trier parmi les indices supérieur à un seuil donné quels sont les sons qui sont des occurrences d'un même instrument et quels sont les "déchets".
En regardant simplement les valeurs absolues des fonctions de corrélations de divers signaux avec divers sons de référence, on se rend compte que les enveloppes des pics de corrélation ont des formes particulières : des formes aiguës plus ou moins symétriques[6].

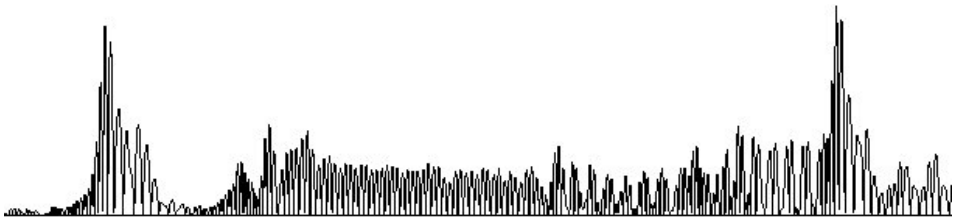


Figure II.2—6 : Extrait de la valeur absolue de la corrélation
entre "A night to remember" et le son de référence "filter1"

PREMIER CRITERE

Ainsi, un premier critère de qualité a été défini de telle sorte que les indices autour desquels la valeur absolue de la fonction de corrélation a une forme de pic soient gardés et les autres rejetés.
La définition mathématique de ce critère pour l'indice N_0 est la suivante :

$$Q1(N_0)=\frac{R_{xy}(N_0)^2}{\frac{1}{\Delta}\sum_{n=N_0-\Delta/2}^{N_0+\Delta/2} R_{xy}(n)^2} \quad (\text{où } \Delta \text{ est la largeur de pic})$$

Équation II.2—3 : Premier critère de qualité

On divise la hauteur du carré du maximum de corrélation par l'énergie de la fonction de corrélation autour de ce maximum. Si le critère est grand, le pic est aiguë, sinon il est obtus. On fixe là-aussi un seuil de décision.

Cependant, comme le montre la Figure II.2—7 suivante, les pics de corrélation d'un son percussif de référence avec des sons percussifs d'amplitude assez élevée possèdent à peu prés tous des enveloppes de formes aiguës.

[6] L'autocorrélation d'un son est exactement symétrique.

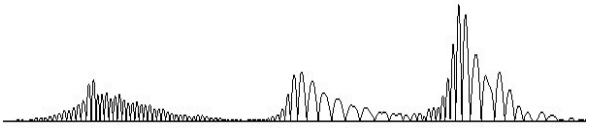


Figure II.2—7 : Extrait de la valeur absolue de la corrélation entre "A night to remember" et le son de référence "filter1"

Dans ce "zoom " sur la valeur absolue de la fonction de corrélation, les deux sons de droite correspondent à des occurrences de grosse caisse, et celui de gauche à une occurrence de caisse claire.
Ainsi, on peut s'attendre à ce que le critère précédemment défini soit efficace pour la ségrégation entre les pics correspondant à des sons percussifs et les autres, mais moins performant pour la ségrégation entre sons percussifs.

DEUXIEME CRITERE

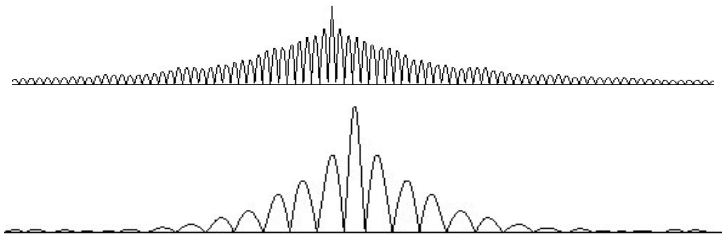
Le deuxième critère mis au point se base à la fois sur la forme de l'enveloppe de la corrélation à un indice donné, et sur la forme de la fonction de corrélation elle-même. On regarde à un niveau plus précis que l'enveloppe.

On fixe d'abord une largeur de pic δ par une itération sur les échantillons de la fonction de corrélation : δ grandit tant qu'il n'y a pas de recoupement de jupes, et tant que la pente de l'enveloppe correspond à un schéma donné (avec une certaine latitude).
On cherche ensuite à déterminer la proximité entre

- l'autocorrélation de l'instrument de référence
- la corrélation de l'instrument de référence avec le signal à un indice donné

A cette fin, on calcule la fonction de corrélation entre ces deux fonctions sur des régions de largeur δ autour de leur maximum respectif. Il s'agit en quelque sorte d'une "corrélation à l'ordre 2".

A titre d'exemple, voici les valeurs absolues des autocorrélations d'un son de caisse claire et d'un son de grosse claire :



Le second critère s'écrit finalement :

$$Q2(N_0)=\sum_{m=N_0-\delta/2}^{N_0+\delta/2}\left(\sum_{n=0}^{\delta}R_{yy}(n)\times R_{xy}(n-m)\right)$$

Équation II.2—4 : Deuxième critère de qualité

Dans le cas où deux pics ayant des enveloppes à peu près similaires, mais étant de formes différentes (comme dans la Figure II.2—7), le pic ressemblant au mieux à la valeur absolue de l'autocorrélation du son de référence aura un meilleur score que l'autre.

Une expérience de comparaison des cas de réussites ou d'échec des deux critères de qualité décrit ci-dessus a été menée sur une large base de donnée d'extraits musicaux.
Cette expérience a montré que, pour une très grande part des cas, le deuxième critère était le plus efficace dans la phase de ségrégation des indices fournis par l'algorithme de simple corrélation.

II.2.1.d. LES CONTRAINTES DE LA PHASE DE DETECTION

Définir un critère de qualité, comme dans le paragraphe précédent, afin d'aiguiller la phase de détection, correspond à la mise en oeuvre pratique de l'application de contraintes sur la détection.
Dans ce paragraphe, nous allons nous intéresser aux contraintes imposées au processus de détection qui ne sont pas laissées à notre choix, mais qui sont inhérentes à l'utilisation de la corrélation entre le signal musical et un son de référence.

Nous avons vu dans les exemples du paragraphe II.2.1.a qu'imposer un seuil en amplitude sur la valeur absolue de la fonction de corrélation entre le signal musical et un son de référence permet de définir une mesure de proximité entre sons (voir Équation II.2—2). Cette mesure implique des contraintes sur la détection.
D'une manière générale, à partir de la donnée d'une mesure de distance entre éléments, la théorie de l'analyse des données (voir [YOUNG]) nous enseigne qu'il existe un espace implicite de représentation des éléments associé à cette distance. Effectuer une discrimination en fonction des distances entre éléments s'apparente à faire des classifications dans cet espace multidimensionnel.
Dans notre cas, on impose des *contraintes* sur la phase de détection en imposant un *seuil en amplitude* et un *son de référence*. En envisageant le calcul de la mesure de proximité entre les sons d'un signal et un son de référence comme la projection des sons de ce signal dans un espace multidimensionnel, on peut voir ces contraintes comme la sélection géographique d'une région de cet espace. Changer le seuil ou bien le son de référence implique des modification des contraintes de la phase de détection ; ce qui peut être envisagé comme une évolution des frontières de la région privilégiée de l'espace de représentation des sons.

Le problème qui se pose est de déterminer quels sont les *dimensions* de l'espace de projection implicite à notre mesure de proximité.
On est en droit de penser que les dimensions de cet espace correspondent à certains paramètres physiques des sons dont on fait la corrélation. A ce propos deux directions de recherche nous ont occupées :

1. L'extraction de paramètres physiques des sons percussifs permettant de les différencier par familles. C'est l'objet du paragraphe II.3.
2. L'étude de la méthode d'analyse des données nommée MDS (pour *Multidimensional scaling*). Elle permet de trouver le nombre de dimensions de l'espace inhérent à une mesure de distance entre sons (voir [KRU/WISH] et [YOUNG]). Soulignons ici que trouver quels sont les paramètres physiques correspondant à ces dimensions n'est pas pris en compte par cette méthode.

ANALYSE MULTIDIMENSIONNELLE (*MULTIDIMENSIONAL SCALING*)

Comme on peut le lire dans [YOUNG], "*The essential ingredient defining all multidimensional scaling methods is in the representation of data structure*". Les données à analyser peuvent se présenter sous des natures différentes, et l'on peut définir divers modèles pour l'analyse.
Dans l'exemple décrit en page 16, où l'on fait la corrélation d'une séquence avec les éléments de cette séquence – donc l'exemple le plus simple –, les données se présentent sous la forme de matrices carrées, exactement symétriques et les éléments sont continus. Ce qui permet, dans la théorie développée par YOUNG, d'appliquer un type de modèle particulier (i.e. "*one-mode / two-way*").
Ce modèle a fait l'objet d'une implémentation qui donne des résultats cohérents dans les cas d'écoles de matrices de faibles dimensions. Dans le cas qui nous occupe en page 16, la matrice est de taille moyenne (e.g. 44 par 44) ; et les résultats d'expérimentations sont que le nombre de dimensions importantes est toujours inférieur de 1 au nombre de sons. Ceci n'est a priori pas satisfaisant, et laisse entrevoir des erreurs de programmations. De plus, les données du problème réel ne correspondent pas à ce modèle particulier, car la corrélation s'effectue entre une séquence de sons et des sons complètement étrangers à cette séquence. Pour ce type de donnée, YOUNG propose un modèle "*two-mode / two-way*" qui est théoriquement moins évident, et surtout plus complexe à implémenter que le premier.
Nous n'avons pas eu le temps de mener à terme cette étude. Cependant, nous y faisons référence au paragraphe II.4.3.

II.2.2. UTILITE D'UN ALGORITHME ITERATIF

Notre méthode de détection est basée sur le calcul de la fonction de corrélation entre le signal musical et un son de référence (ou "instrument"). Il est aisé de se rendre compte qu'une telle opération est équivalente au filtrage du signal par un filtre dont la réponse impulsionnelle est égale à la représentation temporelle retournée du son de référence.
En effet, on peut montrer la formule suivante :

$$R_{xy}(n)=x(n)*y(-n)\,.$$

NB : L'instrument y(n) étant à durée finie, nous nous trouvons bien dans le cas des filtres à réponses impulsionnelles finies, dont une propriété importante est qu'ils sont inconditionnellement stables.

L'utilisation du filtrage d'un signal pour détecter les phénomènes transitoires est une technique classique (voir e.g. [LEVINE]). De même, dans un cadre d'analyse rythmique,

SCHEIRER propose un filtrage du signal par bandes de fréquences (au nombre de six), et la recherche de pics d'amplitude ou d'énergie dans ces bandes fréquentielles (voir [SCHEIRER]). Ainsi, la procédure qui consiste en :

1. effectuer une corrélation entre un morceau[7] et un son de référence donné[8], puis
2. déterminer un seuil limite en amplitude sur cette fonction de corrélation pour détecter les indices auxquels un évènement rythmique potentiellement intéressant est présent

peut être rattachée à des techniques "classiques" existantes dans la recherche de détection d'évènements rythmiques.

On voit bien sur le graphique suivant que fixer un seuil en amplitude permettrait de déterminer des occurrences de la grosse caisse plus facilement sur la version filtrée que sur la version originale.

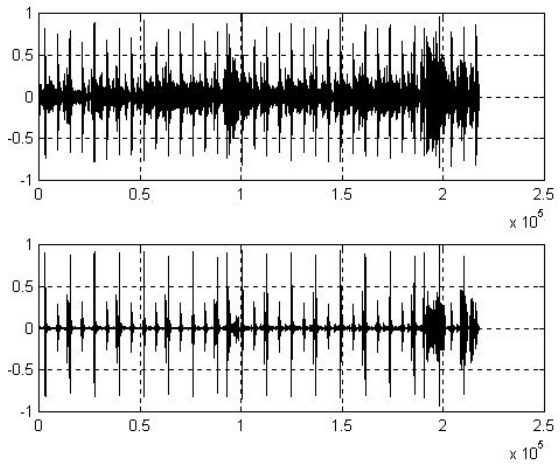


Figure II.2—8 : Extrait de "A night to remember" : original et version filtrée passe-bas

Sur une base de données de plus de deux cents morceaux, une expérimentation a montré l'insuffisance de la méthode consistant en successivement :

1. effectuer la corrélation et déterminer un seuil en amplitude (i.e. une méthode "classique")
2. utiliser le premier critère de qualité décrit page 21 pour améliorer la détection

L'utilisation du deuxième critère de qualité apporte une amélioration notable dans les résultats, mais il faudrait avoir un seuil en amplitude (sur la fonction de corrélation – ou signal filtrée) très bas pour pouvoir repérer toutes les occurrences d'un instrument. Dans un tel cas, la phase de détection donne beaucoup de déchets, souvent même plus que de bons indices. La mesure de la qualité n'est pas assez robuste pour être appliquée à un tel cas.

