Année universitaire 2005-2006
MASTER SIC - Systèmes Intelligents et Communicants
2ème année

Extraction automatique d'une suite d'accords à partir de l'analyse d'un signal audio musical

Hélène Papadopoulos

Encadrant :
Geoffroy Peeters
Responsable :
Xavier Rodet

IRCAM (Institut de Recherche et Coordination en Acoustique/Musique)
1, place Igor Stavinsky
75004 Paris

Je voudrais ici remercier M. Geoffroy Peeters pour avoir consacré une partie de son temps à l'encadrement de mon travail. Il a su orienter mes travaux tout en me laissant la liberté nécessaire à l'acquisition d'une saine autonomie.

Je voudrais aussi remercier M. Xavier Rodet pour l'intérêt qu'il a porté à mon travail.

Merci aussi à Niels pour son attention et sa gentillesse.

Résumé : Nous présentons ici un système qui permet de d'extraire de manière automatique la suite d'accords d'un signal audio musical polyphonique complexe. La première partie du système effectue l'extraction de vecteurs représentant l'importance des différentes hauteurs à un instant donné (vecteur de chroma). Ces informations sont de nature probabiliste ; l'objectif est d'estimer la suite d'accords au cours du temps la plus probable. Pour cela nous développons deux méthodes différentes, toutes deux basées sur des Modèles de Markov Cachés. A travers ces méthodes, nous comparons l'influence de différentes hypothèses faites sur le signal. Le système est évalué à partir d'un ensemble d'extraits musicaux issus de la musique populaire.

Ce document est le compte-rendu du stage que j'ai effectué à l'IRCAM du 20 mars au 20 juillet 2006. Il entre à la fois dans le cadre de mes études d'ingénieur à l'ENSEA, où il fait l'objet de mon travail de fin d'études, et dans le cadre du master recherche SIC (Systèmes Intelligents et Communicants) de l'université de Cergy-Pontoise.
Mais c'est avant tout pour moi le début d'un travail de reherche qui se poursuivra par un doctorat à l'IRCAM.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Ce stage se place dans le contexte de Music Information Retrieval (Recherche en musique) et de la transcription automatique de morceaux de musique. L'objectif est d'estimer à partir d'un signal audio, la suite d'accords composant un morceau de musique.

L'estimation de la suite d'accords est un sujet qui se rattache au domaine de l'indexation musicale. L'extraction de descripteurs audio consiste à trouver des modèles mathématiques qui décrivent les propriétés du son en utilisant les outils du traitement du signal. L'obtention de ces paramètres permet de répondre efficacement à une demande qui ne cesse d'augmenter : les services de distribution de musique en ligne prolifèrent aujourd'hui. Des applications telles que la recherche dans une base de données, ou le traitement du signal par son contenu (par exemple trouver un certain thème dans une grande base de données) peuvent alors être développées.

L'un des paramètres que l'on peut extraire du signal audio musical est la suite d'accords composant le morceau. La succession des accords dans le temps est le coeur de l'harmonie d'une pièce de musique. Annoter manuellement un morceau de musique (transcrire les accords du morceau de manière individuelle) est un travail beaucoup trop fastidieux étant donnée l'ampleur des bases de données dont on dispose, c'est pourquoi il est nécessaire de développer des technologies de transcription automatique. Outre les applications citées précédemment, cela peut être également la base d'applications telles que la segmentation musicale, l'identification de similarités musicales, etc.

La suite de ce rapport sera organisée de la manière suivante : dans une première partie, nous rappelerons brièvement quelques notions de théorie musicale nécessaires à la compréhension de ce rapport ; nous présenterons ensuite l'état de l'art ; les sections 3, 4 et 5 seront consacrées à la desription du travail réalisé pendant le stage ; enfin la dernière partie présentera et analysera les résultats obtenus.

2 Pré-requis musicaux à l'attention des scientifiques non musiciens

Ainsi que le sujet de ce stage l'indique, il s'agit de travailler sur des signaux audio musicaux. Il est donc impossible d'en présenter le contenu sans se référer à des notions et des termes empruntés au langage musical théorique. C'est pourquoi nous commencerons par aborder les quelques notions générales de la théorie de la musique sans lesquelles la suite de ce rapport ne saurait être comprise par un non-musicien.

2.1 Qu'est ce que le son ?

Le son est une vibration mécanique qui se propage dans l'air. Un son musical peut être caractérisé par trois grandeurs : sa hauteur, son intensité et son timbre. Ces trois critères correspondent respectivement à trois caractéristiques de l'onde qui sont sa fréquence, son amplitude et sa constitution harmonique. L'oreille humaine ne perçoit que les sons dont la fréquence est comprise entre 20 et 20 000 Hz .

- La *hauteur* (*pitch*) d'un son est l'une de ses caractéristiques principales. C'est une notion subjective. La hauteur perçue par l'oreille est liée à la fréquence du son. À une fréquence faible correspond un son grave, à une fréquence élevée un son aigu.

- L'*amplitude* est une autre caractéristique importante du son. C'est en particulier de l'amplitude du son que correspond la force perçue. Elle représente une mesure du déplacement des molécules d'air. Plus les molécules d'air frappent avec force la membrane de l'oreille, plus l'amplitude de l'onde est grande et donc plus le son paraît fort.

- Le *timbre* est le terme utilisé en musique pour définir la qualité d'émission d'un son spécifique à un instrument donné ou à la voix. Le timbre d'un son dépend de nombreux facteurs. Le ou les matériaux qui constituent l'instrument (bois, cuivres, cordes) donnent une empreinte particulière au timbre. Deux sons de même hauteur et de même intensité sont différents selon que les vibrations sont émises par frottement ou par soufflement. Le timbre dépend aussi du nombre d'harmoniques et de leurs amplitudes. Le timbre d'un même instrument sonne différemment au moment de l'attaque d'un son, de sa durée, ou de son extinction. L'air ambiant tient également une place dans la spécificité du timbre.

  Le timbre est donc caractéristique d'un instrument de musique. Des sons de même hauteur émis par deux instruments distincts ont un timbre différent qui permet de les distinguer. Le timbre dépend de la composition du son en harmoniques. Un diapason émet une vibration sinusoïdale sans harmoniques, c'est un son pur. Une vibration sonore associée à une note a une amplitude qui varie au cours du temps. L'*enveloppe* est la courbe reliant les maxima des amplitudes du son au cours du temps.

2.2 Notes et intervalles

2.2.1 Notes

Avant de poursuivre, nous rappelons quelques notions sur les notes, qui sont les signes employés pour écrire la musique. Les notes représentent des durées et des hauteurs de son. Les méthodes de division de l'octave en intervalles ont depuis très longtemps donné naissance à la gamme dite heptatonique, c'est-à-dire comprenant sept notes dans un intervalle d'octave. Ces notes ont été nommées, dans le sens as-

cendant :

français	*do*	*re*	*mi*	*fa*	*sol*	*la*	*si*	*do*
anglo-saxon	*C*	*D*	*E*	*F*	*G*	*A*	*B*	*C*

Nous utiliserons par la suite la terminologie anglo-saxonne.

2.2.2 Intervalles

L'oreille identifie des « intervalles ». Un intervalle est une grandeur additive que nous percevons comme une différence de « hauteur », quand la physique identifie des rapports de fréquences.

On appelle *octave* l'intervalle entre deux sons dont l'un est à la fréquence f et l'autre à la fréquence $2f$. L'octave correspond à l'intervalle qui sépare la fréquence fondamentale de la première harmonique.

2.3 Son fondamental et harmoniques

Une vibration sonore est une fonction mathématique périodique. Elle peut donc être décomposée en série de Fourier, c'est à dire en une somme de fonctions sinusoïdales élémentaires. Les sons musicaux sont formés d'un son fondamental et des harmoniques appelées aussi partiels, dont les rapports de fréquence avec la fondamentale sont des quotients de nombres entiers. La hauteur d'un son est mesurée par la fréquence du fondamental.

Lorsque l'on entend un C, on entend aussi la première harmonique qui est le C de l'octave supérieure. Une note de la gamme est ainsi déterminée modulo la multiplication par une puissance de 2 qui détermine l'octave où elle se trouve.

Par exemple l'échelle des A, en Hz est la suivante :

Fréquence (Hz)	55	110	220	440	880	1760
note	*A*1	*A*2	*A*3	*A*4	*A*5	*A*6

Lorsque l'on entend un C de fréquence f, on entend aussi les harmoniques de fréquences 2f, 3f ... :

Fréquence (Hz)	f	$2f$	$3f$	$4f$	$5f$	$6f$
note	*C*	*C*	*G*	*C*	*E*	*G*

2.4 Gamme et tonalité

Une gamme est une série de sons conjoints. La gamme tempérée est de nos jours utilisée de façon presque universelle dans la musique occidentale. Elle est obtenue en divisant l'octave en douze intervalles égaux.

`note`	C	D	E	F	G	A	B	C
`fréquence`	f	a^2f	a^4f	a^5f	a^7f	a^9f	a^11f	$2f$

avec $a = 2^{1/12}$.

On appelle *demi-ton* l'intervalle défini par $(f, 2^{1/12}f)$. Un *ton* peut se diviser en 2 demi-tons.

Pour former une gamme, on utilise sept notes de noms différents. Chacune ayant un rôle déterminé, on lui donne le nom de degré que l'on écrit en chiffres romains. Chaque degré a un nom particulier qui caractérise la position qu'il occupe dans la gamme.

Nous utiliserons par la suite le terme *tonique*, qui correspond au premier degré, ainsi que les termes *médiante*, *sous-dominante*, *dominante* et *octave* qui correspondent respectivement aux 3^{eme}, 4^{eme}, 5^{eme} et 8^{eme} degrés.

2.5 Mode majeur, mode mineur

Deux modes peuvent être distingués dans la musique occidentale : le mode majeur et le mode mineur. Chacun de ces deux modes présente des caractéristiques particulières qui sont liées à la position des tons et demi-tons dans les gammes qui leurs sont associées. Une gamme est composée d'une séquence de notes. Chaque couple de notes forme un certain intervalle. Un intervalle est défini par le rapport entre les fréquences de deux notes f1 et f2. Pour une gamme tempérée, un demi-ton (st) est toujours défini par un rapport de fréquence de $f2/f1 = 2^{1/12}$. Un intervalle de n demi-tons est défini par un rapport de fréquences de $f2/f1 = 2^{n/12}$.

Nous pouvons associer à chaque tonique un mode majeur et un mode mineur. La *tonalité* est l'ensemble des lois qui régissent la constitution des gammes. Il existe donc 24 tonalités (12 majeures et 12 mineures) auxquelles on peut associer 24 accords de trois notes composés de la tonique, la médiante et la dominante (on les nomme accords parfaits). On ne fait pas ici de distinction entre les notes enharmoniques, c'est à dire entre les notes qui sonnent de la même manière mais qui ont un nom différent, par exemple C# et Db.

2.6 Accords

On nomme accord tout ensemble de sons entendus simultanément pouvant donner lieu à une perception globale identifiable. Pour former un accord, il faut au moins l'émission simultanée de trois sons. On distingue les accords de 3 sons (majeurs, mineurs, diminués et augmentés), les accords de 4 sons appelés accords de septième et les accords de 5 sons appelés accords de neuvième.

La connaissance de la composition de ces différents accords (intervalles dont ils sont formés, propriétés ...) était nécessaire pour bien comprendre le sujet du stage et donner des réponses aux problèmes posés. Cependant, elle n'est pas indispensable à

la compréhension de ce rapport, c'est pourquoi nous ne nous étendons pas davantage sur ce point.

2.7 Classes de hauteur ou pitch class

La perception de la hauteur d'un son par l'oreille humaine est périodique. Les hauteurs séparées par un nombre entier d'octaves sont perçues commme "sonnant" de manière équivalente, c'est pourquoi on leur donne le même nom. On dit qu'elles partagent le même *chroma*. L'ensemble des notes partageant un même chroma est appelé *classe de hauteurs* ou *pitch class*.

Les théoriciens de la musique se réfèrent en général aux différentes classes de hauteur en utilisant des nombres. On peut transformer la fréquence fondamentale f d'un son en un nombre réel p selon l'équation suivante(conversion en échelle midi) :

$$p = 69 + 12log_2\frac{f}{440} \tag{1}$$

Le C4 correspond à la note midi 60.

Dans l'espace des classes de hauteur, il n'y a pas de distinction entre les notes qui sont séparées par un nombre entier d'octaves. (p, $p + 12$, $p + 2 * 12$...). Par exemple C4, C5, C6 appartiennent à la même classe de hauteur.

3 État de l'art

La reconnaissance automatique d'accords musicaux à partir d'un signal audio complexe contenant des sons vocaux et percusifs reste à l'heure actuelle un problème qui n'est pas encore résolu.

