HAL
open science

Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique

Gabriel Sargent, Frédéric Bimbot, Emmanuel Vincent

To cite this version:

Gabriel Sargent, Frédéric Bimbot, Emmanuel Vincent. Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique. Journées d'Informatique Musicale 2010, May 2010, Rennes, France. inria-00551365

HAL Id: inria-00551365

https://inria.hal.science/inria-00551365

Submitted on 3 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UN SYSTÈME DE DÉTECTION DE RUPTURE DE TIMBRE POUR LA DESCRIPTION DE LA STRUCTURE DES MORCEAUX DE MUSIQUE

Gabriel SARGENT
INRIA,
Centre INRIA Rennes
- Bretagne Atlantique
gabriel.sargent@inria.fr

Frédéric BIMBOT
CNRS, IRISA
(- UMR 6074)
frederic.bimbot@irisa.fr

Emmanuel VINCENT
INRIA,
Centre INRIA Rennes
- Bretagne Atlantique
emmanuel.vincent@inria.fr

RÉSUMÉ

Cet article présente une chaîne d'inférence automatique d'éléments de la structure musicale de morceaux de musique sous leur forme sonore. Le système, inspiré de l'état de l'art, est constitué de deux étapes : la segmentation du morceau et l'étiquetage des segments trouvés. Deux contributions sont proposées : un critère de détection des ruptures du timbre des morceaux, initialement proposé pour la séparation entre parole et musique dans un flux audio, est introduit lors de la segmentation. On propose par ailleurs de réaliser l'étiquetage en effectuant un regroupement hiérarchique des segments selon leur timbre modélisé par une gaussienne et en ayant recours à des métriques empruntées au traitement de la parole. Cette étude se termine sur les premiers résultats d'évaluation de cette chaîne en utilisant un corpus de 20 morceaux de musique.

1. INTRODUCTION

1.1. Contexte

Les techniques actuelles de compression sonore facilitent la diffusion de la musique (lecteurs mp3, services Internet d'écoute de musique en ligne. . .), ce qui conduit à la constitution de bases de morceaux de plus en plus volumineuses. Une indexation efficace est nécessaire afin de conserver un accès facile et rapide à ces « données ». Ce besoin a conduit au développement du domaine de la recherche d'information dans les morceaux de musique, qui a pour but la description et l'organisation automatique des contenus musicaux.

L'analyse de la structure des morceaux en terme de couplets, refrains, ponts... entre dans le cadre de la description haut niveau du contenu musical et possède plusieurs applications intéressantes. L'identification des parties répétées, qui sont en général les plus caractéristiques des chansons, permet la génération de résumés sonores. En production musicale, le travail de l'ingénieur du son est susceptible d'être facilité par un accès précis et rapide aux parties musicales à mixer. Enfin, dans le cadre de la musicologie, il peut être utile de formuler des caractéristiques relatives à un style musical par l'analyse automatique d'un grand nombre de chansons associées à ce style (analyse à grande échelle).

Le travail présenté dans cet article correspond au développement d'un système de référence réalisé dans le cadre d'un mémoire de Master 2 (spécialité Systèmes Intelligents et Communicants).

1.2. Définition de la tâche de structuration

Du point de vue des Sciences de l'Ingénieur, on peut décrire un morceau de musique comme étant l'émission d'un ensemble d'événements sonores agencés de manière cohérente dans le temps. La musique tonale occidentale, qui constitue le matériau expérimental dans ce travail, peut être décrite suivant différentes « strates musicales » : la tonalité, le tempo, le timbre, le rythme, l'harmonie, la mélodie et les paroles. On fait l'hypothèse qu'il existe une description de la structure d'un morceau de musique rendant compte de groupements logiques sur les différentes strates décrivant les morceaux : couplets, refrains, ponts ou tout autre découpage s'appuyant sur les similarités et répétitions à moyen terme dans le morceau.

Le but du système présenté dans cet article est de rechercher ces éléments structurels sur la strate du timbre du morceau. On désignera par « segmentation timbrale » l'opération consistant à déterminer les parties homogènes du point de vue du timbre. Celles-ci permettent la description partielle de la structure d'un morceau à plus ou moins long terme. Une description à court terme considère les notes jouées et l'on peut passer à des groupes hiérarchiques de notes plus conséquents en regroupant ces notes par motifs cohérents (mélodies, séquences harmoniques) ou selon les mesures musicales.

Lorsque l'on s'intéresse à une échelle à moyen terme, la notion de structure musicale devient incertaine, même si certaines parties sont bien identifiées. La définition précise met généralement en jeu les différentes strates. La structure des morceaux n'étant pas explicitement communiquée par les compositeurs (si tant est qu'ils l'aient consciemment fixée), il est possible de rencontrer des descriptions différentes du point de vue de la taille des segments, de leurs bornes ou encore de leur étiquette (voir à ce sujet le travail de Peiszer *et al.* dans [19]). Ceci est dû au choix des strates musicales supposées pertinentes pour un morceau donné, qui peut varier d'un auditeur à l'autre. Il n'y a ainsi pas de réel consensus sur ce sujet et l'on a

choisi ici de s'intéresser prioritairement au timbre en remarquant que ses ruptures coïncidaient souvent avec les frontières des segments structurels (couplets, refrains...).

La partie 2 de cet article décrit brièvement les différentes méthodes de détection de la structure dans l'état de l'art, la partie 3 traite de la chaîne de structuration qui a été mise en place, constituée de deux sous-tâches : la segmentation et l'étiquetage. La partie 4 présente les premiers résultats obtenus sur le corpus de 20 morceaux : une moitié est utilisée pour l'apprentissage de ses paramètres et l'autre moitié pour le test, puis les deux moitiés sont interverties. Enfin, un premier diagnostic est établi suite à l'évaluation de différentes versions du système proposé.