Comme nous l'avons déjà introduit en page 14, nous proposons d'améliorer cette méthode par un processus itératif, utilisant un nouveau son de référence à chaque

[7] e.g. : "A night to remember"

[8] e.g. : "filter1" est un signal généré par la réponse impulsionnelle d'un filtre passe-bas

itération. Il s'agit ici de préciser la définition du timbre recherché en même temps que l'on améliore la détection de ses occurrences. Plus l'itération est avancée, plus le son de référence est sensé être représentatif d'un timbre percussif présent dans le signal.
Le bon fonctionnement de la détection doit garantir la simplicité de la synthèse, et le bon fonctionnement de la synthèse doit garantir la simplicité de la détection suivante.

Nous verrons que différentes méthodes de synthèse d'un nouveau son de référence sont possibles, et que la phase importante de détermination de la qualité[9] des pics a donné lieu à des recherches supplémentaires. Le nombre d'itérations nécessaires dépend de la méthode de synthèse, et des hauteurs de seuils fixées. Les diverses améliorations sont détaillées au long de ce rapport et l'état actuel du programme est décrit au paragraphe II.6.
Cependant, il peut d'ores et déjà être utile de comparer sur deux exemples précis le résultat donné par :

- Une méthode non-itérative (i.e. "classique"), avec un seuil assez élevé (pour ne pas avoir trop de déchets dans la détection), utilisant de plus la deuxième mesure de qualité.
- Une méthode prenant en compte les aspects itératifs en plus des aspects qualitatifs.

Les résultats de la détection par les deux méthodes des grosses caisses de divers morceaux[10] ont montré que des occurrences d'amplitudes faibles ne sont pas trouvées par la première méthode (non-itérative), mais sont trouvées par la seconde (itérative).

[9] Répondre à la question "Ces sons ont-ils tous été générés par le même instrument?"

[10] e.g. : "A night to remember" ou encore "Everybody here wants you"

II.3. EXTRACTION DE PARAMETRES D'UN SON PERCUSSIF

Comme nous le verrons, des paramètres représentatifs des sons trouvent leur utilité dans des phases de classification (voir paragraphe II.4), mais aussi de synthèse (voir paragraphe II.5).
Le problème est le suivant : nous désirons extraire des paramètres d'un signal relativement court, supposé contenir une occurrence d'un son percussif, mais également, éventuellement, un niveau important de bruit (i.e. bruit d'enregistrement, ou bien signal correspondant à tout instrument autre que celui produisant le son percussif).
Une multitude de paramètres d'analyse peuvent faire l'objet de notre attention, notre choix est de nous intéresser en premier lieu aux paramètres issus d'une segmentation temporelle du signal, puis aux paramètres issus d'analyses spectrales, et enfin à d'autres types de paramètres.

II.3.1. SEGMENTATION TEMPORELLE DU SIGNAL

Extraire des paramètres d'un son sous-entend que l'on connaît les limites de ce son. Plus précisément, le "découpage" d'un son percussif en différentes régions cohérentes est justifiable par la connaissance a priori que l'on a de la manière dont ces sons sont générés. Classiquement, nous définissons un son percussif par une région d'attaque et une région de décroissance, la première correspond à l'établissement des modes propres de l'instrument ainsi qu'aux bruits produits par les vecteurs de leurs excitations (e.g. baguettes), la deuxième région correspond à la progressive perte de puissance de ces modes de vibrations.

NB : Il nous faut ici remarquer que si une telle segmentation du signal se justifie aisément dans le cas d'un instrument percussif, il ne faut pas la généraliser trop rapidement à tous les types de sons. La mise parallèle de la détermination des phases de segmentation et de caractérisation des signaux est un thème de recherche encore fertile (voir e.g. [ANDRE-OBRECHT] ou [BASS./NIK.]).

Nous proposons ici les détails de l'algorithme de segmentation des signaux, phase préliminaire à la détermination de paramètres.

On suppose que les signaux à segmenter sont du type suivant :

- bruit antérieur au son percussif
- attaque du son percussif
- décroissance du son percussif
- bruit postérieur au son percussif

II.3.1.a. TEMPS D'ATTAQUE

La manière la plus classique pour calculer le temps d'attaque d'un son percussif est d'avoir une approche de type "seuillage" du signal : on détermine un seuil au dessous duquel le signal est considéré comme étant le bruit d'enregistrement antérieur au son, et au dessus duquel le signal est considéré correspondre au son. Une telle approche n'est pas

envisageable dans le cas qui nous occupe. En effet, les sons percussifs présents dans les morceaux musicaux ne sont pas isolés, le bruit antérieur peut être très important, il ne correspond pas uniquement au bruit d'enregistrement mais à d'autres instruments que celui produisant le son percussif, et surtout, ce bruit est différent selon l'endroit dans le morceau et selon le morceau, ce qui implique que l'on ne peut pas fixer un seuil générique.
On utilise donc une autre méthode.
On calcule tout d'abord l'enveloppe de la valeur absolue du signal et on en détermine le maximum. Le calcul de l'enveloppe se fait par jonction des maximums locaux du signal, dans des fenêtres successives de signal de 100 échantillons. C'est une méthode grossière, mais suffisante à la détermination du maximum (i.e. *onset* du son percussif).

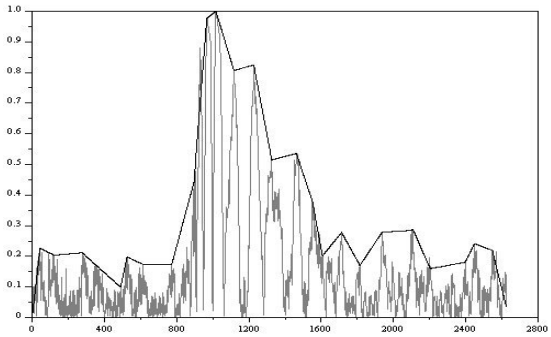


Figure II.3—1 : Enveloppe de la valeur absolue du signal entourant un coup de grosse caisse dans "By and Bye"

Une portion du signal antérieure au maximum est isolée, c'est celle dans laquelle on va calculer le temps d'attaque.
La technique employée génère tout d'abord l'enveloppe de cette courte portion du signal. Puis, on détermine le segment de l'enveloppe ayant la pente la plus grande, l'endroit où son prolongement coupe l'axe des abscisses est considéré comme le début du son percussif ; cette méthode permet de faire face à des proportions importantes de bruit antérieur au son percussif.
Comme nous désirons une bonne précision dans l'estimation du temps d'attaque, il ne peut être acceptable de fixer arbitrairement le nombre de fenêtres couvrant le signal – comme il est fait pour le calcul de l'enveloppe globale. L'écueil à éviter est de prendre un nombre trop grand de fenêtres et de rentrer ainsi dans l'intra-période du signal (comme il est illustré dans à la Figure II.3—2 b)). Le nombre de fenêtres dans lesquelles prendre les maximums locaux est usuellement choisit après une transformée de Fourier du signal : la TF donne la fréquence dominante, son inverse donne un nombre d'échantillons (i.e. la taille de fenêtre) pour lesquels il n'y aura qu'un maximum local. Les signaux étudiés étant très courts (entre 30 et 200 échantillons), une TF ne permet pas une assez bonne précision fréquentielle. La méthode finalement utilisée effectue les calculs d'enveloppes pour un nombre de fenêtres augmentant progressivement, le nombre choisit est le médian de ceux pour lesquels une stabilité du calcul du temps d'attaque est atteinte (voir Figure II.3—2 a)).

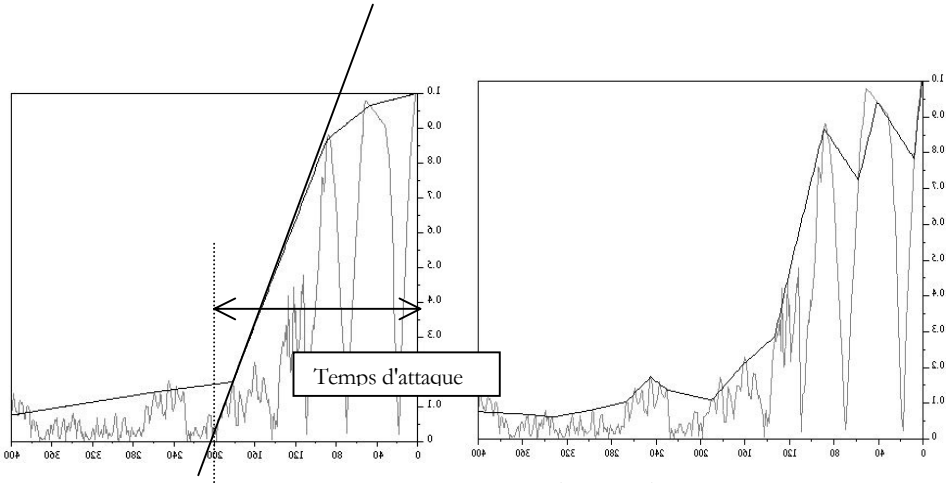


Figure II.3—2 a) et b) : Enveloppe de l'attaque d'un coup de grosse caisse dans "By and Bye". Respectivement bon et mauvais nombre de fenêtres pour le calcul.

II.3.1.b. TEMPS DE DECROISSANCE

La partie de décroissance des sons étudiés ici correspondant à des régions de signal dont les attributs sont beaucoup plus stables que ceux des parties d'attaques. Il nous a donc semblé acceptable de fixer une durée de décroissance des signaux percussifs égal à un multiple du temps d'attaque.
NB : Des travaux ont été commencés ayant comme objectif une définition plus précise des temps de décroissance (voir page 56).

II.3.2. PARAMETRES SPECTRAUX

Etant donné la nature des régions d'attaque et de décroissance des sons percussifs, il nous semble pertinent de chercher à caractériser un son percussif par des paramètres spectraux relatifs à sa partie d'attaque et d'autres relatifs à sa partie de décroissance.
C'est ce que nous allons faire, en développant diverses méthodes d'analyses spectrales : la modélisation de Prony et l'analyse de Fourier.

II.3.2.a. MODELE DE PRONY

Il s'agit ici de modéliser le signal comme une somme de sinusoïdes amorties.

DEVELOPPEMENTS THEORIQUES

Modèle théorique

Le modèle de Prony idéal est défini comme suit :

$$x(n)=\left(\sum_{m=1}^{2L} B_m \times Z_m{}^n\right)$$

Équation II.3—1

où: $Z_m=e^{-\alpha_m}\times e^{j2\pi f_m/Fs}$ et $B_m=A_m\times e^{j\theta_m}$

Paramètres de Prony :
f_m est la fréquence
α_m est le taux d'amortissement
A_m est l'amplitude
θ_m est la phase

On peut montrer (voir annexe V.1) qu'il existe un vecteur $(1,a_1,\cdots,a_p)$ tel que $x(n)+\sum_{k=1}^{p} a_k \times x(n-k)=0 \quad \forall n \geq p+1$, les Z_m étant les racines des a_k.

On peut former la matrice de covariance à partir de N échantillons du signal comme suit :

$$C=M^t M \text{ où } M=\begin{pmatrix} x(p)\cdots & x(0) \\ \vdots & \\ x(N) & x(N-p) \end{pmatrix}$$

Il est prouvé dans [LAROCHE] que si $p>2L$, cette matrice est singulière, de rang 2L.

Il est démontré dans [LAROCHE] que tout vecteur du noyau de la matrice de covariance du signal forme un polynôme dont les racines sont directement liées aux fréquences et taux d'amortissements du modèle.

Ainsi, une méthode simple, basée sur la détermination du noyau de la matrice de covariance, permet de déterminer les paramètres du modèle de Prony.