Des travaux récents sur le sujet de la détection automatique de suites d'accords ont montré que l'on pouvait obtenir des résultats intéressants sur des morceaux polyphoniques complexes sans avoir à passer par une transcription symbolique. L'approche traditionnelle consistait à extraire les notes individuelles présentes à chaque instant dans le signal audio pour en déduire les accords en s'appuyant sur des règles musicales. L'inconvénient de cette méthode est qu'il est difficile d'extraire les notes, d'une part en raison du bruit, d'autre part en raison des harmoniques des différentes notes présentes dans le spectre du signal audio qui se mélangent et se recouvrent. Les algorithmes existants sont peu fiables et les résultats obtenus par cette méthode sont insuffisants pour pouvoir espérer arriver à une transcription automatique.

En 1999, Fujishima introduit la notion de Pitch Class Profiles (PCPs) [1] pour la reconnaissance d'accords. S'inspirant des travaux de Fujishima et de Barsh et Wakefield (2001) [2], Sheh et Ellis obtiennent en 2003 [3] des résultats encourageants pour l'estimation d'accords sans passer par une transcription symbolique. Ils proposent une représenation du signal audio en termes de "Pitch Class Profiles" ainsi que l'utilisation de HMM. Ces méthodes sont reprises par Harte et Sandler [4] puis Bello et Pickens [5] en 2005. Ces derniers obtiennent des résultats satisfaisants sur

des signaux audio polyphoniques complexes.

Les systèmes existants ont le défaut de ne pas tenir compte des harmoniques présentes dans le spectre. Ce problème a déja été soulevé en particulier dans le cas de l'estimation de tonalité, sujet proche de celui de l'extraction de suites d'accords. Deux sortes de solutions peuvent être adoptées : soit on peut retirer les harmoniques présentes dans le spectre ([6], [7]), soit les prendre en compte dans la création des Pitch Class Profiles ([8]).

Pendant ce stage, nous avons repris dans un premier temps les différentes méthodes présentées ci-dessus, que nous avons implantées puis testées sur une base de données composée de deux des premiers albums des Beatles , *Please Please Me* et *Beatles for Sale*. Le choix de ces signaux audio se justifie d'une part par le fait que l'on dispose de transcriptions symboliques précises qui permettent de comparer les résultats obtenus avec les résultats théoriques et d'autre part que, depuis les travaux de Sheh et Ellis sur l'estimation de suites d'accords, l'évaluation des systèmes a été faite sur ces morceaux, ce qui nous permet de comparer nos résultats avec ceux obtenus auparavant.

Nous avons utilisé deux approches pour estimer les suites d'accords. L'une est basée sur la corrélation d'observations avec des modèles théoriques, l'autre est similaire à celle proposée par Bello et Pickens dans [5]. Nous les présentons par la suite et comparons leurs performances.

4 Système initial

Rappelons que l'objectif est d'estimer à partir d'un signal audio, la suite d'accords composant un morceau de musique. La première partie du système effectue l'extraction d'un vecteur représentant l'importance des différentes hauteurs à un instant donné (vecteur de chroma). Ces informations sont de nature probabiliste. A partir de ces observations, l'objectif sera d'estimer la suite d'accords au cours du temps la plus probable. De manière équivalente à la reconnaissance de parole, l'adjonction d'un dictionnaire de grammaire musicale (estimation des probabilités de transition entre accords) permettra de formuler l'estimation sous forme d'une chaîne de Markov cachée.

L'estimation de suites d'accords d'un signal audio commence par une phase d'analyse du signal. Dans la plupart des travaux mentionnés dans l'état de l'art, les techniques utilisées pour répondre au problème partent d'une même base, bien que des variations apparaissent dans la phase d'implantation.

Nous avons dans un premier temps repris les étapes d'analyse communes à tous les systèmes existants. Elles ont pour but d'obtenir des vecteurs d'observation contenant les caractéristiques du signal audio. Elles consistent d'abord à transformer le signal temporel subdivisé en trames dans le domaine fréquentiel. Le spectre est

ensuite transormé dans le domaine des chromas ce qui permet d'obtenir des vecteurs d'observation du signal à travers le temps. Ces observations sont utilisées pour construire la succession des accords au cours du temps, en formulant l'estimation sous forme d'une chaîne de Markov cachée.

Nous présenterons deux méthodes différentes. Leur différence réside dans la manière de calculer les probabilités d'état du système (soit par comparaison instantannée avec un modèle théorique, soit en modélisant la distribution des observations par une gaussienne).

4.1 Observation et traitement du signal

4.1.1 Transformation du signal temporel dans le domaine fréquentiel

Le signal audio est échantillonné à $11025Hz$. S'il s'agit d'un signal au format stéréophonique, il est converti en signal monophonique en prenant la moyenne sur les deux canaux. Il est ensuite divisé en trames se recouvrant partiellement.

Dans un premier temps, nous avons fixé, ainsi que dans [4], [5], la taille des trames à $N = 8212$ points (ce qui correspond à une taille de fenêtre d'analyse de $0.743s$) et le taux de recouvrement à 7/8. Le signal est alors transformé dans le domaine fréquentiel avec une transformée de Fourier discrète ([1], [3], [9]) ou une *constant Q transform* ([5], [4], [10])

4.1.2 Construction du chromagram

La notion de chroma a été introduite par le psychologiste Roger Shepard dans les années 1960 [11]. Les chromas transforment les fréquences en classes d'équivalence d'octaves. Shepard a montré que deux dimensions sont nécessaires pour bien représenter la perception du système auditif humain. Celle-ci peut être reprensentée par une hélice. (Voir figure 1).

La hauteur d'une note (pitch (p)) en Hz peut être décrite par deux valeurs : le chroma (ou Pitch Class) (c) et la hauteur de ton (ou Pitch heigh) (h) .

$$p = 2^{c+h} \tag{2}$$

Le chromagram (ou spectre de chroma) est une extension de la notion de chroma qui inclut la dimension temporelle. Il peut être utilisé pour représenter les propriétés de la distribution du spectre d'énergie du signal à travers les fréquences et le temps. Il s'agit d'une représentation compacte de la représentation spectrale (FFT ou CQ) du signal audio.

Le chromagram s'obtient en effectuant un mapping entre cette représentation spectrale et un vecteur à 12 dimensions représentant les 12 demi-tons de la gamme

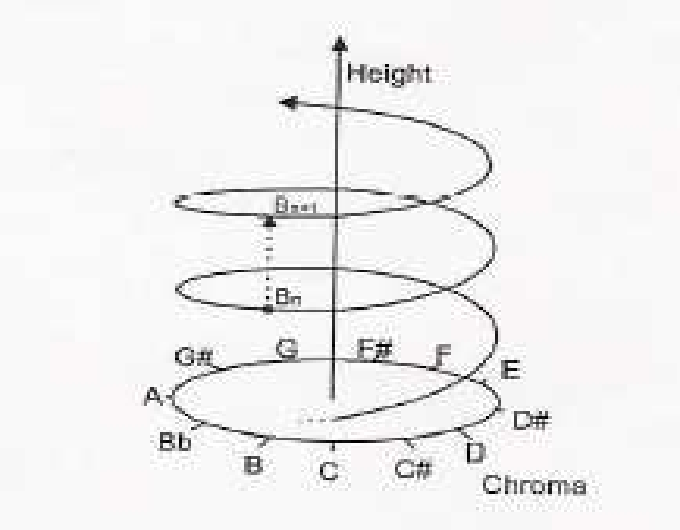

FIG. 1 – Représentation de la perception de la hauteur par l'oreille humaine. B_{n+1} est une octave au-dessus de B_n. Extrait de [4]

chromatique. La représentation sous forme de chromas est très utilisée dans les travaux relatifs à l'estimation automatique de tonalité, d'harmonie ou d'accords. ([12], [6], [13], [7], [4]...). Le calcul du chromagram est basé sur les Pitch Class Profile (PCP) introduits par Fujishima en 1999. La procédure est la suivante :

– Mapping des valeurs du spectre aux 12 demi-tons des pitch-classes. Pour chaque fréquence f_k on a :

$$c(f_k) = mod(12log_2(\frac{f_k}{261.62}), 12) \tag{3}$$

où $c(f_k)$ est la valeur associée à f_k sur l'échelle des chromas et 261.61 correspond à la fréquence d'un C4.

– Calcul du vecteur de chroma à 12 dimensions en additionnant les intensités de la transformée de Fourier des fréquences de même valeur $c(f_k)$:

$$\text{Pour } l = 1, \ldots, 12 \quad C(l) = \sum_{f_k \text{telle que} c(f_k)=l} A(f_k) \tag{4}$$

Ici, le calcul du chromagram a été fait selon la méthode proposée par Peeters dans [9].

Dans un premier temps, les fréquences f_k du spectre sont converties en notes midi correspondant aux hauteurs des demi-tons :

$$n(f_k) = 69 + 12log_2(\frac{f_k}{440}) \tag{5}$$

Le spectre est ensuite fractionné en régions centrées sur les fréquences n' correspondant aux demi-tons appartenant à l'intervalle de fréquences considérées. Nous

avons d'abord considéré l'intervalle $[98Hz; 5250Hz]$ ce qui correpond aux notes allant du G2 au E8, ou encore aux notes midi de l'intervalle $[43; 112]$.

Nous construisons alors ce qui est nommé "spectre de demi-tons" dans [9]. Pour cela, le spectre du signal est multiplié avec un ensemble de filtres centrés sur les fréquences $n' = 43 + \frac{1}{R}, 43 + \frac{2}{R}, \ldots, 112$ où R est un facteur qui définit la résolution du chromagram. Autrement dit, R correspond au nombre de filtres représentant un chroma. Par exemple pour $R = 3$, un filtre représente un tiers de demi-ton. Le filtre $H_{n'}$ centré sur $n' \in [43 + \frac{1}{R}, 43 + \frac{2}{R}, \ldots, 112]$ est défini par :

$$H_{n'} = \frac{1}{2}\tanh(\pi(1 - 2x)) + \frac{1}{2} \tag{6}$$

où x est la distance relative entre le centre du filtre n' et les fréquences de la tansformée de Fourier : $x = R|n' - n(f_k)|$.

Les valeurs N(n') du spectre de demi-tons sont obtenues par :

$$N(n') = \sum_{f_k} H_{n'}(f_k)A(f_k) \tag{7}$$

où les $A(f_k)$ sont les valeurs de la transformée de Fourier.

Le mapping entre les valeurs n du spectre de demi-tons et les chromas c est défini par : $c(n) = mod(n, 12)$.

Les vecteurs de chroma $C(l)$ à 12 dimensions sont alors obtenus en cumulant les valeurs du spectre de demi-tons correspondant à un même chroma :

$$C(l) = \sum_{n' \ tel \ que \ c(n') = \ l} N(n') \quad avec \quad l \in [0, 12[\tag{8}$$

Nous obtenons alors un spectre de chroma ou chromagram. Pour chaque trame, nous avons calculé un vecteur de chroma à 12 dimensions qui correspondent aux 12 notes de la gamme $C, C\#, \ldots, B$. Ces vecteurs de chroma sont nos observations. Ils sont caractéristiques du signal audio analysé.

La figure 2 correspond à un morceau du chromagram obtenu à partir du morceau *I am a Loser* de l'album *Beatles For Sale* des Beatles. Les régions les plus sombres corrrespondent aux intensités les plus importantes. Nous pouvons distinguer ici la suite d'accords A majeur(A C# E), A# majeur(A# D F), D majeur(D F# A), G majeur(G B D), D majeur, G majeur, D majeur, G majeur.

4.2 Estimation de suite d'accords basée sur des HMM

De manière équivalente à la reconnaissance de parole, l'adjonction d'un dictionnaire de grammaire musicale (estimation des probabilités de transition entre accords)

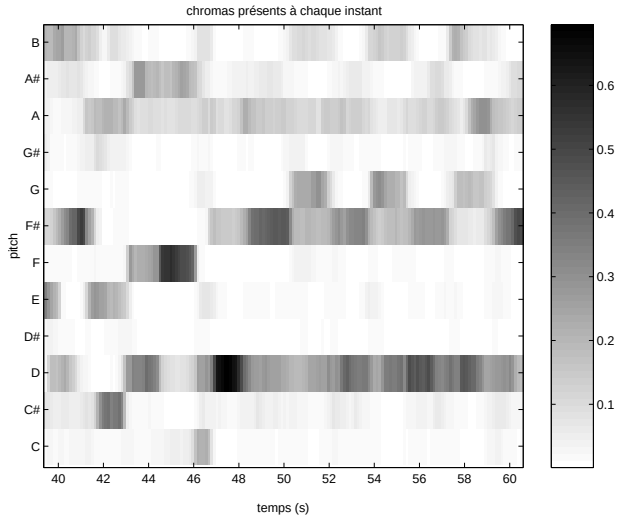

FIG. 2 – Exemple de chromagram extrait de *I am a Loser* de l'album *Beatles For Sale*.

permettra de formuler l'estimation sous forme d'une Chaîne de Markov Cachée.