2. PANORAMA DE L'ÉTAT DE L'ART

Les tâches de segmentation et d'étiquetage, réalisées successivement ou conjointement selon les méthodes, s'effectuent par l'intermédiaire de descripteurs extraits du signal musical à intervalles réguliers. Les plus populaires sont les coefficients cepstraux à l'échelle Mel et les vecteurs de chroma [4].

2.1. Descripteurs

Les coefficients cepstraux à l'échelle Mel, que l'on notera ci-après MFCC (pour *Mel Frequency Cepstral Coefficients*), sont des descripteurs de timbre et sont interprétables en terme d'instrumentation. Les 20 premiers coefficients sont en général retenus car ils contiennent l'essentiel de l'énergie du spectre du signal étudié : les vecteurs MFCC sont donc de taille 20 [4]. L'échelle Mel reflète certaines caractéristiques perceptives de l'oreille humaine.

Le vecteur de chroma est un descripteur basé sur le tempérament égal de la musique occidentale. Pour l'obtenir, le spectre du signal obtenu à un instant donné est découpé en intervalles fréquentiels de la largeur d'un demi-ton, puis ramené à une seule octave (constituée de 12 demi-tons). Le vecteur obtenu est ainsi de taille 12, chaque coefficient correspondant à l'amplitude d'un demi-ton, tous octaves confondus - pour plus de détails, voir [2, 7].

Pour procéder à l'extraction de ces descripteurs, le signal est préalablement découpé en trames régulières de quelques dizaines de millisecondes [6]. Jehan a montré dans [10] qu'étendre les bornes des trames aux pulsations musicales (une échelle de description naturelle des morceaux) pouvait améliorer la précision des frontières détectées.

2.2. Segmentation et étiquetage

Deux approches dominantes se sont développées pour l'extraction automatique de la structure haut niveau : celle ayant recours aux Modèles de Markov à états Cachés (MMC) et celle utilisant les matrices de similarité.

2.2.1. Modèles de Markov à états Cachés

Le Modèle de Markov à états Cachés est un modèle statistique permettant la reconnaissance dynamique et robuste de motifs spécifiques (suite de phonèmes formant des mots et portions de phrases, mélodies, suite d'accords...). Une suite d'observations est supposée être issue d'un modèle constitué de p états inaccessibles, dit cachés. Ce modèle suit un processus markovien tel que la probabilité d'être dans un certain état, à un instant donné, ne dépend que de l'état précédent. Dans notre contexte, les observations sont les descripteurs extraits du signal musical et les p états cachés représentent les étiquettes des segments structurels recherchés. Le problème de structuration revient à estimer le modèle par apprentissage de ses paramètres sur un ensemble de chansons (« étape d'apprentissage »), puis à estimer quelle est la suite d'étiquettes pouvant le mieux représenter les descripteurs (« étape de décodage », réalisée par un algorithme de Viterbi). Pour plus de détails sur les MMC, se référer à [21].

Logan et Chu utilisent une description timbrale (MFCC) du signal musical, découpé en trames régulières de quelques dizaines de millisecondes [14]. Ils proposent une étude comparant les performances d'un algorithme de regroupement hiérarchique de trames voisines de timbre homogène, et un MMC constitué d'un nombre d'états comparable au nombre de segments structurels recherchés. Abdallah et al. utilisent les MMC afin d'associer un premier état - une première étiquette - à chacune des trames du signal [1]. Les auteurs décrivent ensuite le signal par des histogrammes en effectuant des groupes de 15 trames voisines. Un clustering des histogrammes est effectué via des mesures de distances empiriques (telles la distance cosinus ou la version symétrisée de la divergence de Kullback-Leibler), ou par une approche de type « maximum de vraisemblance ». Rhodes et al. introduisent un terme de contrôle de la taille des segments trouvés dans la précédente fonction de coût [22]. Levy et al. proposent de leur côté d'effectuer un clustering par l'algorithme des K-Moyennes floues [13].

2.2.2. Matrices de similarité

Les matrices de similarité offrent une représentation visuelle de la comparaison des descripteurs de toutes les trames du signal entre elles. Un exemple est donné à la figure 1. C'est Foote [6] qui en proposa initialement l'utilisation pour la tâche de structuration en utilisant une description du timbre des morceaux : les segments structurels à identifier sont caractérisés par des zones rectangulaires d'une texture spécifique, lorsque l'on visualise la matrice. La position des bords de ces zones permet de localiser les instants de séparation entre les parties recherchées. Ceci est facilité par le calcul d'une fonction de « nouveauté timbrale » obtenue par corrélation de la matrice avec un noyau en damier, dont les pics correspondent à des ruptures de timbre. Foote et Cooper proposent aussi d'obtenir les segments structurels par clustering spectral sur la

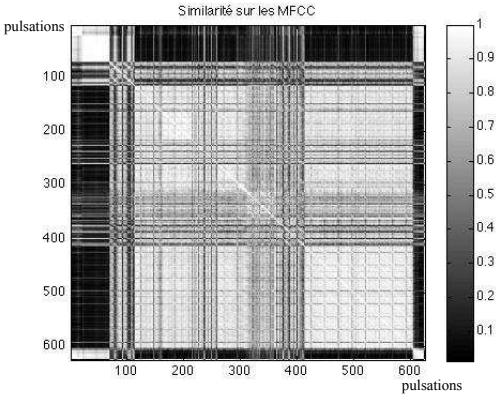

Figure 1. Matrice de similarité calculée sur les MFCC extraits de la chanson « Caribbean Blue » de Enya.

matrice de similarité, puis de modéliser les segments obtenus par des modèles gaussiens comparés par la divergence de Kullback-Leibler [5]. Une autre utilisation possible de la matrice passe par l'utilisation des vecteurs de chroma : il s'agit alors de détecter les séquences de similarité élevées entre trames, visualisables sur la matrice par des bandes sombres, sur ses sous-diagonales. Ces séquences correspondent à la répétition de motifs harmoniques joués lors du morceau analysé. Plusieurs méthodes sont utilisées afin de détecter ces bandes. Lu procède par opérations morphologiques 2D [15]. Goto utilise une matrice temps-décalage dans le cadre précis de la détection de refrains [7]. Shiu propose un noyau de corrélation pour étudier les séquences harmoniques à court terme [24].