L'étude de ce modèle et la détermination des paramètres restent théoriques. Il faut introduire un autre modèle pour se rapprocher de signaux réels. En effet, il est important de souligner qu'aucun signal (à l'exception des signaux de synthèse) n'est une somme exacte de sinusoïdes amorties. Seuls certains signaux se rapprochent de ce modèle. La démarche de modélisation paramétrique d'un signal a ceci de dangereux qu'elle ne donne aucune information sur la validité du choix du modèle, mais donne dans tous les cas des valeurs aux paramètres (même quand le choix du modèle n'est pas pertinent)!

Modèle #1

Le signal est maintenant considéré comme la somme d'un signal de Prony idéal et d'un bruit blanc e(n), terme appelé « erreur de modèle »:

$$x(n)=\left(\sum_{m=1}^{2L} B_m \times {Z_m}^n\right)+e(n)$$

Équation II.3—2

La matrice de covariance du signal a le même ensemble de vecteurs propres que celle du signal idéal. Mais le problème est maintenant que le noyau de la matrice de covariance n'est plus trivial à trouver. En effet, quelque soit la taille de cette matrice, aucune valeur propre n'est exactement nulle, contrairement au cas précédent.

Nous avons implémenté la méthode appelée *Truncated Singular Value Decomposition* développée dans [KUM/TUFTS]. Elle permet de déterminer le rang de la matrice de covariance, et donc le nombre L de sinusoïdes, qui définissent le signal de Prony idéal $\tilde{x}(n)$ se rapprochant le plus du signal x(n), c'est à dire qui minimise la norme quadratique de la différence entre $\tilde{x}(n)$ et x(n), i.e. l'erreur de modèle e(n).

Il faut ici rappeler que l'application que nous cherchons à développer s'effectue sur des signaux extrêmement bruités, dans le sens où tout autre signal que le son percussif est considéré comme du bruit dans notre analyse. Malheureusement, l'estimation des paramètres d'un modèle de Prony développée ci-dessus est très sensible au bruit.

Plus précisément, les artefacts de cette méthode dus à la présence de bruit sont les suivants :

- Le nombre estimé de composantes sinusoïdales diminue avec le niveau de bruit. Le niveau maximal de bruit ajouté au modèle de Prony au-delà duquel deux sinusoïdes proches ne peuvent plus être séparées et sont amalgamées en une seule, est lié au conditionnement du signal[11]. Or, dans un cas réel, le niveau de bruit nous est imposé, tout comme le conditionnement du signal.
- La manière dont deux fréquences sont amalgamées en une seule consiste en un moyennage des fréquences, i.e. on estime une seule fréquence là où deux fréquences existent, et sa valeur estimée se situe entre les deux valeurs originales.

Il apparaît donc nécessaire d'introduire un nouveau type de modélisation.

Modèle #2

Considérer le bruit (que l'on sait important) comme erreur de modèle engendre des analyses médiocres, le signal est donc maintenant considéré comme la somme d'un signal de Prony idéal x(n), d'une erreur de modèle e(n), et d'un terme de bruit perturbateur w(n) (cf [GOUYON]).

$$x(n)=\left(\sum_{m=1}^{2L} B_m \times Z_m{}^n\right)+e(n)+w(n)$$

Équation II.3—3

L'objectif initial était d'introduire à l'algorithme précédant un module réduisant l'effet du bruit $w(n)$ sur la phase d'estimation. Ainsi, dans le cas idéal où un tel module pourrait effectivement détecter ce qui dans le signal à analyser est $w(n)$, il ne resterait plus qu'à minimiser l'erreur de modèle e(n), ce qui est la tâche qu'effectue avec succès l'analyse de Prony définie précédemment. Cependant, il n'a pas été possible de trouver un algorithme permettant de faire une distinction nette entre les termes perturbateurs appartenant à e(n) et ceux appartenant à w(n). Nous avons donc plutôt orienté notre recherche vers des algorithmes permettant de prendre en compte des niveaux de bruits très élevés, amalgamant ainsi le terme d'erreur de modèle dans celui du bruit additif.
Nous avons vu page 29 qu'un signal de Prony idéal x(n) vérifie :

$x(n)+\sum_{k=1}^{p} a_k \times x(n-k)=0 \quad \forall n \geq p+1$, ce qui correspond à la réponse impulsionnelle d'un filtre ARMA, après l'ordre p. L'expression de la réponse impulsionnelle d'un filtre ARMA pour tous les indices temporels étant :

$$x(n)+\sum_{k=1}^{p} a_k \times x(n-k)=\sum_{k=0}^{p} b_k \times \delta(n-k) \quad \forall \mathrm{n} \geq 0$$

Nous avons par conséquent utilisé une approche concernant l'estimation des paramètres d'un processus ARMA lorsqu'il lui est ajouté un bruit blanc de forte puissance. La méthode développée par MAYNE et FIROOZAN a été étudiée, elle est appelée *Three-Stage Least Square method* et est explicitée dans [KAY] ; elle est relative à l'estimation des paramètres

[11] Rapport de la valeur propre maximale sur la valeur propre minimale de la matrice de covariance

d'un modèle ARMA et se base sur l'estimation préliminaire du bruit blanc en entrée, puis détermine en deux étapes successives les parties AR et MA du processus.
Les implémentations successives ont porté sur :

1. l'estimation des paramètres d'un processus ARMA synthétique
2. l'estimation des paramètres d'un signal de Prony synthétique (donc idéal) auquel divers bruits blancs de puissances importantes ont été ajoutés.
3. l'estimation des paramètres de Prony à partir d'un signal réel (supposé correspondre au modèle Prony+bruit additif important)

Le fait d'appliquer la même méthode pour les points 2 et 3 illustre l'amalgame qui est fait entre l'erreur de modèle et le bruit perturbateur. Cependant, cette petite entorse par rapport à l'objectif initialement fixé trouve une justification dans les résultats pratiques recueillis. La méthode ainsi définie semble fournir des résultats meilleurs que celle explicitée dans le modèle #1, comme le montrent les exemples du paragraphe suivant.

NB : Il est important de souligner le fait que la suprématie de cette méthode sur la précédente est directement liée à l'importance du bruit. Si le signal correspond exactement à une somme de sinusoïdes amorties et qu'il n'y a aucun bruit perturbateur, c'est alors le premier modèle qui est optimal. L'objectif étant d'extraire de l'information d'occurrences de sons percussifs issus de signaux musicaux bruts, et donc mélangés la plupart du temps à d'autres sons, il nous semble cohérent d'utiliser plutôt la dernière méthode en ce qui concerne la modélisation de Prony.
Les détails des méthodes d'estimations des fréquences, amplitudes, taux d'amortissements et phases du modèle de Prony sont explicités en annexe V.1, page 60.

EXEMPLES

Avant de se demander si les paramètres issus d'une analyse de Prony sont pertinents ou pas (voir paragraphe II.4 et paragraphe II.5 pour les utilités des paramètres), il faut s'assurer de la validité des cette analyse. On peut le faire en écoutant les sons de synthèses dont les paramètres pilotant la synthèse sont issus de phases d'analyse de sons réels par Prony.
Pour les sons dont les segmentations en régions d'attaque et de décroissance sont parfaites (effectuées manuellement par exemple), on peut se demander si le modèle de Prony fournit des résultats qui semblent cohérents pour chacune de ces régions.
De nombreux sons ont été analysés puis resynthétisés en utilisant le modèle de Prony. Il s'agit de sons de synthétiseurs (propre), mais également de sons directement issus de signaux audios polyphoniques (donc plus ou moins bruités).
Nous ne proposons ici que quelques figures illustrant les synthèses..

Parties d'attaques des sons percussifs

Sons isolés (monophoniques) (issus de *drumkits*, i.e. banques sonores des synthétiseurs)

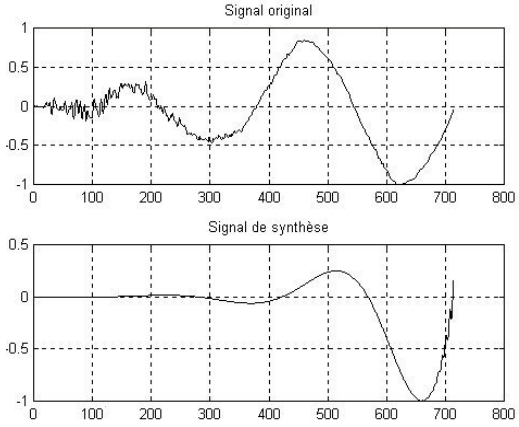


Figure II.3—3 : Synthèse par Prony de la partie d'attaque d'un son de grosse caisse isolé (63 ms)

Parties de décroissance des sons percussifs

Sons isolés (monophoniques)

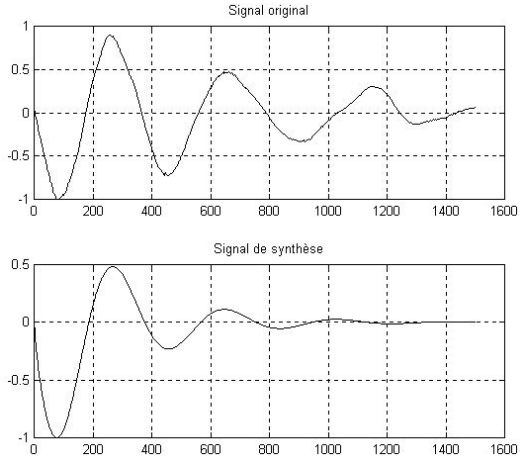


Figure II.3—4 : Synthèse par Prony de la partie de décroissance d'un son de grosse caisse isolé (145 ms)

II.3.2.b. COMPARAISON AVEC L'ANALYSE DE FOURIER

POINT DE VUE THEORIQUE

Tout d'abord, il faut rappeler que ces deux techniques d'analyse des signaux appartiennent à des familles différentes : la *modélisation paramétrique des signaux* et la *représentation des signaux*. Le modèle de Prony est un modèle paramétrique, la FFT ne modélise pas le signal mais le représente d'une manière utile dans les problèmes d'analyse spectrale. L'analyse de Fourier permet de visualiser les parties sinusoïdales et non-sinusoïdales du signal en proposant un continuum de données spectrales (dans la définition théorique de la transformée de Fourier, l'indice de la somme est infini). La détermination de paramètres tels que les fréquences et amplitudes de partiels doit se faire par un algorithme postérieur à la FFT (i.e. détection de pics, puis paramétrisation). Au contraire, le modèle de Prony fait l'hypothèse de "sinusoïdalité" du signal dès le départ (dans la définition, l'indice de la somme est fini). C'est ici que se gagne de la *précision fréquentielle*. D'un autre coté, et

contrairement à l'analyse de Fourier, la modélisation n'offre aucun moyen de savoir si l'hypothèse initiale est justifié ou non.

POINT DE VUE EXPERIMENTAL

Parties d'attaques des sons percussifs

Pour un son dont l'attaque est de 213 échantillons[12] (i.e. à peu près égale à 19 ms, ce qui est l'ordre d'idée pour un son de grosse caisse), la précision fréquentielle d'une analyse de Fourier est d'environ 51 Hz.
Pour un son dont l'attaque est de 56 échantillons (i.e. 5 ms), la précision fréquentielle est pratiquement de 200 Hz!
On se rend compte que si une information fréquentielle importante est contenue dans les parties d'attaques des sons percussifs, une analyse par FFT ne nous permettra vraisemblablement pas d'y avoir accès avec précision.
Des expérimentations ont montré que les analyses par FFT de régions d'attaque de sons percussifs ont des précisions fréquentielles telles que, quel que soit le son, les paramètres issus de FFT sont toujours très proches. Ce qui n'est évidemment pas acceptable.

Parties de décroissance des sons percussifs

Ces parties sont plus longues que les parties d'attaques, ainsi une analyse de Fourier y trouve plus de sens en terme de précision fréquentielle.
Cependant, des expérimentations ont montré que si un son est composé de plusieurs composantes très proches, une analyse de Fourier ne permet pas une assez bonne précision pour pouvoir les séparer. Ainsi les battements d'amplitudes[13] caractéristiques de certains sons ne peuvent être analysés.
Un autre élément caractéristique des signaux percussifs qui ne peut pas être pris en compte par une analyse de Fourier est la présence de composantes fréquentielles qui s'éteignent rapidement. Une analyse par FFT attribue à de telles composantes une amplitude très faible. Par contre, dans la modélisation de Prony, les taux d'amortissements sont des paramètres totalement décorrélés des amplitudes, ce qui permet de rendre compte des composantes s'éteignant rapidement, que leurs amplitudes soient faibles ou importantes.