L'utilisation de Modèle de Markov Caché (Hidden Markov Model, HMM) pour estimer automatiquement les accords, la structure, l'harmonie ou la tonalité d'un morceau est assez courante ([3], [5], [9], [7]). Pickens et Bello en particulier obtiennent de bons résultats pour l'estimation de suites d'accords en utilisant cette méthode. Les vecteurs de chroma sont utilisés pour entraîner un HMM avec un état pour chaque accord pouvant être distingué par le modèle.

4.2.1 Rappels théoriques sur les HMM

Les modèles de Markov cachés (Hidden Marov Models, HMM) ont été introduits par Baum et ses collaborateurs dans les années 1960-70. D'abord utilisés en reconnaissance de la parole à partir des années 80 (Rabiner, [14], Gold et Morgan [15]), ils ont ensuite été appliqués à la reconnaissance de textes manuscrits et à la bioinformatique.

Un HMM est un automate probabiliste, c'est à dire une structure composée d'états et de transitions, et d'un ensemble de distributions de probabilités de transition. Chaque état génère une observation. Les états de transition suivent la propriété markovienne qu'étant donné un état présent, le futur est indépendant du passé.

Définition : Un processus de Markov est un processus à temps discret se trouvant à chaque instant dans un état parmi N états distincts. Les transitions entre les états se produisent entre deux instants discrets consécutifs selon une certaine loi de probabilité. La probabilité de chaque état ne dépend que de l'état qui le précède immédiatement.

Un modèle de Markov caché représente de la même façon qu'une chaîne de Markov un ensemble de séquences d'observation dont l'état de chaque observation n'est pas observé mais associé à une fonction densité de probabilité (pdf). Il s'agit donc d'un processus doublement stochastique, dans lequel les observations sont une fonction aléatoire de l'état et dont l'état change à chaque instant en fonction des probabilités de transition issues de l'état antérieur.

On distingue trois problèmes fondamentaux qui concernent les HMM :

- Évaluation de la probabilité d'une séquence d'observations étant donné un modèle d'HMM.
- Étant donné un modèle, comment déterminer la séquence d'états optimale qui a donné naissance à la séquence d'observations ?
- Étant donnée une séquence d'observations, comment ajuster les paramètres du modèle pour avoir la meilleure explication ?

Dans les problèmes de reconnaissance, on s'intéresse en particulier à la séquence d'états qui a donné une séquence d'observations.

Terminologie :

Dans la suite, nous noterons :

- N le nombre d'états du modèle.
- q_t état à l'instant t et $Q = (q_1, \ldots, q_T)$ l'ensemble des états.
- o_t observation à l'instant t et $O = (o_1, \ldots, o_T)$ l'ensemble des observations.
- $a_{ij} = P[q_{t+1=j|q_t=i}], 1 \leqslant i, j \leqslant N$ la probabilité de transition de l'état i à l'état j. Ce sont les probabilités de passer d'un état i à un état j $P(q_i|q_j)$. Elles sont stockées dans une matrice de transition A dont les coefficients sont les $a_{ij} = P(q_i|q_j)$.
- $B = b_j(k)$ avec $b_j(k) = P[o_t = v_k|q_t = i], 1 \leqslant k \leqslant M$ la distribution des symboles observés.
- $\pi = \pi_i$ avec $\pi_i = P[q_1 = i], 1 \leqslant k \leqslant N$ la distribution initiale.

Pour retrouver la séquence d'états qui a donné naissance aux observations, un critère d'optimalité consiste à choisir la séquence d'états (ou chemin) qui apporte un maximum de vraisemblance en respectant le modèle donné. La séquence d'états peut être déterminée à partir de l'algorithme de Viterbi. Celui-ci nous donne deux

résultats importants :

- la sélection parmi tous les chemins possibles dans un modèle considéré, du chemin qui correspond à la séquence d'états la plus probable au sens de la probabilité de vraisemblance de la séquence d'observations X.

- la probabilité de vraisemblance sur le meilleur chemin.

On pourra trouver une description détaillée de l'algorithme de Viterbi dans [14].

On s'intéresse par la suite au problème de trouver une séquence d'états optimale (qui explique le mieux les observations) étant donnée une séquence d'observation $O = (o_1, \dot{;} o_T)$ et le modèle $\lambda = \pi, A, B$.

4.3 Utilisation des HMM pour la détection de suites d'accords, modélisation gaussienne

L'une des deux méthodes que nous avons utilisées pour estimer de manière automatique la suite d'accords de nos signaux audio consiste à modéliser la distribution des observations par une gaussienne à 24 états. Nous nous refèrerons à cette méthode en parlant de « méthode par modélisation gaussienne des observations ».

4.3.1 Étiquettage des accords

Nos vecteurs d'observation sont de dimension 12. Nous voulons associer un label à chacune des observations. Pour cela, on définit un dictionnaire d'accords contenant des modèles d'accords théoriques auxquels vont être comparées les observations. Le dictionnaire utilisé est formé des 24 accords de trois sons parfaits majeurs et parfaits mineurs. Par simplicité, nous avons choisi d'utiliser un dictionnaire de taille limitée ainsi que dans [5]. Dans [3] ou [4], d'autres types d'accords sont ajoutés à ce dictionnaire (diminués et augentés, 7^{eme} ou 9^{eme} par exemple). Nous avons donc un système à $N = 24$ états.

4.3.2 Distribution des états initiale

A priori nous n'avons pas de raison de préférer un état par rapport à un autre (le morceau est susceptible *a priori* de commencer par n'importe quel accord). Cest pourquoi on choisit une distribution initiale π uniforme : $\frac{1}{24}$ pour chacun des 24 états.

4.3.3 Matrice de transition $[A]$

Dans leurs travaux, Sheh et Ellis utilisent une initialisation aléatoire de la matrice de transition. Bello et Pickens ont montré que l'introduction de connaissance musicale permet d'améliorer les résutats de façon significative. En effet, on ne peut

savoir si par exemple un accord de C majeur sera suivi par E majeur ou A mineur. Par contre, les règles musicales permettent d'affirmer que ces hypothèses sont plus plausibles que $F\#$ majeur.

Dans la plupart des styles de musique, en particulier dans la musique populaire, les changements d'harmonie se font selon des règles bien établies. De manière analogue à [5], nous intégrons cette notion en initialisant la matrice de transition avec le double cycle des quintes (voir figure 3). En musique, on appelle cycle des quintes, une succession, ascendante ou descendante, de notes séparées par des intervalles de quinte juste.

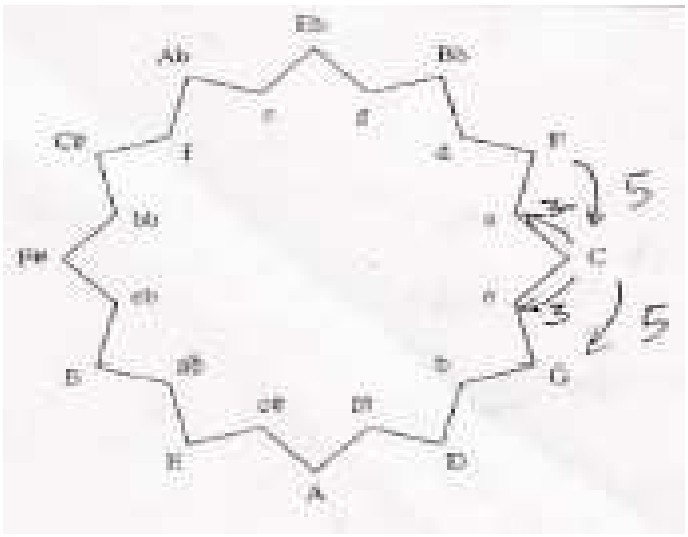

FIG. 3 – Figure du double cycle des quintes.Extrait de [5]

Nous donnons à la transition de C → C une probabilité de $\frac{12+\epsilon}{144+24\epsilon}$, où ϵ est une constante d'atténuation très petite. La probabilité de transition C → E vaut $\frac{11+\epsilon}{144+24\epsilon}$, et de même jusqu'à la probabilité de transtion C → F# qui vaut $\frac{0+\epsilon}{144+24\epsilon}$. À partir de ce point, les probabilités de transition augmentent à nouveau de $C \rightarrow B^b = \frac{1+\epsilon}{144+24\epsilon}$ à $C \rightarrow A = \frac{11+\epsilon}{144+24\epsilon}$.

Pour chaque état on calcule de même les probabilités de transition avec chacun des 23 autres états en donnant la valeur la plus élevée à la probabilité de transition d'un état avec lui-même puis en donnant un poids aux transtions entre cet état et les autres états en fonction de leur distance sur le cercle des quintes. (Voir figure 4 pour la marice de transition).

4.3.4 Distribution des observations $[B]$

On suppose la fonction de distribution des observations continue. Pour la modéliser, on utilise une gaussienne à 24 états corespondant au 24 accords mineurs et majeurs, chacun décrit par un vecteur moyen μ et une matrice de covariance Σ.

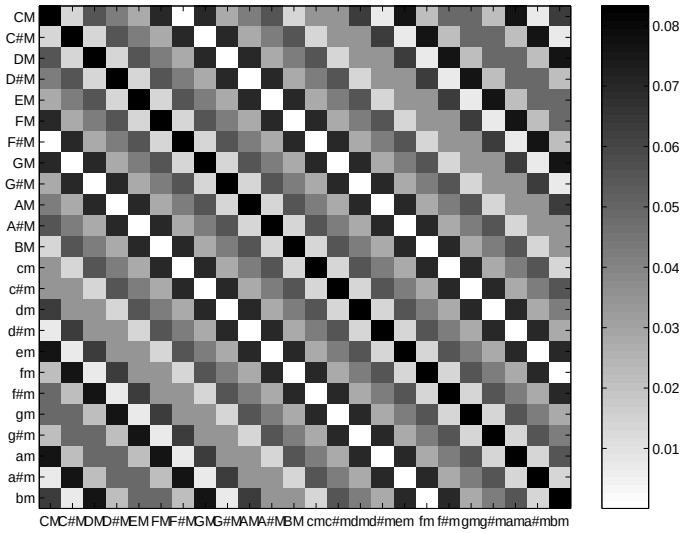

FIG. 4 – Matrice de transition des états

Dans [3], l'initialisation de μ et de Σ est aléatoire. Bello et Pickens ont montré dans [5] que l'introduction de connaissance musicale dans la distribution des observations permet également d'améliorer l'estimation des accords.

Les vecteurs moyens, qui représentent les accords théoriques doivent donc refléter cette connaissance musicale. Ainsi, de même que dans [5], pour l'état de C majeur, on initialise μ à 100010010000 : 1 pour les dimensions correspondant aux notes présentes dans l'accord C, E et G, 0 ailleurs. On initialise de la même manière les vecteurs moyens des 23 autres états, par permutation circulaire. (Voir la figure 5).

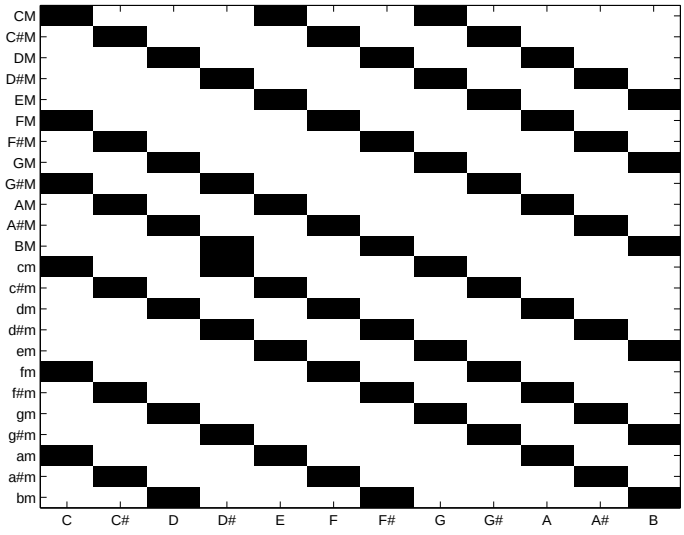

FIG. 5 – Matrice des observations théorique

La covariance est intuitivement une mesure de la variation simultanée de deux variables aléatoires. La covariance devient plus positive pour chaque couple de va-

leurs qui diffèrent de leur moyenne dans le même sens et plus négative pour chaque couple de valeurs qui diffèrent de leur moyenne dans le sens opposé. Les notes qui appartiennent à un même accord sont liées entre elles et l'on va décrire dans notre matrice de covariance comment ces notes varient simultanément.