2.2.3. *Approches mixtes*

D'autres méthodes empruntent aux deux approches, comme Peeters [18], qui s'appuie sur la similarité timbrale entre trames voisines pour effectuer une première segmentation du signal, puis utilise les K-Moyennes et l'apprentissage d'un MMC décodé par l'algorithme de Viterbi afin d'obtenir la meilleure séquence de segments étiquetés décrivant les morceaux. Jensen représente l'ensemble des différentes hypothèses de segmentation possibles par un graphe acyclique orienté et ramène le problème à la recherche du chemin de moindre coût le traversant. Il effectue cette étude sur 3 descripteurs différents (rythme, timbre, harmonie) et remarque que la segmentation issue des descripteurs de timbre est la plus proche de sa segmentation de référence [11]. Paulus et Klapuri, dans [16], segmentent le signal musical par l'utilisation de la fonction de « nouveauté timbrale » de Foote et recherchent le meilleur étiquetage des segments par une fonction de coût utilisant leur description harmonique.

2.2.4. *Récapitulation*

Des diverses méthodes proposées dans la littérature, il n'existe pas à notre connaissance d'étude comparative permettant de les départager.

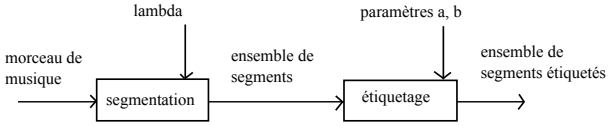

Figure 2. Schéma fonctionnel de la chaîne de structuration. λ est le seuil de détection de rupture timbral (voir 3.1), a et b sont des paramètres d'ajustement de l'étiquetage (introduits en 3.2.3).

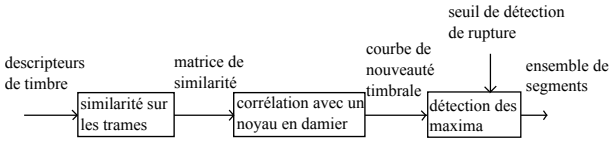

Figure 3. Etape de segmentation

Dans le travail présenté ici, nous nous concentrons sur une approche utilisant des matrices de similarité qui servira de méthode de référence à laquelle nous pourrons comparer par la suite d'autres approches que nous développerons ultérieurement.

3. SYSTÈME DE SEGMENTATION TIMBRALE MIS EN PLACE

L'organisation globale du système de traitement étudié dans ce travail est présenté dans la figure 2. Il est composé d'une étape de segmentation et d'une étape d'étiquetage. Celles-ci font intervenir 3 paramètres à régler, qui peuvent être « appris » sur un corpus d'apprentissage .

Les descripteurs choisis pour l'analyse, les MFCC et vecteurs de chroma, sont exprimés à l'échelle des pulsations des morceaux étudiés.

3.1. Segmentation

La méthode de segmentation utilisée est inspirée de celle de Foote [6] et utilisée sur les 20 premiers coefficients MFCC. La figure 3 rend compte des différentes étapes la composant : similarité sur les trames, puis corrélation avec un noyau en damier et enfin détection des maxima, dont le seuil est déterminé par le paramètre λ. Les trames utilisées sont des fenêtres temporelles centrées sur les pulsations rythmiques. Pour chaque pulsation, la taille de la fenêtre associée est la moitié de la durée entre la pulsation qui la précède et celle qui la suit.

La matrice de similarité, notée A par la suite, est construite en comparant les MFCC des trames du morceau. Les coefficients de A sont obtenus par la formule suivante :

$$[A]_{i,j} = 0.5 + 0.5 \frac{\langle v_i, v_j \rangle}{\|v_i\| \, \|v_j\|} \tag{1}$$

où v_i et v_j sont les descripteurs associés aux trames i et j, $\langle ., . \rangle$ et $\|.\|$ représentent respectivement le produit scalaire et la norme euclidienne des vecteurs considérés.

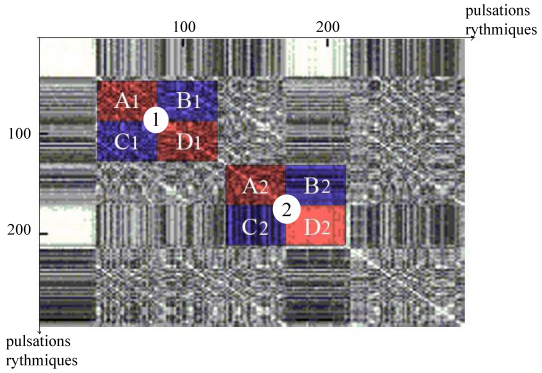

Figure 4. Exemples de position du noyau de corrélation le long de la diagonale de la matrice de similarité (support audio : extrait de Karma Police de Radiohead). La position 1 correspond à une faible valeur de nouveauté (l'instant sur lequel est centré le noyau est contenu dans une zone de texture régulière), la position 2 correspond à un pic de nouveauté, l'instant sur lequel est centré le noyau étant le point de séparation de deux zones de textures différentes.

La courbe de nouveauté timbrale est ensuite calculée par corrélation de la matrice avec un noyau « en damier ». Par exemple un tel noyau, de taille égale à 4, est tel que :

$$K = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \tag{2}$$

Ce type de noyau permet de comparer les valeurs des similarités contenues dans les zones passée et future à l'instant considéré (appelées « *similarités propres* », zones A1-2 et D1-2 de la figure 4), avec les zones de similarité comparant les trames passées avec les trames futures (ou « *similarités croisées* », zones B1-2 et C1-2 de la figure 4). L'instant considéré est celui sur lequel le noyau est centré : localisé sur la diagonale, il représente la frontière entre deux trames voisines.