II.3.2.c. CONCLUSIONS SUR LES PARAMETRES SPECTRAUX

Afin de conclure cette partie concernant l'extraction de paramètres issus d'analyses spectrales des signaux, il nous faut introduire les défauts de notre implémentation du modèle de Prony.
Le défaut majeur est que les modélisations effectuées ne se sont pas montrés robustes par rapport à la précision de segmentation du signal. Il est évident qu'aucune méthode de segmentation automatique ne peut certifier que les limites trouvées correspondent à l'échantillon prêt aux limites réelles des différents segments du son, la méthode d'analyse spectrale se doit donc d'être robuste aux petits changements (i.e. de quelques échantillons) des limites des segments d'attaque ou de décroissance. Dans certains cas, le

[12] Rappelons que la fréquence d'échantillonnage est de 11025 Hz, comme dans tout le rapport.

[13] Ces battements sont dus à la présence de deux composantes fréquentielles très proches.

modélisation de Prony implémentée donne des résultats différents si la limite des régions du signal est changée (e.g. de 10 échantillons).
Un autre défaut consiste dans le fait que les signaux à analyser sont inégalement bruités, et notre méthode de Prony est construite pour faire face à des signaux n'étant pas bruités à l'extrême, mais n'étant pas non plus sans bruit (pour lesquels une modélisation de Prony du type décrit en page 30 est plus adapté). Les signaux réels issus de régions différentes de morceaux de musique étant pour certains très bruités et pour d'autres non bruités (et oui, il y en a!), les performances de notre algorithme s'avèrent inégales (voir les exemples de synthèses au paragraphe II.5).

II.3.3. AUTRES PARAMETRES

II.3.3.a. ENERGIE

L'énergie moyenne d'un signal est un paramètre simple à calculer et qui s'avère souvent très utile.
Nous calculons l'énergie moyenne du signal x de longueur N comme suit :

$$E(x)=\frac{1}{N}\sum_{n=0}^{N-1}x(n)^2$$

II.3.3.b. TAUX DE PASSAGE A ZERO

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. De plus, nous avons fait en sorte de tenir compte des bruits additifs de différents type :

- Hautes fréquences : un seuil en amplitude est fixé de telle sorte que les composants fréquentiels de hautes fréquences et d'amplitudes inférieures de 30 dB au son percussif ne soient pas considérés dans le calcul du ZCR. En effet, nous situant dans un environnement polyphonique, il est très probable que le son percussif dont on cherche à extraire les caractéristiques ne soit pas le seul à participer au fait que le signal franchisse l'abscisse zéro. A l'endroit où la note d'un instrument percussif est jouée, les autres instruments sont considérés comme du bruit pour la mesure du ZCR. Le point de vue adopté ici est que la méthode de segmentation autour du maximum d'amplitude (i.e. *onset*) du son percussif est assez fiable pour considérer que dans les régions où l'on calcule les ZCR, les autre instruments sont plus faibles de 30 dB.
- Basses fréquences : Les instruments jouant des notes de très basses fréquences[14] contribue non pas à des passages intempestifs du zéro, mais à un élèvement (ou abaissement) du niveau moyen du signal, ce qui peut avoir des conséquences sur le ZCR. Un module de détection puis de suppression d'un tel phénomène est présent dans l'algorithme de calcul du ZCR.

[14] Par rapport à l'inverse de la taille de la fenêtre

Un intérêt a priori du ZCR est que, bien que calculé dans le domaine temporel, il est corrélé avec des aspects spectraux du signal, KEDEM (voir [KEDEM]) le considère comme une mesure de la *fréquence dominante*. Ainsi, même dans les régions très courtes (e.g. les attaques peuvent être inférieures à 20 ms), nous disposons d'une mesure corrélée au contenu spectral des signaux.

II.4. CLASSIFICATION DE SONS PERCUSSIFS

Etant donnée une série d'indices temporels auxquels sont présents divers évènements sonores, le but du module de classification est de "faire le tri" entre ces évènements. L'algorithme de détection basé sur la corrélation n'est pas optimal, il faut déterminer quels sont les indices correspondant aux évènements sonores proches du son de référence et ceux correspondant à des évènements qui en sont plus éloignés.
La méthode la plus instinctive est certainement de calculer un critère de qualité comme il est décrit dans le paragraphe II.2.1.c.
Cependant il nous a semblé important de poursuivre des recherches relatives aux thèmes de la classification et de la reconnaissance des formes.

Posons le problème comme suit : étant donnés plusieurs sons percussifs[15], nous voulons définir une méthode permettant de les séparer en deux classes, ceux proches du son de référence, et ceux éloignés.
Nous suivons la démarche suivante :

1. Identification des dimensions significatives des sons percussifs par une méthode de classification en mode supervisé sur une base de donnée relativement petite (voir II.4.1)
2. Confrontation du résultat avec une base de donnée beaucoup plus importante, en utilisant une méthode de classification en mode non-supervisé (voir II.4.2)

NB : Les expériences suivantes ne concernent que des morceaux de musique populaires, dont les éléments percussifs sont regroupés en caisses claires ou grosses caisses.

II.4.1. METHODE D'ANALYSE FACTORIELLE DISCRIMINANTE (AFD)

L'AFD correspond à la phase de prétraitement de méthodes de classification supervisée (voir annexe page 62 sur les méthodes de classification).
Dans les cas qui nous occupe, nous nous situons clairement dans une problématique d'analyse en mode supervisé. En effet, de par le fait que nous ayons une intuition sur les différentes classes auxquelles nous auront à faire face (e.g. grosses caisses et caisses claires), et que nous possédions des représentants de ces classes (e.g. dans tout morceau ou tout synthétiseur), nous disposons d'une base d'apprentissage de données expertisées.

La philosophie de l'AFD est de calculer le plus grand nombre possible de paramètres à partir des signaux expertisés. Puis, parmi tous ces paramètres, on cherche à déterminer les plus pertinents dans la tâche de classification selon les classes connues. L'idée initialement introduite par FISHER (voir [TOURNERET]) consiste à déterminer des axes (ou axes factoriels) tels que les projections des données expertisées sur ces axes permettent de séparer les classes.

[15] issus de la phase de détection par corrélation

Comme point de départ, les sons sont donc projetés dans un espace multidimensionnel redondant constitué de 19 paramètres[16] :

- Temps d'attaque
- Temps de décroissance
- Temps entre maximum d'amplitude et indice de pente d'enveloppe maximum dans la région d'attaque
- Nombre de sinusoïdes trouvées par modélisation de Prony de la région d'attaque
- Nombre de sinusoïdes trouvées par modélisation de Prony de la région de décroissance
- Fréquence d'amplitude maximale trouvée par modélisation de Prony de la région d'attaque
- Fréquence d'amplitude maximale trouvée par modélisation de Prony de la région de décroissance
- Taux de croissance exponentielle de la composante fréquentielle d'amplitude maximale trouvée par modélisation de Prony de la région d'attaque
- Taux d'amortissement de la composante fréquentielle d'amplitude maximale trouvée par modélisation de Prony de la région de décroissance
- Fréquence d'amplitude maximale trouvée par FFT de la région d'attaque
- Fréquence d'amplitude maximale trouvée par FFT de la région de décroissance
- Fréquence d'amplitude maximale trouvée par FFT de tout le son
- Energie moyenne calculée sur la région d'attaque
- Energie moyenne calculée sur la région de décroissance
- Energie moyenne calculée sur tout le son
- Proportion entre les énergies moyennes de la région d'attaque et de décroissance
- Taux de passage à zéro (ZCR) de la région d'attaque
- ZCR de la région de décroissance
- ZCR de tout le son

On calcule grâce au critère de FISHER les pouvoirs de discrimination de chacun de ces paramètres. On peut ainsi diminuer la dimensionalité de notre espace de représentation en ne gardant que les dimensions pertinentes à la séparation de classes.

IDENTIFICATION D'UNE DIMENSION PERTINENTE

Les signaux expertisés sont fournis par la banque de sons General MIDI du synthétiseur Korg05RW.
Ces sons se divisant en deux classes, nous ne voulons utiliser qu'un paramètre parmi les 19 calculés.
Le résultat de l'AFD donne le *ZCR de la région de décroissance* comme étant le paramètre le plus adapté.

NB : Le second paramètre le plus pertinent à la classification est le ZCR de la région d'attaque.

[16] NB : Dans [Scheir/Slan], il y a 13 paramètres

II.4.2. VALIDATION PAR CLASSIFICATION HIERARCHIQUE ASCENDANTE

Afin de déterminer si les résultats de l'AFD ont effectivement une valeur générale, nous proposons d'effectuer des tests de classification d'autres sons percussifs.
Des expériences de classifications hiérarchiques ascendantes (ou encore *agglomerative clustering*) nous paraissent adaptées à cette démarche. Cette méthode de classification en mode non-supervisé suppose qu'aucune base d'apprentissage n'est disponible. Les éléments à classer sont représentés par une mesure selon une certaine métrique (ici, c'est la mesure du ZCR des régions de décroissance). Au départ, à un élément correspond une classe, puis on diminue progressivement le nombre de classes en associant les éléments entre eux selon leur distance. Le dernier regroupement forme une seule classe. Des distances inter-groupes doivent être définies à partir de la distance inter-éléments. La distance entre deux groupes peut être définie comme la distance entre les barycentres de ces groupes, ou encore entre les éléments les plus éloignés de chaque classe. Ces deux distances ont été étudiées.
Comme on ne considère pas de base d'apprentissage, si cette méthode – poussée jusqu'au regroupement des tous les éléments en deux classes – forme deux classes correspondant aux sons de type caisse claire et à ceux de type grosse caisse, alors la mesure du ZCR de la région de décroissance sera considérée pertinente.
Les deux expériences suivantes nous poussent à penser qu'elle l'est.

SONS MONOPHONIQUES (PROPRES)

Les premiers sons utilisés pour la validation sont des sons des kits percussifs du Korg05RW (à l'exception du kit General MIDI).
En utilisant l'une ou l'autre des mesure de distances inter-groupe définies précédemment, le regroupement en deux classes a résulté en une classe de caisses claires et une classe de grosses caisses avec un taux de réussite de 94.5%.

SONS REELS (BRUITES)

Les autres bases de sons utilisées pour la validation sont constituées de sons réels, directement issus de morceaux de musique.
Plaçons-nous dans le cas où l'on cherche à détecter les occurrences de grosse caisse, le son de référence utilisé pour la corrélation est donc construit comme la réponse impulsionnelle d'un filtre passe-bas de large bande.
Une région quelconque extraite d'un morceau de musique populaire a une grande probabilité d'être un signal polyphonique, l'instrument percussif (e.g. grosse caisse) que l'on cherche à repérer est donc noyé dans un "bruit" composé en grande partie des autres instruments (guitare, voix,...).

NB1 : Rappelons qu'il n'est pas question ici de faire une première séparation des instruments percussifs dans un flot instrumental complexe, puis de déterminer postérieurement les caractéristiques qui permettraient de les distinguer. Le problème de la séparation d'instruments est bien trop compliqué pour l'envisager comme une phase préliminaire de notre projet. De plus, les timbres récurrents qui nous permettent de percevoir les rythme d'un morceau ne correspondent pas à de simples instruments, mais

sont des sons "de type" grosse caisse ou caisse claire. On s'intéresse au signal comme un timbre dont une importante caractéristique est la nature percussive, et l'on cherche à extraire des paramètres permettant de caractériser la nature récurrente d'un timbre participant à notre perception du rythme.

NB2 : Notons également que l'on ne cherche pas à déterminer un hypothétique espace des timbres percussifs universel, mais plutôt à déterminer un moyen de différencier deux timbres présents à plusieurs endroits dans *un* morceau de musique donné.
Ainsi, cette expérience ne porte pas sur le regroupement des grosses caisses d'un grand nombre de morceaux (ni de même avec les caisses claires). On effectue *une expérience par morceau de musique*.