La théorie musicale et les résultats obtenus empiriquement par Krumhansl dans [16] ont montré que la dominante est plus importante que la médiante dans la caractérisation d'un accord de trois sons. Nous avons dans un premier temps repris les valeurs fixées empiriquement dans [5].

Pour l'état de C majeur, les notes C, E et G sont fortement corrélées avec elles-même, on fixe donc à 1 la variance de ces notes. En accord avec les résultats obtenus par Krumhansl, on fixe à 0.8 la valeur de la covariance de la tonique avec la dominante ainsi que de la médiante avec la dominante et à 0.6 celle de la covariance de la tonique avec la médiante. Les valeurs de la diagonale autres que celles corespondant aux notes de l'accord sont fixées à 0.2 afin d'assurer que la matrice est semi-définie, positive. Les autres valeurs de la matrice de covariance sont fixées à 0 pour indiquer que l'on considère qu'il y a indépendance entre les notes de l'accord et celles qui ne lui appartiennent pas. Les 11 autres matrices de covariance du mode majeur se déduisent de celle correspondant à C majeur par permutation circulaire.

Les matrices de covariance du mode mineur sont construites de la même manière à partir de la matrice de covariance de C mineur. Celle-ci diffère de C majeur par la tierce : on fixe à 1 la variance de C, D# et G et non plus C, E, G. (Voir les figures 6 et 7).

4.4 Estimation de suite d'accords par corrélation

Cette méthode s'inspire de celle proposée par Harte et Sandler pour l'estimation de suite d'accords dans [4]. Les vecteur d'observations sont comparés à un ensemble de templates ou modèles théoriques qui correspondent aux états dans lequel le système peut se trouver. Par soucis de simplicité, nous nous sommes limités à un dictionnaire de 24 accords de trois sons (12 majeurs, 12 mineurs). Ce sont des vecteurs à 12 dimensions contenant des 1 lorsque la note est présente dans l'accord et 0 sinon. Par exemple pour C majeur le modèle est 100010010000. Ces templates sont notés $\widehat{c_i}$, où $i \in [1; 24]$.

Pour chaque trame du chromagram, nous calculons le résultat de la multiplication du vecteur de chroma par les 24 templates distincts. Nous obtenons donc un ensemble de 24 valeurs que nous normalisons de manière à ce que la somme fasse 1, ce qui nous fournit un ensemble de "pseudo probabilités" notées $P(c_i)$ avec $i \in [1; 24]$. Le système peut se trouver dans l'un des 24 états correspondant aux 24 accords majeurs ou mineurs possibles. La valeur de $P(c_i)$ la plus élevée correspond bien sûr à l'accord le plus probable.

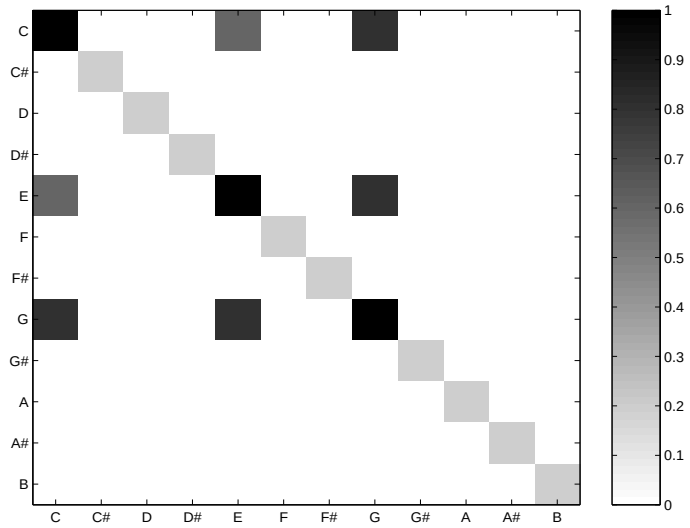

FIG. 6 – Matrice de covariance pour C majeur

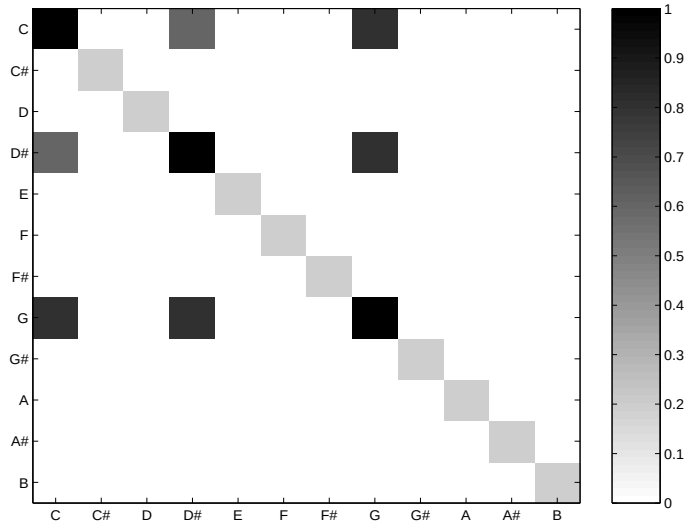

FIG. 7 – Matrice de covariance pour C mineur

Nous pouvons alors faire une estimation instantanée des accords du morceau en choisissant à chaque instant l'accord correspondant à la valeur donnée par $maxP(\widehat{c_i})$.

Cependant une estimation instantanée ne correspond pas à la réalité puisque la progression des accords dans le temps suit une logique liée aux règles musicale. Ainsi que nous l'avons vu, il est plus probable qu'un accord de C majeur soit suivi par un accord de G majeur que par un accord de A# majeur. L'introduction d'une matrice de transition va nous permettre de tenir compte de ces propriétés. Nous prenons ici la même matrice de transition que celle utilisée dans le cas de la « méthode par modélisation gaussienne des observations ».
Pour chaque trame d'analyse, nous obtenons 24 valeurs comprises entre 0 et 1 qui correspondent à la probabilité d'être dans l'un des 24 états possibles. Trouver la suite d'accords la plus probable dans le temps revient à choisir à chaque instant le

meilleur chemin parmi les 24 possibles, ce que nous faisons en effectuant un décodage Viteri.

Nous nous référerons à cette méthode en parlant de « méthode par corrélation avec des templates ». [1]

5 Développements du système

Les résultats obtenus par le système initial ne sont pas très satisfaisants (le taux de bonne reconnaissance des accords atteint seulement 30% environ). Nous avons cherché à identifier les défauts du système et à les corriger. Ce travail fait l'objet de cette section.

5.1 Améliorations de la partie signal

5.1.1 Tuning

Le tuning des enregistrements dont on dispose n'est en général pas parfait. En effet, il est possible que les instruments utilisés pour l'enregistrement n'aient pas été accordés selon le tuning classique de $440Hz$. D'autre part, le tuning peut être modifié par l'enregistrement. La différence entre le tuning d'une pièce et la position théorique des pics d'énergie du signal peut avoir une grande influence sur l'estimation des accords. C'est pourquoi il est nécessaire de réestimer le tuning de chaque pièce afin de pouvoir utiliser de manière cohérente le système que nous avons construit. Nous utilisons la méthode proposée par Peeters dans [9]

Le tuning de la pièce est supposé rester constant au cours du temps. Nous cherchons quel est le tuning qui explique le mieux le spectre d'énergie du signal. Un ensemble de tunings compris entre $427Hz$ et $452Hz$, fréquences correspondant aux quarts de ton au dessous et au dessus du A4 est testé. Pour chacun, nous calculons la part du spectre d'énergie pouvant être expliquée par l'énergie située aux fréquences correspondant aux demi-tons basés sur ce tuning. Pour cela, pour chaque tuning t et chaque trame m on calcule l'erreur suivante : (il s'agit de mesurer le rapport entre l'énergie du spectre expliquée par le tuning testé et l'énergie totale du spectre)

$$\epsilon(t,m) = 1 - \frac{\sum_{f_t} A(f_t,m)}{\sum_f A(f,m)} \tag{9}$$

où A est l'amplitude de la transformée de Fourier et f_t la fréquence correspondant aux demi-tons basés sur le tuning testé. Le tuning est choisi tel que sa valeur minimise l'erreur calculée dans le temps.

[1] J'ai donné à cette méthode le nom de « méthode par corrélation avec des templates »par analogie avec une méthode utilisant un calcul de corrélation. La multiplication avec les templates aurait pu être remplacée par le calcul d'une corrélation entre vecteur de chroma et templates ; cependant, c'est par la méthode employée que nous obtenons les meilleurs résultats.

Le signal est ensuite reéchantillonné de manière à le ramener à un tuning de 440 Hz, ce qui permet de baser le reste du système sur un tuning de 440Hz.

La figure 8 représente l'histogramme des tunings estimés pour chacun des deux albums des Beatles *Please Please Me* et *Beatles for Sale*. Nous pouvons voir que le tuning de l'ensemble des morceaux du premier CD est très loin de 440 Hz. L'ajout du tuning améliore de 10% relatifs environ les résultats.

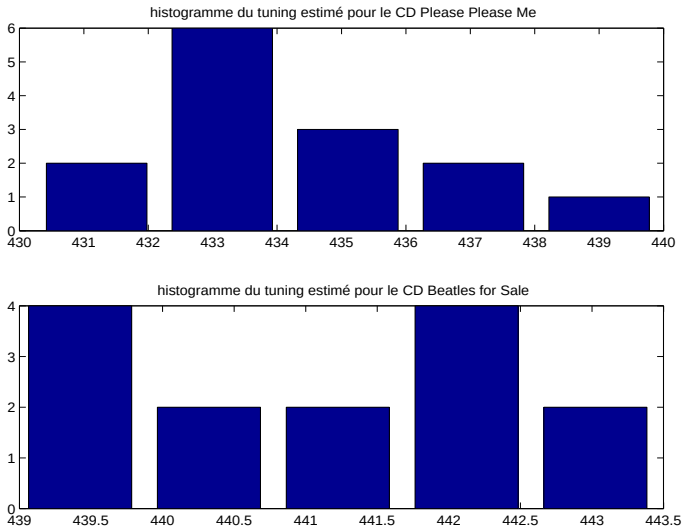

FIG. 8 – Histogramme des tunings estimés pour les albums Please Please Me et Beatles for Sale

5.1.2 Taille de la fenêtre, résolution fréquentielle/temporelle

Le signal audio est fenêtré avec une fenêtre d'analyse de type Blackman. La résolution fréquentielle définie par $\Delta f = fs/N$, où fs est la fréquence d'échantillonage en Hz et N la taille de la fenêtre en nombre d'échantillons, dépend de la taille de la fenêtre d'analyse.

Celle-ci a une influence sur les observations obtenues à la sortie du système. Jusqu'à maintenant, nous avons utilisé une fenêtre de taille $0.743s$, ce qui est relativement élevé en termes d'harmonie musicale. Les trames successives se recouvrent avec un pas de $\frac{1}{8}$, afin d'améliorer la résolution temporelle. Cependant nous avons remarqué que l'on peut obtenir de meilleurs résultats sans utiliser une fenêtre de taille aussi importante.

Nous ne prenons pas en compte les très hautes fréquences dans notre analyse, ces parties du spectre étant en général très bruitées en raison des sons percussifs, de la friction des cordes, etc.

Nous avons choisi la taille de la fenêtre en fonction de la résolution nécessaire à l'intervalle de fréquences considéré. Plus on s'approche des basses fréquences, plus

les notes sont rapprochées en fréquence les unes des autres. Pour séparer ces notes, le critère classique utilisé est la largeur de la fenêtre à puissance moitié (à $-3dB$). Pour pouvoir séparer tout couple de pics du spectre, il faut que la différence entre deux lobes principaux soit supérieure à la largeur de bande à puissance moitié (voir figure 9).