Formellement, si l'on note N la courbe de nouveauté et K le noyau utilisé, on a :

$$N(i) = \sum_{m=-L/2}^{L/2} \sum_{n=-L/2}^{L/2} K(m,n)\, A(i+m, i+n) \tag{3}$$

Si la zone dans laquelle se trouve un instant t est homogène à l'échelle du noyau, les *similarités propres* et *similarités croisées* sont proches en valeurs. Ainsi la différence des deux types de zones donnera une valeur de nouveauté proche de 0 (exemple à l'instant 1 de la figure), tandis que si l'instant considéré se trouve à la frontière de deux types de textures différentes, on obtient un pic de nouveauté (instant 2 de la figure). On utilise un noyau en damier de forme gaussienne dont l'allure est donnée à la figure 5.

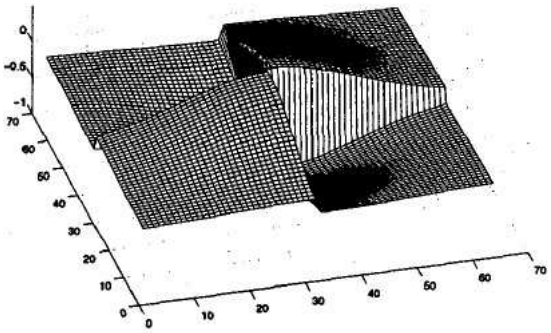

Figure 5. Allure du noyau en damier de forme gaussienne proposé par Foote, extrait de [6].

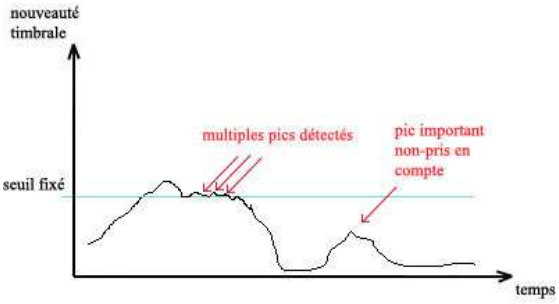

Figure 6. Exemple de détection de maxima par rapport à un seuil fixé.

On est maintenant amené à détecter les maxima de la courbe obtenue, qui représentent les variations importantes du timbre. On les identifie aux frontières des segments structurels recherchés.

Pour détecter les « pics de nouveauté » pertinents, on pourrait se baser sur le seuillage de l'ensemble des maxima locaux détectés [16]. Cette méthode peut cependant conduire à un grand nombre de fausses détections, notamment lorsque la courbe fluctue au voisinage de ce seuil et n'aboutit pas à la détection des maxima les plus pertinents pour la détection de frontières entre segments structurels (voir par exemple la figure 6) : on recherche ici des segments de grande taille. Il est nécessaire d'utiliser un critère qui permet de mettre en valeur les pics de nouveauté qui dominent localement leurs voisins.

Seck *et al.* proposent un tel critère dans [23], basé sur l'analyse d'un indice de présence de rupture statistique. Cet indice, que l'on notera I, mesure la similarité entre les zones passée et future au voisinage de chaque trame du signal (elle est ainsi associable à la courbe de nouveauté timbrale vue précédemment). Le critère de décision C est calculé de manière à prendre la valeur 0 pour tous les instants qui ne sont pas des maxima locaux de l'indice de rupture, et des valeurs plus ou moins élevées suivant l'importance relative du pic par rapport à son contexte.

La méthode de calcul de C est la suivante : on cherche, à partir de chaque instant t, les premiers instants voisins $\tau_1(t) < t$ et $\tau_2(t) > t$ vérifiant $I(\tau_1(t)) > I(t)$ et $I(\tau_2(t)) > I(t)$. On recherche ensuite les quantités :

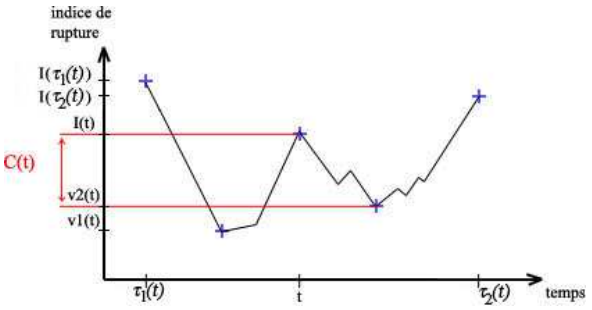

Figure 7. Obtention du critère proposé par Seck *et al.* par l'analyse de l'indice de présence de rupture, associable à une courbe de nouveauté.

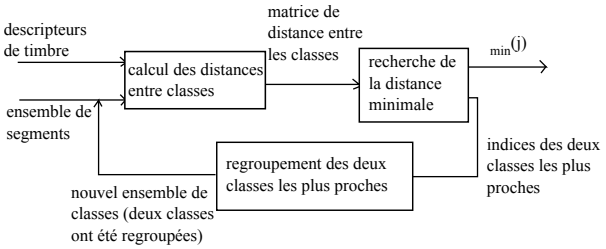

Figure 8. Description du clustering hiérarchique ($\mu_{\min}(j)$ est la distance minimale entre deux classes à l'étape j)

$v_1(t) = \min_{\tau_1(t)<i<t} I(i)$ et $v_2(t) = \min_{t<i<\tau_2(t)} I(i)$ afin de calculer :

$$u(t) = \max(v_1(t), v_2(t)) \tag{4}$$

Le critère de présence de rupture est défini par :

$$C(t) = I(t) - u(t) \tag{5}$$

(voir figure 7). Enfin la décision de la présence de rupture à l'instant t est prise par comparaison à un seuil fixe λ : il y a rupture statistique si $C(t) \geq \lambda$. Ce seuil peut être réglé par apprentissage sur un corpus de morceaux de musique annotés. Dans nos expérimentations initiales, l'utilisation de ce critère s'est avéré être un facteur de robustesse.