Pour chaque extrait de 20 secondes des morceaux étudiés, le nombre d'occurrences de sons percussifs (grosses caisses et caisses claires) varie entre 19 et 63. Pour chacun de ces morceaux, la classification hiérarchique ascendante – utilisant la mesure du ZCR sur les régions de décroissance – a sût discriminer les sons avec des taux de réussite variant entre 87.5% et 97%.

Nous basant sur ces résultats, nous pensons qu'il est justifiable, dans un cadre de classification, d'utiliser comme mesure représentative des sons percussifs le calcul du ZCR sur leurs parties de décroissance.

II.4.3. VERS UNE METHODE DE DETECTION ET CLASSIFICATION OPTIMALE?

La classification effectue la projection du signal musical – aux indices fournis par la phase de détection – dans un espace que nous avons toute latitude de déterminer. Comme nous l'avons vu au paragraphe II.4.1, le choix que nous avons fait est celui d'une certaine simplicité : un espace uni-dimensionnel dont l'axe correspond au taux de passage à zéro des parties de décroissance des sons percussifs.
Attribuer aux sons issus de la phase de détection un score en fonction de leur positionnement sur cet axe permet une meilleure mesure de la "qualité" d'un son que les critères de qualité intuitifs développés en page 21.
Cependant, le fait même d'appliquer une méthode en deux temps (corrélation puis mesure de qualité de certains sons du signal) est la conséquence d'une intuition. Notre travail au long du stage nous pousse à croire qu'il peut y avoir un moyen d'unifier les processus de mesure de proximité par corrélation et de discrimination entre sons.
En effet, si l'on pouvait :

1. déterminer quelles sont les dimensions de l'espace dans lequel on projette le signal sonore lorsque l'on en calcule la valeur absolue de la corrélation avec un son de référence donné,
2. faire correspondre à ces dimensions des paramètres physiques caractéristiques des sons percussifs, et permettant de les classifier,

alors on posséderait un cadre d'analyse regroupant toutes les contraintes appliquées à la phase de détection : celles imposées et celles désirées.

Pour ceci, il faudrait tout d'abord poursuivre l'étude théorique des contraintes inhérentes au à l'utilisation de la corrélation, en menant à terme l'analyse de type MDS introduite au paragraphe II.2.1.d.

II.5. SYNTHESE DE SONS PERCUSSIFS

La méthode de synthèse d'un nouveau son de référence doit répondre à plusieurs critères. Elle doit utiliser comme paramètres d'entrée :

1. Le son de référence de l'itération précédente
2. Les sons réels présents dans les régions du signal musical aux indices fournis par la phase de détection

Le seul moyen véritablement efficace de juger de la valeur d'une méthode de synthèse est de pratiquer une écoute comparative entre les sons analysés et les mêmes sons resynthétisés.
Cependant, il ne faut pas perdre de vue que nous abordons le problème de la synthèse sous un angle particulier. Il ne s'agit pas de définir la méthode d'analyse/synthèse la plus fidèle de toutes, mais celle répondant au mieux à nos besoins. Faisant l'hypothèse que les indices fournit par la phase de détection correspondent à des occurrences du même instrument, nous essayons de synthétiser un son à partir de paramètres extraits de différentes régions du signal. Les différentes occurrences d'un instrument ne sont pas toutes exactement similaires ; de plus, elles sont noyées dans des types de bruits différents. Ainsi, la méthode de synthèse ne peut évidemment pas fournir une resynthèse parfaite d'un son percussif, son objectif est plutôt de capturer la nature *récurrente* d'un instrument percussif.
La phase de classification des sons percussifs extraits d'un signal musical polyphonique (voir paragraphe II.4) a précisément comme objectif de déterminer à partir d'une série d'indices temporels quels sont ceux qui correspondent à des sons du type caisse claire, et quels sont ceux qui correspondent au type grosse caisse. Cependant, comme nous l'avons vu, des paramètres tels que les taux de passage à zéro des parties de décroissance et d'attaque des sons sont plus discriminants que les paramètres de nature spectrale. Bien qu'il soient pertinents dans un cadre de classification, ces paramètres ne sont pas suffisants pour effectuer des synthèses sonores. Les paramètres spectraux sont plus facilement utilisables dans une méthode de synthèse, ce sont ceux que nous utiliserons.
Un effort devra donc être poursuivi pour déterminer des paramètres permettant une très bonne discrimination entre type de sons percussifs, et pouvant également piloter un modèle de synthèse (le paragraphe IV.2 résume les directions devant être poursuivies pour l'achèvement du projet général).

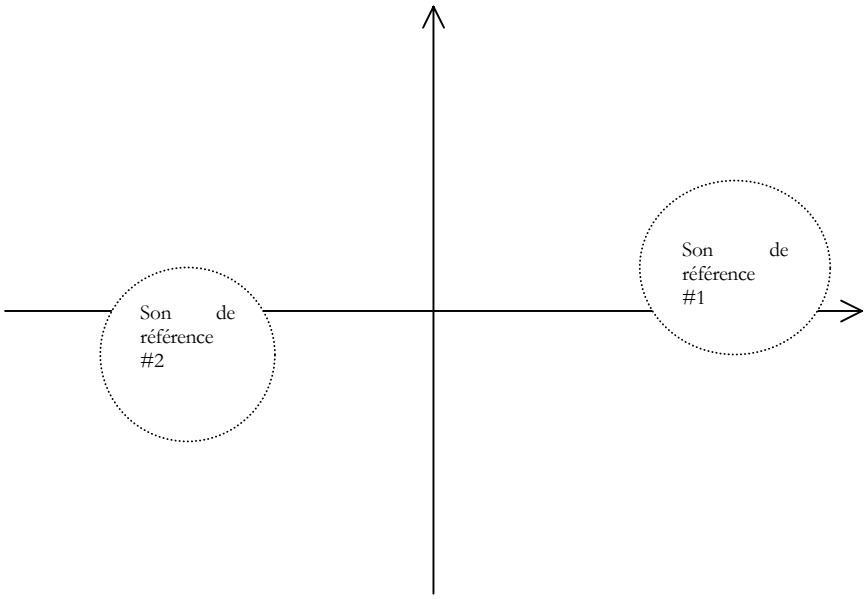
Poursuivant l'intuition qu'une étude plus approfondie de la modélisation de Prony fournira vraisemblablement des tels paramètres, nous proposons une méthode d'analyse/synthèse basée sur une segmentation des sons percussifs en parties d'attaque et de décroissance, une modélisation de Prony étant appliquée sur chacune de ces parties.
Comme il a été introduit plus haut (page 34), certaines resynthèses sont acceptables, d'autres souffrent d'artefacts de la méthode de Prony.

Les conclusions que nous tirons des synthèses effectuées sont les suivantes. En dehors du fait que notre implémentation du modèle de Prony doit être améliorée, nous remarquons tout d'abord que la modélisation de la partie d'attaque d'un son percussif comme une somme de sinusoïdes exponentiellement croissantes ne semble pas pertinente.

Remarquons ici que bien que l'importance des régions d'attaque des sons soient généralement considérées comme primordiales, aucune méthode de modélisation flexible de ces régions n'existe. En revanche, il semble plus acceptable d'utiliser le modèle de Prony pour les régions de décroissance des sons percussifs. Une part très importante de ces sons est le bruit, pour lequel aucun modèle n'a été implémenté pendant notre stage, ceci pourrait faire partie d'un travail futur, se basant par exemple sur les travaux de SERRA (voir [SERRA]).

CONTRAINTES SUR LA SYNTHESE

Si on dispose d'un espace de représentation[17] pour la synthèse des sons percussifs (dans le schéma suivant il s'agit d'un espace bi-dimensionnel pour des raisons de clarté), il est possible de définir des contraintes géométriques sur la synthèse des nouveaux sons de référence. Par exemple, on peut définir des zones de l'espace entourant le point de représentation de chaque son auxquelles on interdit de se chevaucher (e.g. des sphères comme dans le schéma suivant).
De cette manière, on garantit que les deux sons seront effectivement différents, et donc que les phases de détections des occurrences des deux types de timbres seront aiguillées différemment. On s'assure du fait que les deux séries temporelles finales seront différentes.



[17] Un paramètre de synthèse correspond à une dimension de l'espace

II.6. ETAT ACTUEL DE L'ALGORITHME D'EXTRACTION D'EVENEMENTS RYTHMIQUES

Les paramètres d'*entrée* du programme d'extraction consistent en :

- Un extrait d'une vingtaine de seconde d'un morceau de musique
- Deux sons de références

La *sortie* du programme consiste en deux séries d'indices temporels auxquels deux types différents d'évènements rythmiques perceptivement important sont présents.

Pour chacun des sons de références, l'algorithme suivant est appliqué :

1. On effectue la valeur absolue de la corrélation entre l'extrait musical et le son de référence. Puis on prend sa valeur absolue et on normalise afin que le maximum soit égal à 1.
2. On fixe un seuil en amplitude égal à 1
3. On diminue progressivement ce seuil en amplitude jusqu'à ce que le nombre d'indices supérieurs au seuil, et correspondant à des sons issus du même instrument, soit égal à un nombre fixé (e.g. 4 indices pour 10 secondes).
4. On construit un nouvel instrument de référence à partir du son initial et de l'analyse des extraits du signal aux indices issus de l'étape 3 (soit par simple addition de signaux, soit par la méthode introduite au paragraphe II.5).
5. On effectue la valeur absolue de la corrélation entre l'extrait musical et le nouveau son de référence, suivie par la normalisation de sa valeur absolue.
6. On effectue des itérations selon les étapes suivantes:
    - Le seuil en amplitude est baissé progressivement jusqu'à ce que le nombre d'indices supérieurs à celui-ci soit supérieur ou égal à deux fois le nombre d'indices à l'itération précédente.
    - Parmi ces indices, on ne garde que ceux correspondant aux sons issus tous du même instrument.
    - On construit un nouvel instrument de référence à partir du son initial et de l'analyse des extraits du signal aux indices issus de l'étape précédente. La proportion d'utilisation (pour la synthèse) des sons réels issus du signal augmente avec le nombre d'itérations.

    Cas d'arrêt des itérations (l'un ou l'autre) :
    - Le nombre maximum d'itérations fixé (e.g. 6) est dépassé
    - Deux itérations successives donne les mêmes résultats

L'algorithme est itératif et adaptatif comme on peut le constater dans ses diverses étapes.
Dans les étapes 3 et 6, le test devant être effectué sur les sons présents aux indices potentiellement importants[18], et dont l'objectif est de déterminer si ils sont tous issus du même instrument, peut être effectué par différents algorithmes :

[18] i.e. dont la hauteur du pic de corrélation est supérieure au seuil en amplitude

- Une mesure de qualité des pics de corrélation peut être effectuée de différentes manières (voir II.2.1.b et II.2.1.c).
- Une phase de classification (voir II.4) des sons peut être utilisée pour opérer à une discrimination entre les sons

La plupart des expériences d'extraction d'évènements percussifs ont été effectuées avec la deuxième mesure de qualité des pics de corrélation décrite au paragraphe II.2.1.c. C'est le cas de l'expérience décrite au paragraphe III.1.
Cependant, il nous paraît vraisemblable que l'utilisation de l'algorithme de classification (par mesure du taux de passage à zéro sur la partie de décroissance des sons, voir II.4) au lieu de la mesure de qualité des pics de corrélation donnera des résultats encore meilleurs. Nous n'avons pas trouvé le temps de faire une expérience comparative de grande échelle entre ces deux méthodes, aussi nous proposons d'inclure ceci au paragraphe concernant les travaux futurs : IV.2.1.a.

II.6.1. DISCUSSIONS

II.6.1.a. LES SONS DE REFERENCE

Le choix du son de référence initial est primordial dans le processus d'extraction des occurrences d'un timbre percussif. Nous avons vu qu'un son conçu comme la réponse impulsionnelle d'un filtre passe-bas et à large bande permettait d'aiguiller la détection vers des occurrences de grosses caisses. Au contraire, si le filtre est passe-bande de fréquence centrale plutôt aiguë, la détection s'oriente vers des sons générés par des caisses claires.
Le son de référence initial doit être assez général pour permettre de détecter in fine un grand nombre de timbres percussifs différents, mais il doit également être particulier dans le sens où l'on veut qu'il aiguille assez rapidement la détection vers un timbre particulier. C'est le cœur du processus itératif d'accepter de ne trouver que quelques occurrences au premier passage, et de raffiner par étapes successives la détection en même temps que la définition du son de référence.