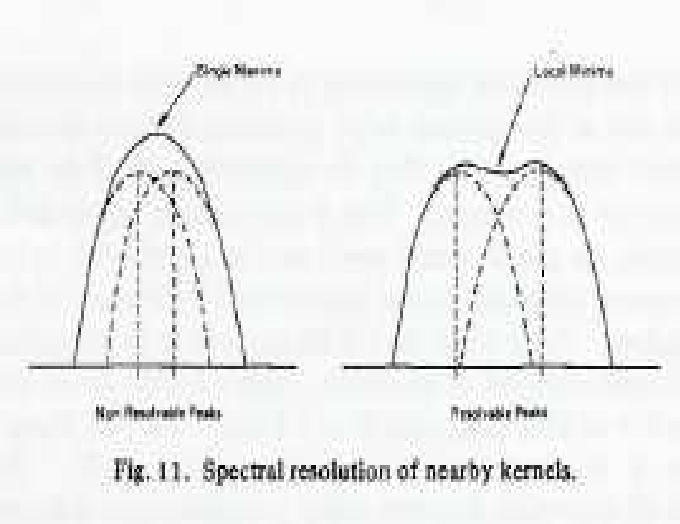

FIG. 9 – Résolution spectrale.A gauche, pas de séparation des pics. A droite, séparation des pics. Extrait de [17]

Par exemple considérons les fréquences du signal à partir de 40 Hz. Nous voulons pouvoir séparer tout couple de notes adjacentes. Les fréquences en Hz du $D\#_1$ et du E_1 son égales respectivement à $38.9Hz$ et $41.2Hz$. La différence de fréquence entre ces deux notes adjacentes de $2.3Hz$. Pour une fenêtre de Blackman, la largeur de bande à $-3dB$ est égale à 1.68 Bins. Pour pouvoir séparer ces deux notes, il faut que la fenêtre soit de taille supérieure à $\frac{1.68}{3.3} = 0.51s$.

5.1.3 Filtrage médian

On applique un filtrage médian sur 10 points au signal à la sortie des filtres, avant le mapping du spectre de demi-tons vers le spectre de chromas dans le calcul du chromagram. Ce filtrage médian permet de réduire les transitoires et les bruits tels que les sons de batterie qui faussent le calcul du chromagram. En effet, ces bruits perturbent le spectre du signal en se mélangeant et en recouvrant les composantes harmoniques du signal.

Appliquer un filtrage médian à la sorte du chromagram permet d'améliorer très légèrement (de l'ordre de 0.5% relatifs) les résultats car des changements d'accords en des très courts intervalles de temps sont peu probables.

5.1.4 Échelle utilisée dans la représentation spectrale : énergie, amplitude, sones

Nous avons testé l'influence de l'échelle utilisée dans la représentation spectrale. Le choix de cette échelle peut influencer de façon importante les observations obtenues en sortie du chromagram et donc les résultats. Les différentes échelles utilisées sont :

- Échelle d'amplitude
- Échelle d'énergie
- Échelle des sones. Il s'agit d'une échelle quantitative qui permet de mesurer l'intensité sonore relative entre deux sons, autrement dit de préciser si un son est deux fois plus fort ou moins fort qu'un son de référence. Nous utilisons la formule suggérée par Bladon et Lindbloom en 1981

$$A_s(k) = 2^{\frac{1}{10}(A_{db}(k)-40)} \quad \text{if } A_{db}(k) > 40 \tag{10}$$
$$= 1/40 A_{db}(k)^{2.642} \quad \text{else} \tag{11}$$

5.2 Introduction du modèle de Gomez dans le système

5.2.1 Présentation du modèle

La plupart des méthodes utilisées jusqu'à présent pour l'estimation d'accords adoptent une approche similaire : on calcule d'abord la corrélation entre les vecteurs d'observation et des templates représentant la distribution de chroma théorique pour chaque accord.

L'un des principaux défauts du système que nous avons construit jusqu'à présent est qu'il ne tient pas compte des harmoniques du signal observé. En effet, dans la représentation spectrale du signal audio, on observe, non pas l'intensité des différentes notes composant le signal, mais un mélange de leurs harmoniques. De même, ces harmoniques sont présentes dans le chromagram qui est une représentation compacte du spectre.

Par exemple, dans un accord de C majeur composé des notes C, E et G, la troisième harmonique du C qui est un G renforce la valeur du G dans le chromagram. Lorsque trois notes sont jouées simultanéent, il y aura dans le chromagram un certain nombre de partiels d'intensité non négligeable à des pitch class autres que celles correspondant aux notes de l'accord. La présences de ces harmoniques n'étant pas prise en compte, il en résulte des erreurs dans l'estimation des accords.

La figure 10 illustre ces propos. Elle représente les valeurs du chromagram pour un accord de C majeur joué par un violoncelle (G4), une flûte (C5) et une trompette (E5). Nous pouvons voir qu'il y a de l'énergie présente à chaque point du chromagram, bien que les instruments ne jouent que les notes de l'accord. Cette énergie

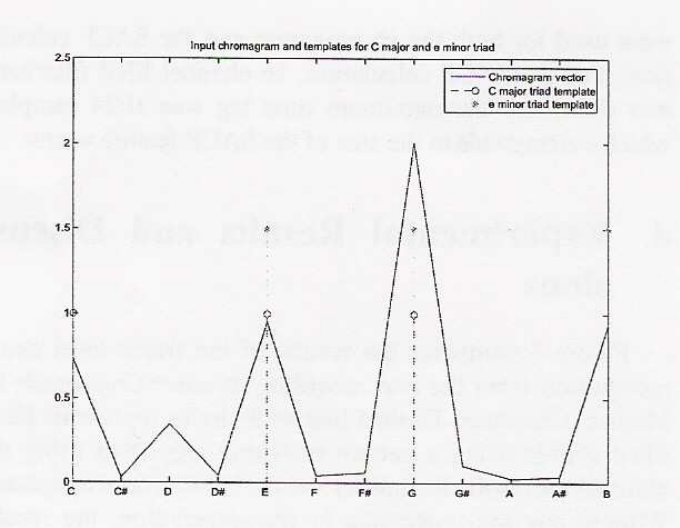

FIG. 10 – Chromagram pour un accord de C majeur et templates pour C majeur et E mineur.Extrait de [18]

vient des harmoniques contenues dans le spectre.
Nous pouvons voir que la classe de hauteur correspondant à B contient plus d'énergie que celle correspondant à C. Ainsi, si on calcule la corrélation entre le vecteur de chroma obtenu et les templates correspondant à C majeur (100010010000) et E mineur (000010010001), le template correspondant à E mineur sera plus fortement corrélé au vecteur de chroma de C majeur que le template correspondant à C majeur.

Remarquons que ce problème a déja été soulevé par [18], dans le cas de la reconnaissance automatique d'accords. Lee propose de remplacer l'utilisation des vecteurs de chroma par une nouvelle approche basée sur la perception humaine et d'utiliser une fonction qu'il nomme *Summary Autocorrelation Function*. Nous avons décidé de continuer à utiliser le chromagram mais de prendre en compte l'influence des harmoniques.

Dans [8], Gomez propose de remplacer les vecteurs de chroma ou Pitch Class Profiles de Fujishima [1] par des Harmonic Pitch Class Profiles (HPCP). Ceux-ci prennent en compte les harmoniques présents dans le spectre en utilisant une enveloppe spectrale théorique qui détermine leurs amplitudes. L'enveloppe spectrale est choisie de manière à ce que la contribution des harmoniques décroit avec la fréquence. En notant f_0 la fréquence fondamentale (première harmonique) d'une note, l'amplitude de la h^{eme} harmonique vaut s^{h-1} où s est un facteur de décroissance spectrale fixé empiriquement à 0.6. Dans [8] et [7], seules les 4 premières harmoniques du signal ont été prises en compte. Nous avons introduit ce modèle dans notre système avec 4 et 6 harmoniques. L'emploi de 6 harmoniques permet d'obtenir les meilleurs résultats, ainsi que nous le verrons par la suite.

5.2.2 Évaluation du modèle de Gomez

Avant d'introduire dans notre système le modèle proposé par Gomez, nous l'avons testé sur un ensemble d'accords construits à partir de notes jouées par divers instruments. L'enveloppe spectrale théorique du son produit par un instrument dépend normalement non seulement de l'instrument mais aussi de la tessiture, de l'intensité du jeu, etc.

Nous construisons un ensemble d'accords tests (majeurs et mineurs) à partir de notes jouées par un ensemble d'instruments variés (cordes, cuivres, bois...). Cet ensemble d'accords tests est représentatif de l'étendue de la tessitures des instruments. Pour chaque accord test c_i, nous calculons la corrélation de cet accord avec les 24 templates correspondants aux 24 accords majeurs et mineurs $\widehat{c_i}$.(Pour la création de ces templates, voir section 4.4, paragraphe « Vecteurs moyens »). L'accord identifié par le système est celui dont la valeur de la corrélation est la plus élevée parmi les 24 valeurs calculées.

Les résultats obtenus montrent que les accords majeurs sont correctement identifiés à 75% (résultats pour 6 harmoniques pris en compte). Ces résultats sont équivalents en moyenne à ceux obtenus en utilisant des templates sans prendre en compte les harmoniques. C'est pourquoi nous prendrons soin de toujours comparer par la suite les résultats que nous obtenons en tenant compte ou non des harmoniques. Il faut noter que les erreurs surviennent uniquement lorsque l'accord testé appartient à la limite de la tessiture des instruments (extrême aigü ou grave). Dans le cas de la musique populaire en particulier (où les instruments sont en général sollicités dans le médium de leur tessiture), ce modèle sera donc *a priori* satisfaisant.

5.2.3 Introduction du modèle de Gomez dans les matrices de moyenne et de covariance

Les vecteurs moyens et les matrices de covariance des états correspondant à C majeur et C mineur sont construits puis on en déduit les paramètres des autres modèles par permutation circulaire.

5.2.3.1 Vecteurs moyens Nous rappelons ici les premières harmoniques des notes composant les accords de C majeur et C mineur et donnons la valeur de leur amplitude dans le spectre selon le modèle de Gomez dans les tableaux 1 et 2

C majeur						
note	harmoniques					
C	C	C	G	C	E	G
E	E	E	B	E	G#	B
G	G	G	D	G	B	D
amplitude attribuée	1	0.6	0.36	0.216	0.1296	0.0778

Tab. 1 – Contribution des premières harmoniques pour un accord de C majeur

C mineur						
note	harmoniques					
C	C	C	G	C	E	G
D#	D#	D#	A#	D#	G	A#
G	G	G	D	G	B	D
amplitude attribuée	1	0.6	0.36	0.216	0.1296	0.0778

Tab. 2 – Contribution des premières harmoniques pour un accord de C mineur

On obtient les vecteurs moyens présentés dans le tableau 3.

C majeur			C mineur		
notes	4 harmoniques	6 harmoniques	**notes**	4 harmonique	6 harmoniques
C	1.816	1.816	**C**	1.816	1.816
C#	0	0	**C#**	0	0
D	0.36	0.4378	**D**	0.36	0.4378
D#	0	0	**D#**	1.816	1.816
E	1.816	1.9456	**E**	0	à.1296
F	0	0	**F**	0	0
F#	0	0	**F#**	0	0
G	2.176	2.2538	**G**	2.176	2.3834
G#	0	0.1296	**G#**	0	0
A	0	0	**A**	0	0
A#	0	0	**A#**	0.36	0.4378
B	0.36	0.5674	**B**	0	0.1296

Tab. 3 – Amplitudes des notes des templates d'accords avec modèle de Gomez

Ces vecteurs moyens correspondent également aux 24 templates de référence utilisée par la « méthode par corrélation avec templates ».

5.2.3.2 Matrices de covariance Nous avons considéré l'influence des 4 premières harmoniques dans la matrice de covariance. Nous donnons à nouveau les résultats pour l'accord de C majeur, les autres cas s'en déduisant aisément.

Jusqu'à présent, nous avons considéré que seules les notes C, E et G composant l'accord étaient corrélées entre elles. Nous allons maintenant considérer que leurs 4 premières harmoniques sont également corrélées entre elles. Nous raisonnons de la manière suivante : si C change, G change obligatoirement aussi ; de même, si E change, B change également, ainsi que D lorsque G change. Nous ajoutons alors dans la matrice de covariance des valeurs correspondant à la corrélations entre ces notes. Celles-ci sont fixées empiriquement, mais en respectant les règles d'ordre établies lors de la construction de la matrice de covariance dans la section 4.