3.2. Étiquetage

La partie précédente permet d'obtenir un ensemble de segments dont le contenu timbral est supposé homogène. On cherche maintenant à les associer entre eux en comparant leur timbre : deux segments similaires de ce point de vue sont rassemblés dans une même « classe de segments » désignée par un numéro. Ce numéro correspond à leur étiquette. On utilise un clustering hiérarchique pour y parvenir ; le processus est décrit à la figure 8.

3.2.1. Clustering hiérarchique

Les regroupements hiérarchiques sont effectués comme suit.

Initialement, chaque segment est caractérisé par une gaussienne et est contenu dans une classe différente. A chaque itération, on calcule la distance entre toutes les classes deux à deux. Les deux classes de distance minimale sont regroupées : les descripteurs des segments qu'elles contiennent sont regroupés et l'on cherche le modèle gaussien correspondant à cette nouvelle classe. Pour ne pas avoir à calculer les moments à partir des descripteurs regroupés différemment, on utilise les moyennes et matrices de covariances déjà calculées.

Si l'on fusionne deux gaussiennes modélisant les classes S_x et S_y, de nombres d'échantillons de dimension p, de moyennes, et de matrices de covariance respectives $(n_x, \bar{x}, X)$ et $(n_y, \bar{y}, Y)$, on obtient une gaussienne modélisant une classe S_z ayant les paramètres suivants :

- nombre d'échantillons

$$n_z = n_x + n_y \tag{6}$$

- moyenne :

$$\bar{z} = \frac{n_x\bar{x} + n_y\bar{y}}{n_x + n_y} \tag{7}$$

- matrice de covariance :

$$Z = \frac{n_x\left(X + \bar{x}\bar{x}^T\right) + n_y\left(Y + \bar{y}\bar{y}^T\right)}{n_x + n_y} - \bar{z}\bar{z}^T \tag{8}$$

Il est ainsi possible de comparer le modèle de cette nouvelle classe avec les classes restantes.

Le procédé est répété jusqu'à satisfaire un certain critère d'arrêt.

3.2.2. Comparaison des segments : modèle gaussien de leur contenu timbral

On fait l'hypothèse que la distribution des MFCC de chaque segment suit une loi gaussienne. Cette approche diffère de celle de [5] au niveau de la métrique de comparaison des modèles correspondants.

La mesure de vraisemblance gaussienne symétrisée utilisée dans cette étude a été introduite dans le cadre de l'identification du locuteur dans des enregistrements numérisés [3]. Elle est définie ci-dessous.

Pour simplifier les prochaines formules, on pose :

$$\delta = \bar{y} - \bar{x}, \Gamma = X^{-\frac{1}{2}} Y X^{-\frac{1}{2}}, \rho = \frac{n_y}{n_x} \tag{9}$$

On considère par la suite que le modèle de S_x est le modèle de référence et celui de S_y est le modèle estimé.

Notons $\{y_t\}_{1\leq t\leq n_y}$ les descripteurs de S_y. La vraisemblance que y_t ait été généré par le modèle de S_x s'écrit :

$$L(y_t|\bar{x}, X) = \frac{\exp\left(-\frac{1}{2}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})\right)}{(2\pi)^{\frac{p}{2}}(\det(X))^{\frac{1}{2}}} \tag{10}$$

Sous l'hypothèse d'indépendance *a priori* de ses éléments, on définit la log-vraisemblance moyenne comme suit :

$$\bar{l}\left(y_{1\ldots n_y}|\bar{x}, X\right) = \frac{1}{n_y}\log\left(L\left(y_{1\ldots n_y}|\bar{x}, X\right)\right) \tag{11}$$

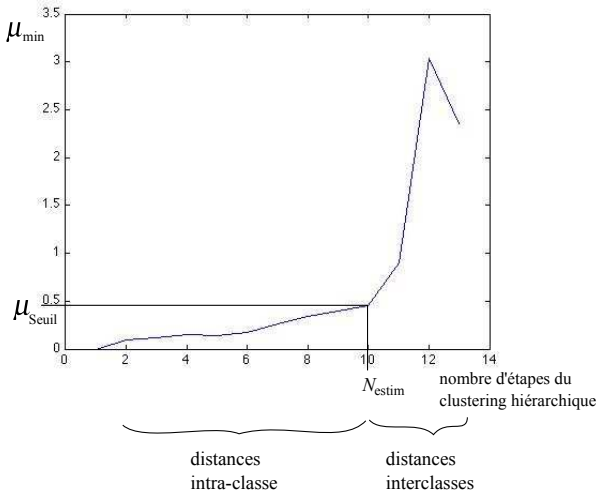

Figure 9. Evolution de $\mu_{\min}$ en fonction du nombre d'étapes du clustering hiérarchique

La mesure de vraisemblance gaussienne est définie par :

$$\mu\left(S_y|\bar{x},X\right)=\frac{1}{p}\left[\text{tr}\Gamma-\log\left(\det\left(\Gamma\right)\right)+\delta^T X^{-1}\delta\right]-1 \tag{12}$$

et [3] montre que :

$$\arg\max_{(\bar{x},X)}\left(\bar{l}\left(y_{1\ldots n_y}|\bar{x},X\right)\right)=\arg\min_{(\bar{x},X)}\left(\mu\left(S_y|\bar{x},X\right)\right) \tag{13}$$

La recherche du paramètre de gaussienne $(\bar{x},X)$ maximisant la probabilité que le modèle ait produit les descripteurs de S_y revient à rechercher $(\bar{x},X)$ qui minimise la mesure de vraisemblance gaussienne. Cette mesure, asymétrique, est positive et telle que $\mu\left(S_y|\bar{x},X\right)=0$ si les éléments de S_y ont été générés par le modèle gaussien de paramètre $(\bar{x},X)$. Cette mesure doit être symétrisée pour pouvoir être interprétable en terme de distance. Pour cela, on effectue une moyenne de la mesure avec sa mesure duale :

$$\mu_{[0.5]}\left(S_x,S_y\right)=\frac{1}{2}\mu\left(S_x|\bar{y},Y\right)+\frac{1}{2}\mu\left(S_y|\bar{x},X\right) \tag{14}$$

3.2.3. Estimation du critère d'arrêt

On considère l'ensemble des « distances cumulées » $\mu_{\min}=\{\mu_{\min}(j)\}_j$, avec $\mu_{\min}(j)$ la distance minimale liant deux classes ou groupes de classes à l'étape j du clustering hiérarchique. On suppose que cet ensemble est séparable en deux classes : la classe des distances « intra-classe », distances de faible valeur liant deux classes (ou groupe de classes) appartenant à la même étiquette, et la classe des distances « interclasses », de valeurs élevées, liant deux segments (ou groupe de segments) appartenant à deux étiquettes différentes. Ceci est illustré par la figure 9.