On pourrait penser prendre un son réel de grosse caisse (ou de caisse claire) de synthétiseur comme son de référence initial, mais cette option ne fonctionne généralement pas car ces sons ne sont pas assez généraux pour pouvoir aiguiller la détection vers *tout* son de grosse caisse ; et ils ne seront a fortiori pas capables d'aiguiller la détection vers des sons autres que des grosses caisses.
En effet, les exemples considérés jusqu'à présent ne concernent pratiquement que des sons de grosses caisses et caisses claires, mais nous savons d'ores et déjà qu'il faudra être capable de trouver des occurrences de sons autres que ceux-là (percussions, ou tout son très percussif et récurrent[19]). Dans un exemple particulier[20], la détection itérative dont le son de référence initial était un filtre passe-bas a permis de repérer les occurrences d'une percussion, très différente d'une grosse caisse.

[19] Même ceux pouvant avoir une nature harmonique (accord récurrent de guitare saturée, etc.)

[20] Dans le morceau "Didi"

II.6.1.b. CONCLUSIONS SUR L'EXTRACTION D'EVENEMENTS RYTHMIQUES

On a vu au paragraphe II.4.3 qu'une amélioration probable de la phase de détection pouvait consister en l'unification des processus de mesure de proximité par corrélation et de discrimination entre sons. Le moyen d'effectuer cette amélioration est de rentrer dans le cadre de la classification des sons par projection dans un espace multidimensionnel, la définition des contraintes appliquées à la phase de détection s'en trouvent simplifiées. Les dimensions de cet espace devraient pouvoir être trouvées par les méthodes introduites page 24.

Rappelons que la philosophie de l'extraction d'évènements rythmiques se base sur le partage d'informations entre un module de détection et un module de synthèse – au sein du processus itératif. Comme nous l'avons vu au paragraphe II.5, les paramètres permettant de synthétiser les sons percussifs doivent encore être améliorés. Insistons ici sur le fait que les paramètres utilisés pour la représentation des sons ne sont pas forcément des paramètres utilisables dans un cadre de synthèse sonore. Par exemple, le taux de passage à zéro est un bon paramètre pour effectuer une discrimination entre sons, mais il ne peut pas être un "levier de commande" d'un outil de synthèse.

En conclusion, il serait pratique autant qu'élégant d'utiliser des paramètres similaires pour la représentation des sons du signal musical (puis leur discrimination), et pour la synthèse des sons percussifs permettant d'effectuer l'itération suivante. Il nous a été possible de définir les termes de ce problème, d'avancer dans des directions de recherche qui lui sont associées (voir MDS page 24, et détermination de paramètres pertinents page 37), mais malheureusement pas de mettre un point final à ce problème.

III. REPRESENTER LE RYTHME – APPLICATIONS

Dans ce chapitre, nous partons de l'hypothèse que la méthode d'extraction des séries d'indices temporels et des sons originaux fonctionne convenablement. Soulignons ici le fait qu'une telle tâche peut d'ailleurs être effectuée manuellement, c'est d'ailleurs le cas dans [BROWN1]. A partir de la donnée des évènements rythmiques d'extraits musicaux, le problème qu'il reste à résoudre est celui de l'utilisation de ces informations dans un but de représentation pertinente de ces différents extraits, permettant donc de les différencier.

III.1. REPRESENTER LE RYTHME

Nos ambitions concernant la représentation automatique des rythmes ont été exposés en page 11, ce paragraphe décrit l'état actuel d'avancement du projet.

III.1.1. TEMPO ET ESPACE DE REPRESENTATION

Une analyse basique du rythme d'un morceau musical, tel que celui dont un extrait est présenté en exemple, fondée sur l'écoute, permet de déterminer aisément la pulsation, le tempo et la métrique.
Il peut donc être intéressant de comparer une telle analyse faite "à l'oreille" à une analyse automatique effectuée à partir de l'étude des fonctions de corrélation entre les séries générées par la méthode décrite au chapitre précédent.

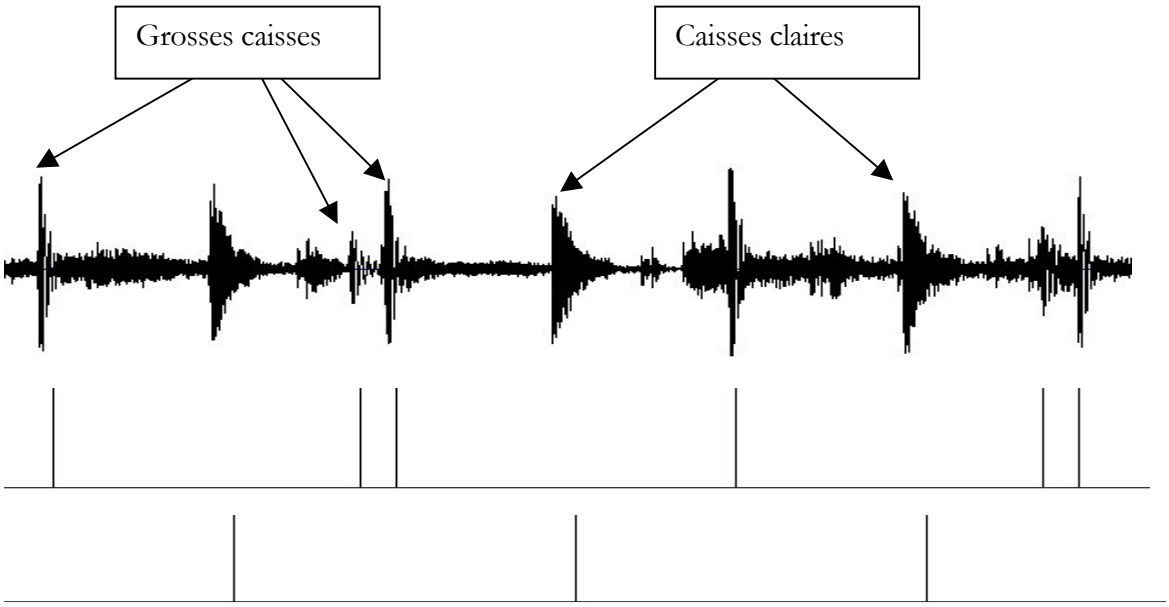


Figure III.1—1 a), b) et c) : Extrait du signal audio "A night to remember", puis extraits des séries temporelles obtenues par la méthode itérative avec, comme son de référence initial, respectivement un filtre passe-bas et un filtre passe-bande.

L'écoute du morceau puis son analyse rythmique basique nous indique qu'une noire sépare le premier coup de grosse caisse du premier coup de caisse claire, deux noires séparent les deux premiers coups de grosse caisse d'amplitudes importantes, de même, deux noires séparent des coups de caisses claires consécutifs, et une double-croche sépare le coup de grosse caisse d'amplitude faible du suivant d'amplitude importante.
Instinctivement, un auditeur taperait du pied en marquant les noires. Le *tempo* est de 105 noires par minute.
Une notation conventionnelle de ce morceau se ferait avec une *signature* en 4/4 car la *pulsation se fait au niveau des noires* (d'où numérateur = 4) et le motif rythmique de base utilise 4 noires (d'où dénominateur = 4).
NB : La Figure III.1—1 a) illustre une mesure et demi.

Penchons-nous maintenant sur les informations que l'on peut obtenir à partir de l'étude des séries temporelles illustrées sur la Figure III.1—1 b) et c).
Nous possédons deux liste d'indices d'occurrences, ce qui peut nous fournir trois fonctions de corrélations : l'autocorrélation de la première série, celle de la deuxième et l'intercorrélation des deux séries.

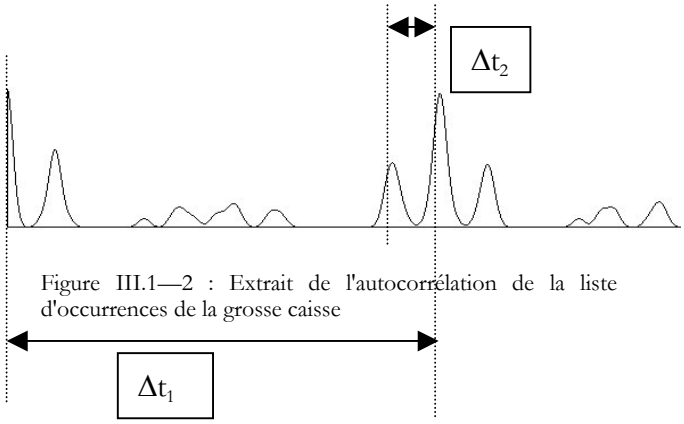


Figure III.1—2 : Extrait de l'autocorrélation de la liste d'occurrences de la grosse caisse

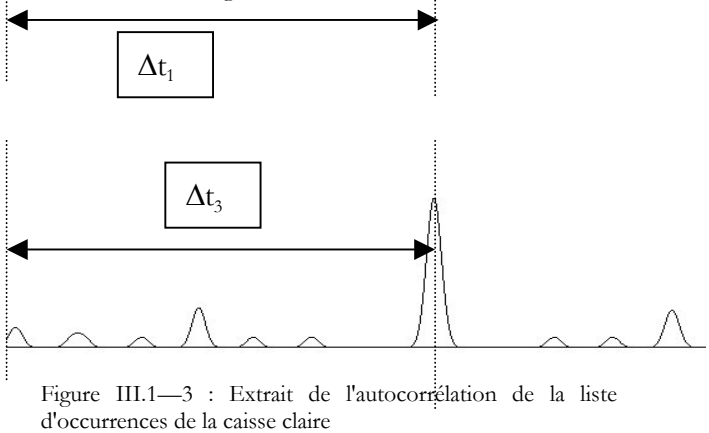


Figure III.1—3 : Extrait de l'autocorrélation de la liste d'occurrences de la caisse claire

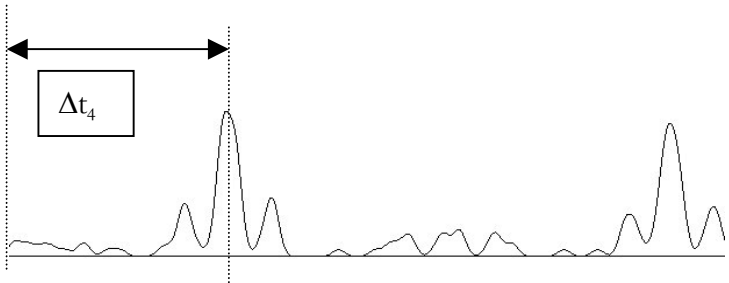


Figure III.1—4 : Extrait de l'intercorrélation des listes d'occurrences de grosse caisse et de caisse claire

NB : Les corrélations ne sont pas discrètes car les listes d'occurrences ne fournissent pas des indices espacés d'écarts *exactement* similaires. Nous avons donc introduit un lissage par fenêtres de Gauss de la fonction de corrélation calculée.

Des informations temporelles peuvent être extraites postérieurement à une détection de maximums. Dans l'exemple étudié, il s'agit de Δt_1, Δt_2, Δt_3 et Δt_4. Ces données sont exprimées en échantillons.

On peut vérifier que Δt_1 correspond bien à l'écart temporel entre deux grosses caisses d'intensité similaires (forte ou faible) ; cet écart est effectivement récurrent dans le morceau. La quantité Δt_2 correspond à l'écart entre les coups forts et faibles de grosse caisse, cet écart est également récurrent. De même, la quantité Δt_3 montre la récurrence de l'écart temporel entre les caisses claires. Et finalement, Δt_4 montre la récurrence de l'écart entre les grosses caisses et caisses claires de fortes intensités.

Dans le cadre d'une analyse automatique, nous déterminons la *pulsation* N, en nombre d'échantillons, comme le premier maximum supérieur à un seuil donné dans l'intercorrélation entre les deux séries, il s'agit ici de Δt_4.
Nous proposons de définir le *tempo* comme suit :

$$tempo = 60 \times \frac{F_e}{N}$$

Il paraît évident que les autres informations temporelles que l'on peut extraire des corrélations entretiennent un lien étroit avec la notion de *métrique*[21] que l'on peut appréhender par une analyse subjective de l'extrait musical. Nous n'avons cependant pas encore pu définir un cadre rigoureux pour la détermination automatique de la métrique.
Dans l'état actuel de l'algorithme nous définissons trois fonctions à partir des trois corrélations tronquées entre 0 et 2N. Puis, nous superposons à chacune 7 peignes de longueur 2N comprenant 1 à 7 pics également espacés, ce qui nous permet de déterminer 21 coefficients. Ces coefficients caractérisent les participations des instruments à diverses diviseurs ou multiples de la pulsation. Par exemple, un résultat pourrait être le suivant : la grosse caisse participe beaucoup à une division temporelle égale à 2 fois la pulsation, mais participe peu à la division temporelle égale à 1/3 pulsation, etc.