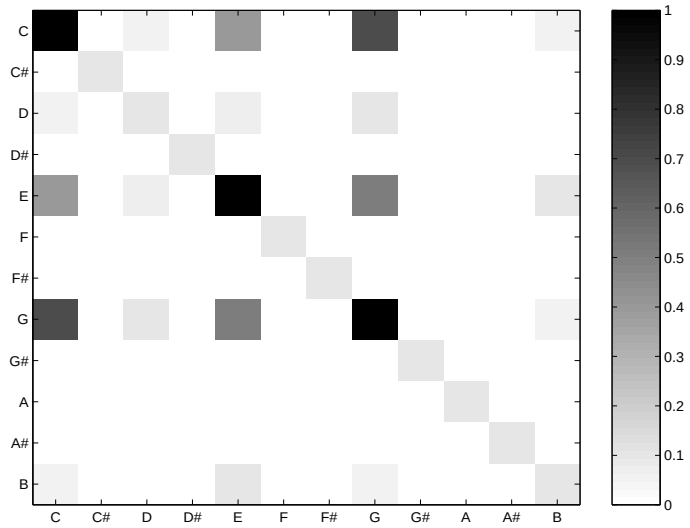

FIG. 11 – Matrice de covariance pour C majeur, avec prise en compte des 4 premires harmoniques

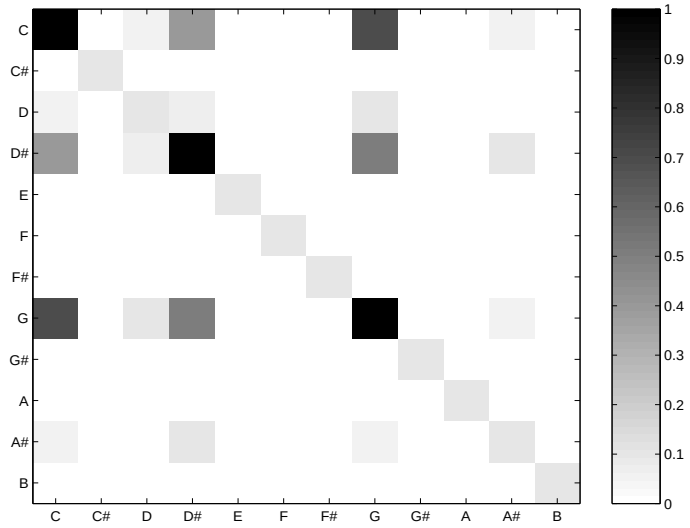

FIG. 12 – Matrice de covariance pour C mineur, avec prise en compte des 4 premires harmoniques

5.3 Synchronisation sur les tactus

Bello et Pickens reprennent dans [5] l'idée de Bartsch et Wakefield [2], de calculer le chromagram en se synchronisant sur les battements (tactus) du morceau. Pour cela, après avoir calculé le chromagram trame par trame, on fractionne celui-ci en segments dont le début et la fin correspondent à un battement puis on prend la valeur moyenne des vecteurs de chroma se trouvant entre les deux tactus.

Nous obtenons également de meilleurs résultats en nous synchronisant sur les tactus. Cependant nous n'obtenons qu'une amélioration relative de 1.5% environ. Cela peut s'expliquer par le fait que nous avons ajouté un filtrage médian lors de la construction du chromagram, ce qui diminue l'influence du fait de moyenner le

chromagram sur les tactus.

Remarque : Etant donné que nous ne disposons pas encore pour le moment des tactus pour chacun des morceaux, les résultats ci-dessous seront donnés sans synchronisation sur les battements.

5.4 Rappels sur l'analyse linéaire discriminante (ALD)

Nous faisons ici quelques rappels sur l'analyse linéaire discriminante dont nous nous servirons par la suite.

L'analyse linéaire discriminante est une technique utilisée dans de nombreux domaines qui sert à déterminer la contribution des variables qui expliquent l'appartenance d'éléments à un groupe. On étudie les données provenant de groupes connus *a priori*. Deux buts principaux :

- Parmi les groupes connus, quelles sont les princiales différences que l'on peut déterminer à l'aide des variables mesurées ?
- Peut-on déterminer le groupe d'appartenance d'une nouvelle observation uniquement partir des variables mesurées ?

Dans notre cas, nous cherchons à discriminer deux groupes (majeur et mineur). Les variables sont des vecteurs à 12 dimensions (12 variables mesurées). On cherche une combinaison linéaire des variables qui permettrait de maximiser la discrimination entre les deux classes majeur et mineur. Il s'agit donc de trouver un vecteur qui sépare du mieux possible les groupes. Nous noterons :

- n : nombre d'observations, ici nombre de trames utilisées pour l'analyse discriminante,
- p : nombre de variables mesurées, ici 12,
- k : nombre de groupes, ici 2,
- B : matrice de variabilité entre les groupes (),
- W : matrice de variabilité dans les groupes (),
- T : matrice de variabilité totale.

Nous cherchons un vecteur u de manière à ce que la projection sur ce vecteur transforme l'espace des observations (vecteurs de chroma) en un nouvel espace où la discrimination est maximum. Nous devons pour cela maximiser la variabilité intergroupes par rapport à la variabilité totale. u doit maximiser le rapport : $\frac{u'Bu}{u'Tu}$

Cela revient, après calculs à résoudre :

$T^{-1}Bu = \lambda u$

et

$u'Tu = 1$.

Le vecteur recherché est le vecteur propre associé à la plus grande valeur propre de $T^{-1}B$.

•aspect clasification
Supposons que l'on a de nouvelles observations que l'on veut classer dans l'un des deux groupes. Une observation sera classée dans le groupe pour lequel la probabilié d'appartenir à ce groupe étant donné les valeurs observées est maiximum. En pratique, on ne peut calculer ces probabilités que si les observations proviennent d'une loi multinormale (ou s'en rapprochent le plus possible).

6 Implantation

L'implantation des différentes méthodes vues auparavant a été réalisée sous matlab. Pendant ce stage, une très large part du temps a été consacrée à la programmation des algorithmes.

6.1 Schéma du système complet

Nous donnons ici (voir figure 13) le schéma global du système afin d'illustrer le rôle des fonctions programmées sous matlab.

6.2 Fonctions programmées sous matlab

On trouvera ci-dessous un bref descriptif des fonctions programmées.

Les morceaux des deux CD des Beatles sont classés par la fonction *Fdatabase_beatles* . Celle-ci nous permet d'accéder au fichier son. wav, à la transcription .lab, aux tactus ainsi qu'à la valeur de décalage éventuel entre la transcription des accords et la le signal audio dont on dispose.

La fonction principale prend en entrée le numéro du morceau dont on veut extraire la suite d'accords et nous renvoie l'estimation insantanée des accords ainsi que la suite d'accords la plus probable dans le temps obtenue par décodage Viterbi. Elle nous donne également le pourcentage des erreurs qui ne sont pas trop graves (voir section suivante).

Le schéma de la procédure est le suivant :

- Recherche du morceau dans la base de données avec *Fdatabase_beatles*,
- Lecture du fichier .wav, transformation au format mono en prenant la moyenne sur les deux canaux,

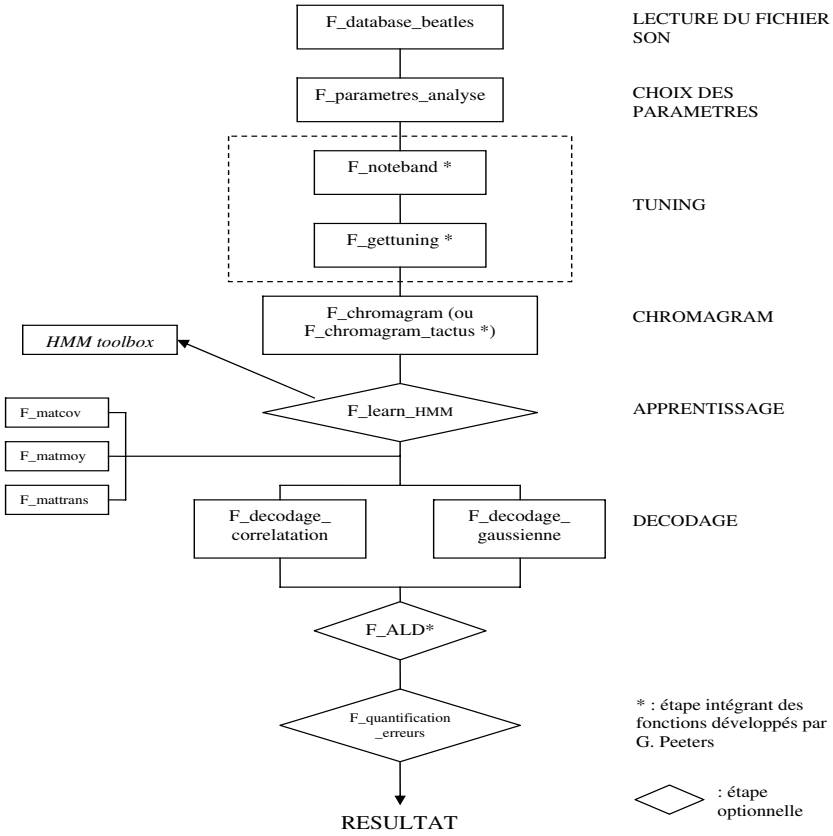

FIG. 13 – Schéma récapitulatif des principales fonctions implantées

– Soustraction de la composante continue (moyenne du signal) éventuelle,

– Entrée des paramètres d'analyse à l'aide de la fonction *Fparametres_analyse*,

  – T_n : longueur totale du signal
  – L_sec : longueur de la fenêtre utilisée en secondes
  – L_n : taille de la fenêtre en nombre d'échantillons

  - STEP_n : pas d'avancement en nombre d'échantillons
  - nbr_frame : nombre de trames
  - fenetre_v : fenêtre d'analyse
  - N : nombre de points de la FFT
  - minfreq_hz : borne minimale des filtres
  - maxfreq_hz : borne maximale des filtres
- Calcul des filtres avec la fonction *Fnoteband*,

- Tuning avec la fonction *Fgettuning* qui nous donne le meilleur tuning estimé parmi ceux testés, le coefficient de reéchantillonnage nécessaire pour le ramener à un tuning de $440Hz$, le signal reéchantilloné et le paramètres correspondant,

- Calcul du chromagram avec la fonction *F_chromagram*

- Vecteurs moyens et matrice de covariane obtenus par les fonctions *Fmatrice_obsth* et *Fmatrice_cov*,

- Distribution des probabilités initiales et matrice de transition obtenue par *Fmatrice_trans*,

- Décodage Vierbi avec la fonction *Ftest_viterbi*,

- Évaluation des erreurs avec la fonction *Fquantification_des _erreurs* qui nous donne la pourcentage des erreurs correspondant aux parallèles majeur/mineur, relatives majeur/mineur ainsi que les confusions avec la dominante ou la sous-dominante.

Il est possible de rajouter d'autre fonctionnalités telles qu'effectuer l'apprentissage de la matrice de transition ou faire une analyse discriminate.

Lors de ce stage ont été programmées environ 60 fonctions différentes qui ont servi pour tester les différentes méthodes proposées, y compris certaines que l'on n'a pas retenues ensuite (*constant Q* par exemple). L'ensemble du système final est constitué d'une trentaine de fonctions.

6.3 Exemple

La figure 14 montre un exemple des résultats obtenus par notre système. Elle a été obtenue par la « méthode par corrélation avec des templates ». Le résultat obtenu pour ce morceau est de 82% de bonne reconnaissance des accords sur les trames.

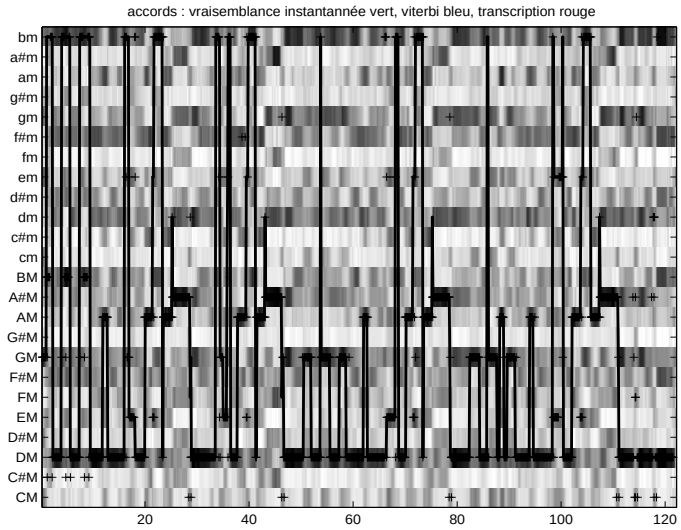

FIG. 14 – Exemple de résultat. I am a Loser, Beatles for Sale (82% de bonne reconnaissance sur les trames). Vert : vraisemblance instantanée. Bleu : décodage Viterbi. Rouge : Transcription théorique

7 Évaluation des résultats

Les résultats sont évalués sur une base de données annotée composée de deux CD des Beatles présentés précédement. Nous disposons pour chacun des morceaux :

- du signal audio au format .wav,
- de la transcription théorique .lab,

Le calcul du taux de bonne reconnaissance des accords est fait en moyenne sur l'ensemble des trames (environ 33000 trames pour des durées de morceaux comprises entre 90s et 150s).