On partitionne une première fois ces distances selon ces deux classes via l'algorithme des K-Moyennes (avec $K=2$).

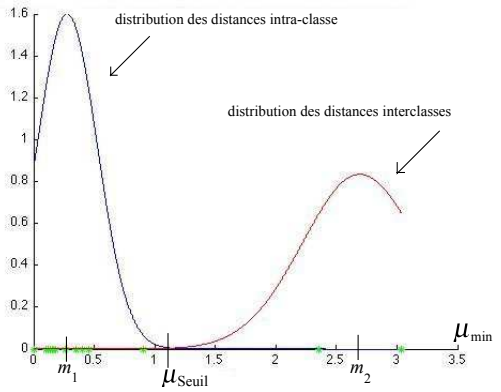

Figure 10. Modélisation gaussienne des classes de distance « intra-classe » et « interclasses », de moyennes respectives m_1 et m_2.

On modélise ensuite chaque classe obtenue par l'intermédiaire d'une gaussienne, comme l'illustre la figure 10. La distance seuil, notée μ_{seuil}, est estimée comme étant la distance correspondant au « croisement des modèles gaussiens ». Cette distance sépare plus finement les distances « intra-classe » des distances « interclasses ». Le nombre de regroupements hiérarchiques est estimé par le nombre de distances « intra-classe » et noté par la suite N_{estim}. L'ensemble des distances cumulées possède autant d'éléments que le nombre maximum de regroupements hiérarchiques pouvant être fait. Le nombre de segments des chansons étant de l'ordre d'une douzaine, la classe des distances « interclasses » peut ne contenir qu'un seul élément ; dans ce cas on ne considère que le regroupement par les K-Moyennes, sans passer par la détermination de la distance seuil.

On suppose enfin que si N_{estim} et le nombre optimal de regroupements N_{opt} ne sont pas égaux, ils peuvent être linéairement dépendants. On introduit ainsi deux paramètres a et b tels que :

$$N_{\text{opt}}=E\left(a*N_{\text{estim}}+b\right) \tag{15}$$

$E\left(.\right)$ représente la fonction partie entière. Les valeurs de a et b sont réglées par apprentissage sur un corpus de morceaux.

4. ÉVALUATION

4.1. Corpus d'étude

Un premier corpus d'expérimentation, que l'on note R, a permis la construction de la chaîne de structuration. Il est constitué de cinq morceaux de différents styles, issus de la base *Real World Computing (RWC) Music database*, qui est une base de données musicale libre de droits créée pour la recherche et proposée par le *RWC partnership* au Japon [1] comme base commune de travail. Les références des morceaux sont rassemblées au sein du tableau

1 . http ://staff.aist.go.jp/m.goto/RWC-MDB/ et [8, 9]

nom de la chanson	auteur	style
Jinsei konnamono	Fevers	pop
Once in a lifetime	Shinya Iguchi	pop
Everyday lovin'	Takeshi Wada	rock
Woman	Suguru Kawagoshi	heavy-metal
Life gage	Kazuo Nishi	house

Table 1. Référence des morceaux ayant servi de support pour le développement de la chaîne de segmentation.

1 ; les dénominations des morceaux choisis pour la suite de l'étude provient de la section et du numéro qui leur sont attribués dans la base RWC. Les créateurs de la base proposent leurs annotations sur leur site Internet, incluant des annotations structurelles [2] .

Un deuxième corpus, noté Q, est constitué d'une liste de 20 morceaux de styles variés (pop, rock, techno) et de leurs annotations fournies par l'IRCAM. On nommera par la suite ces annotations « annotations de référence ». La liste est proposée dans le cadre de la tâche « Evaluation of music structuring and summarizing » du projet QUAERO [3] . Une première moitié (Q1) de ce corpus est utilisée pour l'apprentissage des paramètres de l'algorithme, la deuxième moitié (Q2) constituant le corpus de test de la chaîne de structuration ; le procédé est repris en échangeant le rôle des deux moitiés.

4.2. Mesures de performance

La segmentation est évaluée par des mesures de précision, de rappel et de F-mesure. La précision est le rapport du nombre de frontières correctement estimées au nombre total de frontières estimées, et le rappel est le rapport du nombre de frontières correctement estimées sur le nombre total de frontières de la vérité terrain ; une frontière estimée est « correcte » avec une tolérance de 3 secondes (non-fixé par l'évaluation, mais fréquemment trouvé dans la littérature [12]), c'est-à-dire que l'écart maximum entre la frontière estimée et celle des annotations de référence qu'on lui associe ne dépasse pas cette durée. La F-mesure est une moyenne harmonique de la précision et du rappel. Cette quantité se détériore lorsque l'un des deux critères diminue ; en particulier elle suit le comportement du plus faible des deux dans le cas où l'un est négligeable devant l'autre.

L'évaluation globale de la structure inférée (segments étiquetés) est effectuée par le calcul de l'erreur de modélisation [17, 20]. Celle-ci effectue une mise en correspondance entre les annotations et les estimations.