Bien que le développement d'un algorithme permettant de déterminer automatiquement la métrique et les 21 coefficients représentatifs d'un extrait musical ne soit pas encore achevé, il nous semble intéressant d'introduire en quoi un tel algorithme pourrait être utile:

- Les analyses musicologiques trouveraient certainement un intérêt dans la détermination automatique de la métrique d'un morceau (cf. paragraphe III.2.2).
- Les applications de descriptions musicales basées sur le contenu (cf. paragraphe III.2.1) peuvent trouver un intérêt dans l'utilisation d'un espace – e.g. à 21 dimensions – représentatif des rythmes des morceaux.

D'un autre côté, l'extraction automatique du tempo fonctionnant apparemment bien, nous proposons dans le paragraphe suivant une expérience de validation de l'algorithme d'extraction des séries et de calcul du tempo.

[21] Comme nous l'avions pressenti en page 11, il y a une redondance dans les informations fournies par les deux séries, nous pouvons en effet vérifier que $\Delta t_2 = \Delta t_3$.

III.1.2. EXPERIMENTATION SUR UNE BASE DE DONNEES

Une base de donnée de 206 extraits[22] de 20 secondes de morceaux de musique populaire a été soumise à une expérience dont l'objectif était de définir deux timbres percussifs par morceau, de leur associer deux listes d'indices temporels, puis d'en déduire le tempo.

Il est difficile d'envisager une comparaison exhaustive de séries issues de l'algorithme avec des séries subjectives. La quantité de données à comparer est bien trop importante et leurs qualités trop compliquées pour que la comparaison puisse s'envisager de manière automatique et rigoureuse. En effet, à un extrait de 20 secondes peuvent correspondre deux séries de plus de 50 indices temporels chacune. De plus, deux personnes pourraient définir différemment les séries pour un même morceau.
En revanche, le tempo est en général une notion consensuelle qui se prête bien à une comparaison automatique des résultats objectifs issus de l'algorithme et des résultats subjectifs d'écoute par un auditeur. En effet, les données à comparer sont simples et en nombre réduit : à un morceau correspond un tempo.
Ainsi, nous avons déterminé subjectivement les tempi des 206 morceaux et les avons comparés aux tempi résultant de l'expérience.

75% des tempi ont été correctement trouvés.

NB : L'écoute de certaines séries[23] nous a fait percevoir qu'à certains morceaux ne peut correspondre qu'une série temporelle cohérente (s'il n'y a qu'une grosse caisse et pas de caisse claire par exemple). Les analyses sur ce type de morceaux résultent le plus souvent en une bonne série et une deuxième série ne contenant que quelques indices temporels n'ayant pas de sens. Le plus souvent, la détermination du tempo a fonctionné même dans ces cas-là.

[22] fichiers au format Wave, Fe = 11025 Hz, 16 bits

[23] par l'intermédiaire de ce que nous avons appelé les "clics témoins" : signaux de très courte durée ajoutés au signal musical étudié aux indices d'occurrences d'un timbre particulier, ils se perçoivent facilement comme des "clics".

III.2. APPLICATIONS

III.2.1. DESCRIPTIONS MUSICALES FONDEES SUR LE CONTENU

La naissance du nouveau format standard d'indexation et de transfert d'informations audios[24] MPEG7 est caractéristique de la volonté actuelle de définir des descripteurs des contenus sémantiques des données audios. De tels descripteurs peuvent être extrait manuellement, mais il est évident que cette tâche serait facilitée par des méthodes automatiques ancrées sur le signal audio lui-même.
Afin d'approcher le contenu sémantique d'un signal musical, il pourrait être utile de posséder des descripteurs de l'harmonie de ce morceau, du type d'instrument jouant, du type de voix (féminine, masculine, rauque, aiguë, etc.), du type de mélodie, du rythme, etc.

Notre travail a pour objectif d'extraire de signaux audios des descripteurs rythmiques classiques tels que le tempo et la métrique, mais également de proposer un espace pertinent de représentation du rythme dans lequel des morceaux de rythmes similaires seront proches et des morceaux de rythmes différents seront éloignés. Les 21 dimensions de cet espace étant continues, il devrait être possible d'envisager des progressions fluides de tels styles rythmiques vers tels autres.

III.2.1.a. PROGRAMMATIONS MUSICALES

Les radios ou discothèques trouveraient évidemment un intérêt dans un tel outil de classification des morceaux de musique. Une application pourrait par exemple consister en la programmation automatique de plusieurs heures de musique.

III.2.1.b. SERVEURS MUSICAUX SUR INTERNET

La distribution de musique sur Internet est en plein essor. Il est suffisant de posséder un ordinateur et un modem pour avoir potentiellement accès aux millions de morceaux musicaux disponibles sur toute la surface du globe. Mais la musique est disséminée de manière souvent anarchique. Pour un internaute désireux de découvrir une musique qu'il serait susceptible d'apprécier, la tâche la plus difficile n'est pas d'avoir accès à l'information, mais d'avoir accès à une information ordonnée. Il n'est pas envisageable d'écouter "toute" la musique de l'Internet, et de faire ensuite ses choix. Il serait donc certainement utile de posséder un outil logiciel utilisant des descripteurs sémantiques musicaux, pouvant ainsi guider l'utilisateur dans ses découvertes musicales.
L'intérêt évident est de ne pas imposer à l'utilisateur l'utilisation de "mots-clés", qui dans un cadre musical sont pour la plupart inadaptés, non consensuel, et toujours dépassés[25], mais d'utiliser une classification interne des styles.
Le projet européen CUIDADO, auquel participent notament l'Ircam et Sony CSL, est précisemment centré autour de la conception d'un tel "navigateur musical sur Internet".

---

[24] MPEG7 concerne également les données visuelles

[25] Que signifie le terme "musique rock" aujourd'hui?

NB : Encore une fois, l'aspect rythmique de la musique n'est qu'un des aspects sur lesquels un tel logiciel doit se fonder.

III.2.2. ANALYSES MUSICOLOGIQUES

Les analyses musicales se fondent souvent sur la notion de métrique et, dans la conception architectonique du rythme, sur les notions de niveaux rythmiques primaires, inférieurs et supérieurs. Un outil de détermination automatique de la métrique d'un morceau trouverait donc un intérêt pour ces analyses. En revanche, le lien que l'on pourrait faire entre une représentation multidimensionnelle du rythme et les notions de niveaux rythmiques n'est pas encore assez clair. Cela nécessiterait certainement de faire partie de nos perspectives de travail.

IV. CONCLUSIONS

Dans ce chapitre, nous tirons les conclusions du travail effectué et nous définissons les directions de recherche qu'il nous semble important de suivre pour mener à bien le projet d'extraction et de représentation des rythmes.

IV.1. CONCLUSIONS

Dans ce rapport, nous développons la compréhension que nous avons de la recherche existante relative au thème de l'extraction du rythme, et nous proposons une approche de ce thème fondée sur l'analyse de *bas niveau* de signaux audios musicaux, permettant la détermination des caractéristiques rythmiques de *haut niveau*.

Le chapitre II détaille la plus grande partie de notre travail de recherche : l'extraction automatique d'évènements rythmiques. Nous explicitons les études théoriques et réalisations pratiques effectuées. L'apport majeur de ce travail est la mise en place d'un algorithme permettant, à partir de deux signaux de références, de définir par étapes progressives deux sons représentatifs d'un morceau de musique populaire donné. Les instants d'occurrences de ces sons sont trouvés avec précision. Les listes d'évènements ainsi définies offrent une information rythmique symbolique des morceaux de musiques, plus propice aux diverses utilisations que les échantillons du signal audio.
Cette information est proposée comme flux d'entrée à une analyse de plus haut niveau, introduite au chapitre III, qui focalise sur les notions de représentation du rythme et d'extraction de descripteurs rythmiques génériques. Les applications visées sont décrites dans ce même chapitre.

Le paragraphe suivant propose un survol des directions qu'il nous semble pertinent de poursuivre dans le but de mettre un terme au projet général d'extraction automatique de descripteurs rythmiques dans les signaux audios de musiques populaires.

IV.2. PERSPECTIVES

IV.2.1.a. TRAVAIL FUTUR

Il nous semble intéressant de consacrer un paragraphe à la description des travaux futur ou en cours, et dont la poursuite mènera vraisemblablement le programme d'extraction d'évènements rythmiques et de représentation du rythme à son terme définitif.

Ces travaux sont relatifs à :

1. Une expérience de grande échelle (i.e. sur la base des 206 morceaux) comparant les qualités et défauts des différentes méthodes de "tri" entre indices proposés par la phase de détection. Ces méthodes sont décrites au paragraphe II.2.1.c et II.4.1.
2. La détermination de nouveaux paramètres des sons percussifs tels que des paramètres cepstraux (cf. [BROWN2]), la distorsion harmonique dans les régions de décroissances (cf. [HERRERA]), des paramètres issus de diverses modélisations, ou bien des paramètres dérivés de méthodes de représentations temporelles, fréquentielles ou conjointes (e.g. représentations en ondelettes) des signaux. Les pouvoirs de discrimination de ces paramètres doivent être testés par la méthode introduite page 37.
3. L'analyse multidimensionnelle (MDS) introduite en page 24, qui devrait permettre de déterminer les dimensions implicites de la mesure de distance entre sons fondée sur le calcul de la fonction de corrélation entre un signal et un son de référence. La connaissance de ces dimensions permettrait de regrouper toutes les contraintes de la phase de détection : celles imposées et celles désirées (voir paragraphe II.4.3).
4. La détermination de paramètres de discrimination des sons percussifs qui puissent également contrôler le module de synthèse (voir II.6.1.b).
5. La recherche dans la direction des modélisations non-linéaires. En effet, il est bien connu que les parties d'attaques des sons sont extrêmement importantes pour leurs perception et catégorisation par l'oreille humaine. Malheureusement, nous ne disposons pas à l'heure actuelle de paramètres (extractibles par des méthodes de traitement du signal) permettant de caractériser et synthétiser ces portions de sons courtes et très instables.
6. L'amélioration de la phase de segmentation précise des sons percussifs. Le calcul de paramètres représentatifs de ces sons se fondant sur les régions limitées par le segmentation, il semble important de posséder une technique robuste. En particulier, la région de décroissance doit être défini par rapport au bruit postérieur au son percussif.
7. L'amélioration de l'analyse de Prony et l'élimination des artefacts introduits au paragraphe II.3.2.c.

IV.2.1.b. OUVERTURES

Nous proposons dans ce paragraphe différentes idées et pistes de recherche qui ont été abordées lors du stage, mais qui, ne paraissant pas primordiale, ou bien trop éloignées de l'objectif fixé, ou encore trop ambitieuse pour un stage court, n'ont pas été menées à leurs termes.

LIEN AVEC L'ANALYSE/SYNTHESE

En relation avec l'objectif de modélisation des signaux fortement non-stationnaires – comme les régions d'attaque des sons percussifs – introduit un peu plus haut, il serait vraisemblablement intéressant d'envisager une procédure de classification des sons percussifs s'intégrant au cadre plus général de l'Analyse / Synthèse. Les méthodes existantes d'Analyse / Synthèse se basant sur des modèles "Sinus + Bruit" (e.g. [SERRA], et logiciel SMS[26]) permettent de bonnes reconstructions de signaux musicaux mêmes complexes et polyphoniques, cependant, elles souffrent toutes du même mal : l'analyse de Fourier se faisant sur une fenêtre de signal, il est implicitement supposé que le signal est stationnaire sur cette fenêtre. Précisément, c'est cette assertion qui nous éloigne de l'analyse (et donc la synthèse) des sons fortement non-stationnaires[27]. Nous intéressant aux sons percussifs, dont les attaques sont fortement non-stationnaires, il devrait donc être intéressant d'effectuer des procédures de classification sur les signaux résiduels d'analyses "Sinus + Bruit". Des expérimentations de séparation des parties déterministes et résiduelles d'extraits musicaux ont montré que les sons percussifs impliquent la perception de rythmes en grande partie par leurs caractéristiques non-stationnaires. En bref, une analyse du rythme, centrée autour des récurrences de timbres particuliers, pourrait sûrement s'effectuer sur des signaux issus d'un "pré-filtrage par analyse en somme de sinusoïdes".