7.1 Résumé

Nous avons regroupé par la suite dans des tableaux les résultats obtenus en reprenant un à un les divers éléments du système présentés précédement afin d'en donner une justification quantitative. Les tableaux indiquent les paramètres du modèle choisi pour chaque expérience et le résultat correspondant pour chaque CD et en moyenne sur les deux. Les paramères considérés sont :

- tuning / non tuning,
- résolution (nombre de filtres considérés),
- filtrage médian / non filtrage médian,
- prise en compte des harmoniques ou non,
- taille de la fenêtre,
- bornes des fréquences,
- échelle (amplitude, énergie, sones).

7.2 Analyse

Les résultats obtenus par notre système sont intéressants par rapport à ceux obtenus par les travaux antérieurs sur la détection de suites d'accords.

La « méthode par modélisation gaussienne des observations »diffère de celle proposée par Bello et Pickens dans [5] par deux points principaux : d'une part nous n'avons pas effectué d'entraînement de la matrice de transition, d'autre part nous avons pris en compte les harmoniques du signal audio dans la création des paramètres du modèle. Les résultats que nous obtenons avec cette méthode sans synchronisation sur les tactus sont comparables et même légèrement meilleurs que ceux obtenus dans [5]. Remarquons que dans [5], aucun résultat n'a été donné sans apprentissage de la matrice de transition.

D'autre part, la « méthode par corrélation avec templates »peut être comparée à celle proposée par Harte et Sandler dans [4]. Là aussi nous avons ajouté deux principales contributions : prise en compte des harmoniques dans l'analyse et décodage Viterbi. Les résultats que nous obtenons par cette méthode à partir du chromagram sont nettement supérieurs à ceux obtenus par la « méthode par modélisation gaussienne des observations ».

Les résultats pour les deux albums des Beatles sont très différents. Le système obtient un résultat de 73% d'accords bien identifiés en moyenne sur les trames, pour le quatrième album *Beatles for Sale* alors que l'on n'obtient que 61% d'accords bien identifiés pour le second album, *Please Please Me*. La différence est probablement due, d'une part au styles des deux CD (dans le premier sont utilisés des guitares électriques ainsi que des harmonicas comme instruments solo alors que dans le deuxième ce sont des guitares acoustiques et des instruments à clavier qui sont en majorité utilisés...), d'autre part au type d' accords qui les composent.

La plupart des accords de la transcription sont des accords majeurs et mineurs de trois sons. Cependant, il y a également des accords de quatre ou cinq sons ainsi que d'autres accords de trois sons (par exemple quinte diminuée). Ces accords sont une source d'erreurs de reconnaissance du système. Un accord de 7^{eme} majeure, par exemple (C E G B) correspond dans notre système à C majeur (C E G) mais aurait pu également correspondre à E mineur (E G B). Cependant, par simplicité, nous n'avons retenu qu'une des deux solutions.
Nous avons remarqué que pour certains morceaux, le système choisit E mineur plutôt que C majeur alors que c'est l'inverse pour d'autres. La manière dont a été effectué le mapping entre la transcription des accords réelle et le dictionnaire utilisé a donc une influence sur les résultats. Remarquons que celle-ci est parfois non négligeable car, si en nombre d'accords, ceux qui n'appartiennent pas à notre dictionnaire ne représentent que 10% du nombre total des accords, leur durée est parfois longue comparée aux autres accords du morceau et pour certains, ils vont représenter jusqu'à un quart de la durée totale du morceau, ce qui, pour les raisons décrites ci-dessus dégrade parfois nos résutats...

7.2.1 Apprentissage

L'utilisation de HMM devrait permettre d'estimer les suites d'accords sans avoir à introduire aucune connaissance musicale dans le modèle (présence d'harmoniques dans le spectre d'une note, probabilités de transition entre accords...). En effet, les paramètres seraient appris à partir d'un ensemble de morceaux d'entrainement.

Les résultats que nous avons obtenus en utilisant les morceaux d'un CD comme ensemble de morceaux d'entrainement et en testant le système sur l'autre album donnent de bien moins bons résultats que sans apprentissage. Il serait intéressant de tester l'algorithme en incluant la phase d'apprentissage, mais en utilisant une plus grande base de données. Les caractéristiques des morceaux dont nous disposons sont très différentes selon les albums auxquels ils appartiennent en raison de la différence de style et de l'instrumentation qui existe entre les deux, ce qui peut expliquer les résultats obtenus avec apprentissage. Nous prévoyons également de tester par la suite un apprentissage "croisé", c'est à dire utiliser 13 des 14 morceaux d'un CD comme ensemble d'apprentissage pour tester celui qui reste, puis permuter ...

7.2.2 Choix de la méthode

Les résultats obtenus montrent que la méthode basée sur le calcul de la corrélation entre les accords réels et les templates suivi d'un décodage Viterbi donne en moyenne un meilleur taux d'identification des accords que la méthode basée sur une modélisation gaussienne de la distribution des observations.

7.2.3 Filtrage médian

L'ajout d'un filtrage médian avant le mapping avec les chromas produit une nette amélioration. Dans le cas de la « méthode par modélisation gaussienne des observations »nous obtenons une amélioration relative de 4.66% en moyenne sur toutes les trames. Concernant la « méthode par corrélation avec templates », l'amélioration relative des résultats est encore plus sensible : elle est de 18.03%. Nous avons donc réussi à éliminer une partie du bruit introduit par les transitoires et les ornementations contenus dans le signal.

7.2.4 Taille de la fenêtre et bornes des fréquences

Les bornes de fréquences utilisées dans les principaux travaux précédents sont $98Hz$ à $5250Hz$ c'est-à-dire que l'on considère une plage de fréquences du $G2$ au $E8$. Il n'est en fait pas nécessaire de considérer une telle plage de fréquences, se restreindre à la plage $[60Hz : 4500Hz]$ (ce qui correspond à [B1 :D8]) permet obtenir de meilleurs résultats en moyenne.

7.2.5 Introduction du modèle de Gomez

La prise en compte des harmoniques dans le calcul du chromagram permet d'améliorer sensiblement les résultats, quelle que soit l'échelle utilisée. Nous analysons ici les cas les plus intéressants. Dans le cas de la « méthode par modélisation gaussienne des observations », nous pouvons noter une amélioration relative de 7% environ si l'on introduit le modèle de Gomez dans la matrice des vecteurs moyens. En introduisant de plus ce modèle dans les matrices de covariance nous obtenons une amélioration relative de 36% en moyenne sur toutes les trames.

7.2.6 Échelle et nombre d'harmoniques utilisées

•A propos du nombre d'harmoniques utilisées

Dans [8], Gomez propose de prendre en compte uniquement la contribution des quatres premières harmoniques d'une note. Nous avons remarqué qu'il peut être intéressant de prendre en compte un nombre plus importants d'harmoniques.

En effet, dans le cas de la « méthode par corrélation avec templates », nous avons testé le modèle avec six harmoniques et cela donne dans tous les cas (échelle, morceau,...) des résultats supérieurs à ceux obtenus en ne tenant compte que des 4 premières harmoniques. La différence est en particulier visible lorsqu'on utilise une échelle d'amplitude. Néanmoins, il est inutile de considérer plus de 6 harmoniques dans la création des vecteurs de chroma correspondant aux templates, les résultats sont les mêmes.

Remarquons que l'introduction du modèle de Gomez donne de meilleurs résulats en moyenne sur l'ensemble des morceaux alors que lorsque nous l'avons évalué sur de simples accords, il n'y avait pas vraiment d'amélioration par rapport au cas sans prise en compte des harmoniques. Nous nous intéresserons dans des travaux futurs à ce problème.

•Échelle

Le choix de l'échelle modifie de façon significative les résultats. Cependant, on ne peut donner de règle générale quant au choix de celle-ci, les résultats obtenus variant selon la méthode employée et le nombre d'harmoniques utilisées.

En ce qui concerne la « méthode par modélisation gaussienne des observations », les meilleurs résultats sont obtenus en prenant une échelle d'amplitude et en tenant compte de 4 harmoniques (voir le tableau 4). Cependant, lorsque l'on ne tient pas compte des harmoniques des notes, l'emploi d'une échelle d'énergie donne des résultats nettement supérieurs à ceux obtenus avec une échelle d'amplitude (57.62% contre 48.52%).

Par contre, dans le cas de la « méthode par corrélation avec templates », l'utilisation d'une échelle d'énergie donne de meilleurs résultats dans tous les cas (voir le tableau 5). Remarquons que lorsque l'on utilise une échelle d'amplitude dans le cas

de cette méthode, le nombre d'harmoniques pris en compte ne change pas les résultats de manière aussi spectaculaire que lorsque l'on utilise une échelle d'amplitude (amélioration relative de 2.5% entre $nbh = 1$ et $nbh = 6$ dans le cas de l'énergie, contre 14.44% dans le cas de l'amplitude).

L'emploi d'une échelle de sones diminue systématiquement les résultats, quelle que soit la méthode employée.

Il faut remarquer que les résultats obtenus sont assez différents de ceux obtenus par Peeters dans [7]. L'emploi d'une échelle de sones en particulier donne les meilleurs résultats dans [7].

E	nbh	CD1	CD2	MOYENNE
amplitude	1	45.86	51.18	48.52
énergie	1	51.35	63.88	57.615
sones	1	24.70	30.56	55.26
amplitude	**4**	**55.93**	**67.27**	**61.6**
énergie	4	51.59	61.11	56.35
sones	4	39.84	49.18	44.81

Tab. 4 – Influence de l'échelle et du nombre d'harmoniques, cas de la « méthode par modélisation gaussienne des observations ». E : échelle, nbh : nombre d'harmoniques prises en compte

E	nbh	CD1	CD2	MOYENNE
amplitude	1	53.80	61.55	57.68
énergie	1	58.75	71.82	65.29
amplitude	4	57.96	67.27	62.62
énergie	4	59.28	71.46	65.37
amplitude	6	59.98	72.04	66.01
énergie	**6**	**60.34**	**73.30**	**66.91**
sones	6	47.32	54.98	51.15

Tab. 5 – Influence de l'échelle et du nombre d'harmoniques, cas de la « méthode par corrélation avec templates ». E : échelle, nbh : nombre d'harmoniques prises en compte

•Conclusion

En conclusion, nous pouvons dire que le choix de l'échelle et celui du nombre d'harmoniques utilisées sont fortement liés. Si nous ne pouvons pas donner de règle générale en ce qui concerne le choix de l'échelle, nous pouvons dire que la prise en compte d'un certain nombres d'harmoniques dans la création des templates améliore nettement les résultats.

Étant donné qu'il existe une différence significative dans les résultats obtenus par la « méthode par corrélation avec templates »entre $nbh = 4$ et $nbh = 6$, nous pouvons penser qu'il serait possible d'améliorer les résultats obtenus par la « méthode par modélisation gaussienne des observations »en prenant 6 harmoniques dans le modèle. Cependant, nous ne disposons pas encore de résultats concluants car la méthode d'introduction des harmoniques dans les matrices de covariance n'est pas assez développée pour le moment.(Voir précedemment)

7.3 Quantification des erreurs

7.3.1 Influence de l'échelle et du nombre d'harmoniques sur les erreurs obtenues

Une grande partie des erreurs provient de confusions entre des accords qui sont harmoniquement proches. Nous pouvons considérer que ce ne sont pas des erreurs trop graves, car si l'on ne trouve pas exactement un accord mais un accord voisin, on peut tout de même utiliser le résultat pour trouver la tonalité ou la structure harmonique du morceau. Leur proportion varie en fonction des paramètres du système, et l'analyse de ces erreurs peut nous guider dans leur choix.