Notons T_e^j l'ensemble des segments temporels estimés étiquetés j et T_a^i l'ensemble des segments temporels annotés étiquetés i. Le nombre d'étiquettes utilisées, noté J pour l'estimation et I pour l'annotation, peut être différent. On cherche la meilleure association entre les segments estimés j et annotés i, c'est à dire les couples (i, j') tel que le recouvrement entre T_a^i et $T_e^{j'}$ soit maximisé. On a ainsi $j'(i) = \arg\max_j \left(T_e^j \cap T_a^i\right)$. Une étiquette annotée j ou estimée i ne peut être associée qu'une fois ou ne pas être associée. Après appariement, l'erreur de modélisation est définie comme la partie du signal non-correctement assignée (ou non-assignée) et prend des valeurs comprises entre 0 et 1. Pour l'évaluation de la performance on s'intéresse à la valeur du « score de modélisation », défini par :

$$\text{score} = 1 - \sum_{T_e^{j'}} \sum_{T_a^i} |T_e^{j'} \cap T_a^i| / N \qquad (16)$$

où N est la durée totale des annotations et où $|.|$ représente le cardinal de l'ensemble considéré.

Cette quantité tend vers 1 lorsque les annotations et les estimations sont très proches.

4.3. Extraction des descripteurs

Les MFCC sont tout d'abord extraits sur des trames de taille 23 ms qui ne se recouvrent pas, puis ramenés à l'échelle des pulsations rythmiques de la manière décrite en 3.1. Le calcul des descripteurs et l'estimation des pulsations sont faits par le logiciel Sonic Visualiser 1.7.1 muni des plugins VAMP 1.6 développés au *Centre for digital music, Queen Mary University of London* [4] .

4.4. Évaluation de la segmentation

La taille du noyau de convolution est fixée à $L = 40$ pulsations, soit 16 s en estimant la durée moyenne entre deux pulsations à 400 ms.

Le paramètre λ fixant le seuil de détection de rupture de timbre est fixé de la manière suivante.

La segmentation est effectuée sur le corpus d'apprentissage : on calcule la F-mesure pour chaque morceau et pour chaque valeur de λ qui varie de 5 à 15 avec un pas de 0.1 (une étude préliminaire sur R a permis d'estimer grossièrement le seuil optimal à 10). La valeur de λ utilisée pour le test est celle qui maximise la F-mesure moyenne sur le corpus d'apprentissage.

Le tableau 2 présente les résultats obtenus pour deux expériences : il s'agit des mesures moyennes de précision, rappel, F-mesure sur les corpus de test.

Dans la première expérience, le corpus R est utilisé à la fois en tant que corpus d'apprentissage et de test. La phase d'apprentissage a permis de fixer $\lambda = 9.7$.

Dans la deuxième expérience, le corpus Q1 sert d'ensemble d'apprentissage et Q2 constitue l'ensemble de test. On a ainsi pu fixer $\lambda = 7.5$ sur Q1 et $\lambda = 9.4$ sur Q2.

2 . http ://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/, voir « chorus sections »

3 . Il s'agit d'un programme de recherche et d'innovation franco-allemand visant à développer des technologies de traitement automatique des contenus multimédia et multilingues, afin de proposer de nouveaux produits et services au grand public et aux professionnels. Site internet : www.quaero.org

4 . sites Internet associés : http ://www.sonicvisualiser.org/, http ://www.vamp-plugins.org/

App.	Test	λ	précision	rappel	F-mesure
R	R	9.7	0.6986	0.6748	0.6798
Q1	Q2	7.5	0.5958	0.7034	0.6273
Q2	Q1	9.4	0.5718	0.6353	0.5840

Table 2. Évaluation de la segmentation : mesures moyennes de précision, rappel et F-mesure obtenues sur les différents corpus de l'étude. R : corpus de développement, Q1 et Q2 : moitiés du corpus servant d'apprentissage (« App. ») et de test.

App.	Test	Segmentation	a	b	ScoreMoy
R	R	connue	0.16	1.67	80.24%
R	R	automatique	0.81	-3.63	64.95%
Q1	Q2	connue	0.49	-0.20	79.07%
Q1	Q2	automatique	0.99	-2.85	59.46%
Q2	Q1	connue	0.23	2.17	73.88%
Q2	Q1	automatique	0.03	4.08	56.27%

Table 3. Évaluation de l'étiquetage : a et b déterminés par apprentissage (« App. ») et scores de modélisation moyens (« ScoreMoy ») sur les corpus d'étude.

4.5. Évaluation de l'étiquetage

Les paramètres a et b sont fixés en déterminant pour chaque morceau du corpus d'apprentissage le nombre de regroupements hiérarchiques optimal N_{opt}, maximisant le score de modélisation, et N_{estim} par la méthode décrite en 3.2.3. L'ensemble des couples $\left(N_{\text{opt}}, N_{\text{estim}}\right)$ obtenus est modélisé par une droite de coefficient directeur a et d'ordonnée à l'origine de la droite b (régression linéaire) qui sont les paramètres utilisés pour la phase de test.

L'étiquetage est évalué en deux phases. Le score de modélisation est tout d'abord calculé sur les segments obtenus avec la segmentation idéale connue : les frontières de ces blocs sont celles des annotations de référence. Ceci permet de s'affranchir des performances de l'étape de segmentation. Ce score est ensuite calculé sur les blocs structurels issus de la phase de segmentation automatique, ce qui permet de mesurer la performance de la chaîne complète. Les valeurs des paramètres fixés par apprentissage et les scores de modélisation associés sont présentés dans le tableau 3.

On observe une différence d'en moyenne 17% entre les scores de modélisation calculés avec la segmentation idéale connue et ceux calculés avec la segmentation obtenue automatiquement.