LIEN AVEC UNE ANALYSE PERCEPTIVE

On pourrait envisager effectuer une étude des dimensions subjectives du rythme dans les musiques populaires.
Ceci impliquerait la confrontation de données objectives physiques des morceaux de musique avec des données subjectives issues de tests psychoacoustiques effectués sur divers sujets.
Les données subjectives consisteraient en des mesures de similarités entre les rythmes de différents titres musicaux récoltées sur les sujets. Une analyse de ce type de données peut être effectuée par une analyse de type INDSCAL (*INDividual SCALing*) comme il est proposé dans [GABRIELSSON].
Nous avons introduit au paragraphe II.2.1.b les analyses multidimensionnelles (MDS) se fondant sur les mesures de similarités. Dans le cas présent, il faut élargir une telle analyse à trois dimensions en faisant l'hypothèse que des dimensions subjectives communes à tous les individus caractérisent l'appréhension du rythme, seules les sensibilités à ces diverses dimensions divergent selon les individus. Ces méthodes, dénommées "*three-way MDS*" sont introduites dans [CAR/WISH] et [CAR/CHANG].
Une étude de ce type a été effectué par GABRIELSSON, qui, en cherchant a interpréter les différentes dimensions trouvées dans l'espace subjectif, a proposé :

1. "Mesure"
2. "Rapidité"
3. "Tempo"
4. "Uniformité–variation ou simplicité–complexité"

---

[26] Spectral Modeling Synthesis (voir http://www.iua.upf.es/~sms/)

[27] Des améliorations de ces méthodes par des modèles de type "Sinus + Bruit + Résiduel" se trouvent dans le littérature récente (e.g. [LEVINE]).

5. "Motif basique"
6. "Caractère de mouvement"

Afin de qualifier les axes déterminés par une analyse psychoacoustique, il serait certainement plus pertinent d'analyser des données objectives physiques calculées à partir des signaux musicaux.

V. ANNEXES

V.1. ANNEXE I : MODELE DE PRONY/REPONSE IMPULSIONNELLE DE FILTRE ARMA

Cette annexe a pour objectif d'expliciter la correspondance qui existe entre un signal correspondant à un modèle de Prony idéal (i.e. sans erreur de modèle) et la réponse impulsionnelle d'un filtre ARMA. Le modèle de Prony est d'ordre L (nombre de sinusoïdes) ; et le filtre ARMA est d'ordre [p,p] (respectivement nombre de zéros et de pôles), avec $p=2L$.

Le modèle de Prony idéal est le suivant :

$$x(n)=\left(\sum_{m=1}^{2L} B_m \times Z_m{}^n\right)$$

où: $Z_m=e^{-\alpha_m}\times e^{j2\pi\cdot f_m/Fs}$ et $B_m=A_m\times e^{j\theta_m}$

La forme générique d'un filtre ARMA est la suivante :

$$H(Z)=\frac{\sum_{k=0}^{p} b_k \times Z^{-k}}{1+\sum_{k=1}^{p} a_k \times Z^{-k}}$$

DETERMINATION DES PARAMETRES DE PRONY A PARTIR DE CEUX D'UN FILTRE ARMA

Rappelons ici que s(n) est le signal brut, i.e. son percussif x(n) et bruit additif e(n) :

$$s(n)=x(n)+e(n)$$

Partons du fait que le signal x(n), correspondant à un modèle de Prony idéal, correspond également à la réponse impulsionnelle d'un filtre ARMA d'ordre [p,p] :

$$x(n)+\sum_{k=1}^{p} a_k \times x(n-k)=\sum_{k=0}^{p} b_k \times \delta(n-k) \quad \forall n\geq 0$$

In fine, le problème concret sera de déterminer les paramètres de Prony ($f_m,\alpha_m,a_m,\theta_m$) à partir de la connaissance des a_k et b_k.

Fréquences et taux d'amortissement

On peut montrer l'expression suivante :

$$x(n)+\sum_{k=1}^{p} a_k \times x(n-k)=0 \quad \forall n\geq p+1$$

On peut montrer que les Z_m sont les racines du polynôme formé par les a_k.

C'est à dire : $\prod_{m=1}^{p}(Z^{-1}-Z_m)=\sum_{k=0}^{p} a_k \times Z^{-k}$

La connaissance des a_k nous permet donc de calculer les Z_m, puis enfin les fréquences f_m et taux d'amortissement α_m :

$f_m = \tan^{-1}\left(\frac{\mathrm{Im}(Z_m)}{\mathrm{Re}(Z_m)} \times \frac{Fs}{2\pi}\right)$, et $\alpha_m = \ln(|Z_m|)$

Amplitudes et phases

$$x(n) + \sum_{k=1}^{p} a_k \times x(n-k) = \sum_{k=0}^{p} b_k \times \delta(n-k) \quad \forall n : 0 \le n \le p$$

On peut en dériver l'expression suivante :

$$b_k = \sum_{m=1}^{2L} \left(B_m \times Z_m^k + a_1 \times B_m \times Z_m^{k-1} + \cdots + a_{k-1} \times B_m \times Z_m + a_k \times B_m \right) \quad \forall k = 0, \ldots, p$$

On a ici un lien direct entre les paramètres des deux modèles. Les Z_m ayant été déterminés précédemment, nous connaissons ici tous les paramètres à l'exception des B_m. Ecrivons donc ce dernier système d'équations sous forme matricielle :

$$\Phi B = b$$

où $B = (B_1, \ldots, B_{2L})^t$, $b = (b_0, \ldots b_p)^t$, et

$$\Phi = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ (Z_1 + a_1) & (Z_2 + a_1) & & (Z_{2L} + a_1) \\ \vdots & & & \\ (Z_1^p + a_1 Z_1^{p-1} + \cdots + a_{p-1} Z_1 + a_p) & & & (Z_{2L}^p + \cdots + a_p) \end{pmatrix}$$

NB : $\dim(\Phi) = (p+1; 2L)$; rappelons nous ici que p=2L

La résolution de ce système d'équation se fait comme suit :

$$B = (\Phi^t \Phi)^{-1} \Phi^t b$$

Finalement, les amplitudes des sinusoïdes sont les modules des éléments de B, les phases en sont les arguments.

V.2. ANNEXE 2 : METHODES DE CLASSIFICATION

Cette annexe propose une introduction générale aux méthodes de classification des signaux. On pourra se rapporter à la lecture de [TOURNERET].

Il existe plusieurs type de classification : en *mode supervisé*, en *mode non-supervisé* et enfin la *classification structurelle*.

Pour le dernier type, il s'agit de déterminer des structures d'agencement de l'information (en chaînes ou en arbres), et de déterminer des coûts "d'éloignements" de ces structures.
La classification en mode supervisé suppose que l'on dispose d'une base d'apprentissage de données expertisées (nous connaissons les différentes classes et avons en notre possession des représentants de ces classes), la tâche est ici de classer un flot important de nouvelles données dans les classes définies par les données expertisées.
La classification en mode non-supervisé suppose qu'aucune base de données expertisées n'est disponible, la tâche est donc de créer les classes en même temps que l'on fait face au flot important de données, on fait évoluer le nombre et les caractéristiques des classes au fur et à mesure que l'on classe. Un exemple est la classification hiérarchique ascendante, un autre la classification par carte de KOHONEN.

Dans la classification en mode supervisé, pour faire face au flot important de données à classifier, des méthodes complexes telles que les réseaux de neurones sont envisageables, mais également des techniques plus simples telles que la règle de décision Bayesienne. Dans tous les cas, il faut d'abord déterminer les caractéristiques des classes à partir des données expertisées, c'est la phase de *prétraitement*.

V.3. ANNEXE 3 : RESUME ETENDU D'ARTICLE

Les trois pages suivantes constituent notre proposition d'article soumis à la conférence DAFX-00, ayant lieu à Verone, en Italie, les 7, 8 et 9 Décembre 2000.

cf. http://www.sci.univr.it/~dafx/

BIBLIOGRAPHIE

[ALLEN/DAN.] Paul Allen, Roger Dannenberg "Tracking musical beats in real time", Proceedings ICMC, 1990

[ANDRE-OBRECHT] Régine André-Obrecht : "Segmentation et parole", document d'habilitation à diriger des recherches IRISA 1993

[BASS./NIK.] Michèle Basseville, Igor Nikiforov : "Detection of abrupt changes: theory and applications", Prentice Hall 1993

[BROWN1] Judith Brown : "Determination of the meter of musical scores by autocorrelation", JASA 94(4), 1993

[BROWN2] Judith Brown : "Computer identification of musical instruments using a pattern recognition with cepstral coefficients as features", Journal of the Audio Engineering Society, 105(3), 1999

[CAR/CHANG] Douglas Carroll, Jih-Jie Chang : "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition", Psychometrika, vol.35(3), 1970

[CAR/WISH] Douglas Carroll, Myron Wish : "Models and methods for three-way multidimensional scaling", in Contemporary developments in mathematical psychology vol.II, 1974

[COOP/MEY] Grosvenor Cooper, Leonard Meyer : "The rhythmic structure of music", University of Chicago Press, 1960

[GABRIELSSON] Alf Gabrielsson : "Similarity ratings and dimension analyses of auditory rhythm patterns", Parts I & II, Scandanavian Journal of Psychology 14, 1973

[GOTO/MUR.1] Masataka Goto, Yoichi Muraoka : "A real-time beat-tracking system for audio signals", Proceedings ICMC, 1995

[GOTO/MUR.2] Masataka Goto, Yoichi Muraoka : "Real-time rhythm-tracking for drumless audio signals – chord change detection for musical decision", IJCAI Workshop on computational auditory scene analysis, 1997

[GOUYON] Fabien Gouyon : "Detection and modeling of transient regions in musical signals", rapport de stage DEA SIC ENSEEIHT, rapport interne CCRMA Stanford University, 1999

[GREY] JM Grey : "Multidimensional perceptual scaling of musical timbres", Journal of the Acoustical Society of America, 61, 1977

[HERRERA] Perfecto Herrera, Xavier Serra, Geoffroy Peeters : "Audio descriptors and descriptors schemes in the context of MPEG7", Proceedings ICMC, 1999

[LAROCHE] Jean Laroche "Etude d'un système d'analyse et de synthèse utilisant la méthode de Prony – Applications aux instruments de musique de type percussif" Thèse ENST, 1989

[LEVINE] Scott Levine : "Audio representation for data compression and compressed domain processing", PhD thesis CCRMA Stanford University, 1998

[KAY] Kay : "Modern spectral estimation", Prentice Hall

[KEDEM] Benjamin Kedem : "Spectral analysis and discrimination by zero-crossings", Proceedings IEEE 74(11), 1986

[KRU/WISH] Joseph Kruskal, Myron Wish : "Multidimensional scaling", Bell Laboratories, Sage publications, 1978

[KUM/TUFTS] Ramdas Kumaresan, Donald Tufts : "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise", Proceedings IEEE, 1982

[SCHEIRER] Eric Scheirer : "Tempo and beat analysis of acoustic signals", JASA 103(1), 1998

[SCHEIR/SLAN] Eric Scheirer, Malcolm Slaney : "Construction and evaluation of a robust multifeature speech/music discriminator", Proceedings IEEE ICASSP 1997

[SCHLOSS] Andrew Schloss : "On the automatic transcription of percussive music – From acoustic signals to high-level analysis", CCRMA internal report, Stanford University 1985

[SERRA] Xavier Serra : "A system for sound Analysis / Transformation / Synthesis based on a deterministic plus stochastic decomposition", PhD thesis CCRMA Stanford University, 1989

[TOURNERET] Jean-Yves Tourneret "Classification et reconnaissance des formes", ed. ENSEEIHT (Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique et d'Hydraulique de Toulouse) 1997

[YOUNG] Forrest Young : "Multidimensional scaling: History, theory and applications", ed. Lawrence Erlbaum Associates, 1987