Nous nous intéressons ici uniquement au cas de la « méthode par corrélation avec templates ».

nbh	E	RCD1	P1	R1	SD1	D1	TE1	RAE1
1	A	53.80	11.06	0.37	3.42	23.23	37.96	71.34
1	E	58.75	11.91	0.66	4.6	21.66	38.83	74.77
4	A	57.96	12.68	0.39	5.09	18.45	36.61	73.35
4	E	59.28	11.95	0.93	9.77	15.70	38.35	74.90
6	A	59.97	15.38	1	8.57	22.9	39.29	75.70
6	E	60.52	13.32	2.55	14.62	17.57	48.06	79.49
6	S	47.32	14.86	0.94	8.25	24.1	48.15	72.69

TAB. 6 – Tableau du pourcentage des différentes erreurs du 2^{eme} CD en fonction des paramètres, cas de la « méthode par corrélation avec templates ». nbh : nombre d'harmoniques prises en compte, E : échelle (amplitude, énergie ou sones), RCD : résultat, P : parallèles mineur/majeur, R : relatives, SD : sous-dominante, D : dominante, TE : total erreurs pas trop graves, RAE : résultat sans compter les erreurs pas trop graves

Dans le cas de la « méthode par corrélation avec templates »l'utilisation d'une échelle d'énergie plutôt que d'une échelle d'amplitude permet de réduire la proportion des erreurs graves (voir les tableaux 6, 7 et 8). Il vaut donc mieux utiliser une échelle d'énergie puisque d'une part, ainsi que nous l'avons vu au paragraphe précédent, c'est avec celle-ci que l'on obtient les meilleurs résultats et d'autre part, c'est également dans ce cas que les erreurs sont les moins graves.

nbh	E	RCD2	P2	R2	SD2	D2	TE2	RAE2
1	A	61.55	6.09	0.59	4.90	24.98	36.56	75.61
1	E	71.82	8.42	1.08	9.23	20.52	39.26	82.88
4	A	67.27	7.43	0.99	8.31	19.88	36.61	79.25
4	E	71.46	8.23	1.68	19.66	14.39	43.96	84.01
6	A	72.04	9.97	2.17	13.46	25.03	50.63	86.20
6	E	73.30	7.91	4.21	27.02	16.03	55.17	88.03
6	S	54.98	8.29	0.83	9.82	30.02	48.96	77.03

TAB. 7 – Tableau du pourcentage des différentes erreurs du 1^{er} CD en fonction des différents paramètres, cas de la « méthode par corrélation avec templates ». nbh : nombre d'harmoniques prises en compte, E : échelle (amplitude, énergie ou sones), RCD : résultat, P : parallèles mineur/majeur, R : relatives, SD : sous-dominante, D : dominante, TE : total erreurs pas trop graves, RAE : résultat sans compter les erreurs pas trop graves

nbh	E	RCD1	RCD2	RMOY	TE1	TE2	ME	RAEMOY
1	A	53.80	61.55	57.68	37.96	36.56	37.26	73.44
1	E	58.75	71.82	65.29	38.83	39.26	39.05	78.84
4	A	57.96	67.27	62.62	36.61	36.61	36.61	76.30
4	E	59.28	71.46	65.37	38.35	43.96	41.16	79.62
6	A	59.97	72.04	66.01	39.29	50.63	44.96	81.29
6	**E**	**60.52**	**73.30**	**66.91**	**48.06**	**55.17**	**51.62**	**83.99**
6	S	47.32	54.98	51.15	48.15	48.96	48.55	74.87

TAB. 8 – Tableau du pourcentages des erreurs pas trop graves en fonction des différents paramètres en moyenne sur les deux CD, cas de la « méthode par corrélation avec templates ».

Remarquons aussi qu'en augmentant le nombre d'harmoniques prises en compte dans la création des templates, la proportion d'erreurs « pas trop graves »par rapport au nombre total d'erreurs augmente ce qui est une bonne chose.

Ainsi, si l'on inclut dans nos résultats les accords obtenus par notre système qui ne sont pas exacts mais harmoniquement proches des accords réels, on obtient un résultat de 84% d'accords identifiés correctement ou voisins en moyenne sur toutes les trames.

7.3.2 Comparaison des erreurs résultant des deux méthodes

Seules 35% des erreurs résultant de la méthode basée sur modélisation de la distribution des observations par une gaussienne sont dues à des confusions avec des accords harmoniquement proches. Par contre, en ce qui concerne la méthode basée sur des corrélations, jusqu'à plus de la moitié des erreurs appartiennent à la catégorie des erreurs « pas trop graves ». (voir le tableau 9).

		METT			METG		
nbh	**E**	**RMOY**	**ME**	**RAEMOY**	**RMOY**	**ME**	**RAEMOY**
1	A	57.68	37.26	73.44	48.52	34.3	66.17
1	E	65.29	39.05	78.84	57.62	33.6	71.86
4	A	62.62	36.61	76.30	61.60	33.62	74.51
4	E	65.37	41.16	79.62	56.35	35.46	71.83
6	A	66.01	44.96	81.29			
6	E	66.91	51.62	83.99			

TAB. 9 – Comparaison de la proportion d'erreurs « pas trop graves »obtenues selon la méthode utilisée : METT(« méthode par corrélation avec templates »), METG(« méthode par modélisation gaussienne des observations »)

Cette analyse montre que la « méthode par corrélation avec templates »est beaucoup plus robuste que l'autre méthode présentée. Non seulement les résultats obtenus sont bien meilleurs lorsque l'on s'intéresse uniquement au taux de reconnaissance exacte des accords, mais la moitié environ des erreurs sont des erreurs « pas trop graves ».

7.4 Analyse linéaire discriminante

Pour plusieurs morceaux, (en particulier dans le premier CD), une grande partie des ereurs identifiées provient de parallèles majeur/ mineur. Par exemple, pour *A Taste of Honey*, on obtient un résulat de seulemnet 46.8% de bonne identification sur les trames, ce qui est notre plus mauvais résultat. Cependant, plus de 40% des erreurs sont dues à des parallèles majeurs/mineurs (Le système trouve EM au lieu de em). On arriverait à un taux de reconnaissance de 68% s'il n'y avait pas ces erreurs.

Nous disposons d'observations à 12 dimensions (vecteurs de chroma) et cherchons à discriminer deux classes (majeur/mineur) en utilisant l'analyse discriminante linéaire dont le principe a été rappelé auparavant.

7.4.1 Application

Dans notre cas, nous pouvons connaître à partir de la transcription le groupe d'appartenance de chaque vecteur de chroma. Nous prenons les morceaux du premier CD et pour chacun calculons le chromagram. Chaque vecteur de chroma est ensuite ramené à C par permutaion circulaire puis classé selon le mode auquel il appartient. Nous construisons ainsi deux matrices, *Mat_chroma_mineur* et *Mat_chroma_Majeur*, qui contiennent respectivemant tous les vecteurs de chroma mineurs et tous les vecteurs de chroma majeurs. (voir figure 15)

Ces matrices sont des descripteurs à 12 dimensions. Nous calculons les nouveaux descripteurs en projetant les descripteurs initiaux sur l'axe discriminant (les nouveaux descripteurs sont à une dimension puisque nous n'avons que deux groupes).

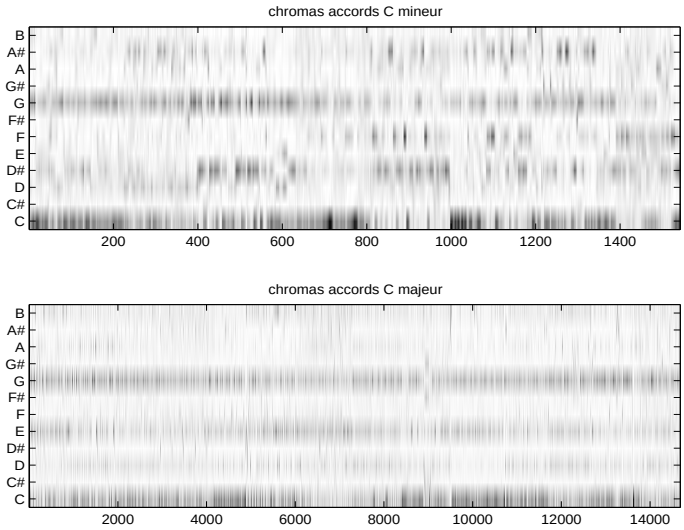

FIG. 15 – Vecteurs de chroma majeurs et mineurs ramenés à C pour le CD *Beatles for Sale*

On calcule ensuite les paramètres gaussiens des nouveaux descripteurs. Nous pouvons calculer pour chacune des deux classes les paramètres de sa fonction de densité de probabilité (c'est à dire le vecteur de moyenne moy_v et la matrice de covariance cov_m) .

Nous obtenons un taux de reconnaissance de 94% pour les accords majeurs et 85% pour les accords mineurs. Il est donc possible *a priori* de discriminer de manière valable les deux classes.

7.4.2 Résultats

Lorsque l'on applique l'analyse discriminante à des morceaux contenant plus de 25% d'erreurs du type parallèles majeur/mineur, le résultat est concluant. En effet, dans ce cas, le nombre d'erreurs instantannées (avant décodage Viterbi) est divisé par deux en moyenne, et le taux de bonne reconnaissance des accords est meilleur en moyenne sur toutes les trames.

L'inconvénient de l'utilisation de l'analyse discriminante dans notre cas est que lorsqu'on l'applique sur l'ensemble du système, les résultats à partir des morceaux contenant peu d'erreurs de type parallèle majeur/mineur sont dégradés, de même que le résultat moyen sur l'ensemble de la base de données. (Le taux de discrimination des accords majeurs/mineurs n'étant pas de 100%, on introduit des erreurs en appliquant l'ALD.) Nous ne pouvons donc pas pour le moment généraliser l'application de l'analyse discriminante. Cependant, nous pensons qu'il y a là une idée intéressante à explorer.

8 Conclusion et perspectives

Sur le plan personnel, ce stage de recherche a été pour moi une riche expérience scientifique. Quatre mois passés au sein de l'équipe analyse/synthèse de l'IRCAM m'ont permis de mieux connaître le traitement du signal audio et surtout de découvrir le domaine de l'indexation musicale.

L'objectif du stage était de construire un système permettant l'extraction automatique d'une suite d'accords à partir de l'analyse d'un signal audio musical. Nous avons présenté l'ensemble du système ainsi que les résultats que nous obtenons. Le système a été implanté sous matlab. Les fonctions codées sont commentées et peuvent ainsi être utilisées facilement. Nous avons également indiqué quels points nous souhaitons développer par la suite (introduction du modèle de Gomez dans la matrice de covariance, utilisation de l'ALD, création de templates à partir d'accords réels...).

Les système n'a été évalué que sur un nombre assez limité d'exemples (cependant, l'intêret de ces exemples est qu'il s'agit de signaux audio polyphoniques complexes). À l'avenir, nous souhaitons l'évaluer sur une base de données beaucoup plus complète, contenant des morceaux de genres et de styles différents.

La détection de suites d'accords n'est pas un problème indépendant. En effet, par exemple, le début et la fin d'un accord sont liés au rythme du morceau ; la suite d'accords peut aider à déterminer la tonalité d'un morceau... C'est pourquoi notre objectif est de poursuivre ce travail par un doctorat dont le but sera de développer un méta-modèle permettant de faire interagir différents estimateurs de paramètres afin d'obtenir une information plus robuste à l'aide du contexte déja extrait.

Références

[1] Takuya Fujishima. Real-time chord recognition of musical sound : A system using common lisp music. *ICMC*, pages 464–467, Beijing, China, 1999.

[2] M.A. Bartsch and G.H. Wakefield. To catch a chorus using chroma-based representations for audio thumbnailing. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 15–18, New Paltz, 2001.

[3] Alexander Sheh and Daniel P.W. Ellis. Chord segmentation and recognition using em-trained hidden markov models. *ISMIR*, Baltimore, MD, 2003.

[4] Christopher A. Harte and Mark B. Sandler. Automatic chord identification using a quantised chromagram. *in AES 118th Convention*, Barcelona, Spain, 2005.

[5] Juan P. Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signal. *ISMIR*, 2005.

[6] S.Pauws. Musical key extraction from audio. *ISMIR*, Barcelona, Spain, 2004.

[7] Geoffroy Peeters. Chroma-based estimation of tonality from audio-signal analysis. *ISMIR*, Victoria, Canada2006.

[8] Emilia Gomez. Tonal description of polyphonic audio for music content processings. *INFORMS Journal on Computing*, February 2004.

[9] Geoffroy Peeters. Musical key estimation of audio signal based on hmm modeling of chroma vectors. *In DAFX, McGill*, Montreal, Canada, September 18-20 2006.

[10] Judith C. Brown. Calculation of a constant q spectral transform. *Acoustical Society of America*, 1990.

[11] Gregory H. Wakefield. Mathematical representation of joint time-chroma distribution. *in SPIE conference on Advanced Signal Procesing Algorithms, Architectures and Implementations IX*, 3807, July Denver, Colorado, 1999.

[12] E. Gomez and P. Herrera. Estimating the tonality of polyphonic audio files : Cognitive versus machine learning modelling strategies. *ISMIR*, pages 92–95, Barcelona, Spain, 2004.

[13] Özgür Izmirli. Template based key finding from audio. *ICMC*, 2005.

[14] L. Rabiner. A tutorial on hidden markov model and selected applications in speech. *IEEE*, 77(2) :257–285, 1989.

[15] B. Gold and N. Morgan. *Speech and audio Signal Processing : Processing and Perception of Speech and Music*. John Wiley & Sons, Inc., 1999.

[16] C.L. Krumhansl. Cognitive foundations of musical pitch. *Oxford University Press*, New Yok, 1990.

[17] Frederic J. Harris. On the use of windows for harmonic analysis with he discreyte fourier transform. *Proceedings of the IEEE*, 66(1), 1978.

[18] Kyogu Lee. Automatic chord recognition using a summary autocorrelation function. *EE391 Special Report*, Spring 2005.