4.6. Evaluation de systèmes annexes et discussion des résultats

4.6.1. Importance du critère de rupture pour la segmentation

Trois variantes du procédé de segmentation proposé sont testées afin d'évaluer l'intérêt de la détection de frontières par seuillage du critère de Seck.

s1					
App.	Test	n_p	précision	rappel	F-mesure
R	R	31	0.5677	0.8269	0.6709
Q1	Q2	38	0.4795	0.9207	0.6168
Q2	Q1	23	0.5217	0.6560	0.5726
s2					
App.	Test	λ	précision	rappel	F-mesure
R	R	14.9	0.7667	0.6615	0.7049
Q1	Q2	11.8	0.5611	0.7841	0.6400
Q2	Q1	12.2	0.5471	0.6507	0.5870
s3					
App.	Test	n_p	précision	rappel	F-mesure
R	R	12	0.6603	0.6748	0.6573
Q1	Q2	20	0.5289	0.7660	0.6105
Q2	Q1	17	0.5454	0.6991	0.5919

Table 4. Évaluation de la performance des systèmes de segmentation s1, s2 et s3.

Les systèmes s1 et s2 se limitent à l'étude de la courbe de nouveauté timbrale. Le système s1 localise les frontières par la détection de tous les pics de nouveauté pour ne garder que les n_p plus grands, tel que n_p maximise la F-mesure moyenne sur le corpus d'apprentissage. Le système s2 ne conserve que les pics dépassant un seuil λ, fixé comme en 4.4. Le système s3 reprend le système s1 mais fixe n_p à partir du critère de Seck.

Les performances obtenues sont référencées dans le tableau 4. Lorsque l'on compare les performances des systèmes s1 et s3, on ne remarque qu'une faible variation des F-mesures moyennes sur les ensembles de test, bien que n_p chute de 14 unités en moyenne d'un système à l'autre. Ceci met en évidence que le critère de Seck parvient à diminuer fortement le nombre de fausses détections et de mettre en valeur les pics de nouveauté pertinents pour notre segmentation. Les performances du système s2 comparées à celles du système évalué en 4.4 ont un comportement similaire, avec néanmoins une légère diminution de la F-mesure moyenne sur ce dernier système. La diminution des fausses détections permet de limiter la présence de segments trop petits par rapport à notre étude à moyen terme de la structure musicale.

Enfin, l'ensemble de ces performances confirme qu'il n'est pas judicieux de fixer le nombre de segments sur un ensemble de morceaux de musique, comparé aux méthodes utilisant le seuillage.

4.6.2. Mesure de vraisemblance gaussienne et rapport de vraisemblance gaussien pour l'étiquetage

Afin de mesurer la pertinence de l'utilisation de la mesure de vraisemblance gaussienne lors du processus d'étiquetage, on utilise cette fois un rapport de vraisemblance gaussien (RVG) en tant que distance entre les classes de segments obtenues au cours du clustering hiérarchique. Le RVG permet de mettre en relation les vraisemblances de 2

App.	Test	Segmentation	a	b	ScoreMoy
R	R	connue	2	-1	87.49%
R	R	automatique	-2	10	72.03%
Q1	Q2	connue	0.58	4.42	75.83%
Q1	Q2	automatique	6.28	-8.11	46.53%
Q2	Q1	connue	0.61	3.77	70.55%
Q2	Q1	automatique	0.22	6.13	53.30%

Table 5. Évaluation de l'étiquetage : cas du rapport de vraisemblance gaussien utilisé en tant que distance entre classes de segments.

modélisations gaussiennes d'un même couple de classes : une modélisation des deux classes par une seule distribution mono-gaussienne, et une modélisation associant à chaque classe une distribution mono-gaussienne différente.

Formellement, si l'on considère les classes S_x et S_y composées respectivement des descripteurs $\{x_t\}_{1\leq t\leq n_x}$ et $\{y_t\}_{1\leq t\leq n_y}$ et modélisées respectivement par les gaussiennes G_x de paramètre θ_x et G_y de paramètre θ_y, on a :

$$\text{RVG}\left(S_x, S_y\right) = \frac{L\left(x_1 ... x_{n_x} y_1 ... y_{n_y} | \theta_z\right)^{n_x+n_y}}{L\left(x_1 ... x_{n_x} | \theta_x\right)^{n_x} L\left(y_1 ... y_{n_y} | \theta_y\right)^{n_y}} \tag{17}$$

θ_z est le paramètre de la gaussienne issue de la fusion de G_1 et G_2 (au sens du paragraphe 3.2.1) et $L\left(x|\theta\right)$ est la vraisemblance des échantillons x étant donné le modèle gaussien de paramètre θ.

On regroupe ainsi à chaque étape du clustering hiérarchique les deux classes dont le RVG est le plus élevé : le couple qui est le plus susceptible d'être modélisé par une distribution mono-gaussienne devient une seule classe de segments. Les scores de modélisation pour les différents ensembles de morceaux sont référencés dans le tableau 5. On constate que l'utilisation de la mesure de vraisemblance gaussienne permet d'obtenir de meilleures performances pour l'étiquetage des segments. Remarquons enfin que les valeurs idéales $a = 1$ et $b = 0$ (pas d'ajustement de N_{estim}) sont rarement atteintes, ce qui signifie qu'une voie d'amélioration possible de la chaîne serait de modéliser plus finement les classes de segments et les classes de distances « interclasses » et « intra-classe », par exemple en ayant recours à des modèles de mélanges gaussiens.

5. CONCLUSION

Cet article a présenté un système de détection de ruptures timbrales pour l'inférence de structure musicale développé au sein de l'équipe METISS. Ce système, inspiré de l'état de l'art, présente quelques nouveautés : un critère de détection de rupture a été introduit lors de l'étape de segmentation ; la comparaison des segments timbraux a été faite par des métriques utilisées en traitement de la parole, que l'on a utilisé afin d'estimer le critère d'arrêt du clustering hiérarchique. On a montré leur intérêt après avoir évalué plusieurs versions du système proposé. La prochaine étape de notre étude consiste à comparer ces résultats avec ceux d'autres algorithmes existants et de mesurer les progrès accomplis par nos travaux par comparaison à ce système de référence.

6. REFERENCES

[1] S. Abdallah, K. Noland, C. M. Sandler, M., and C. Rhodes, "Theory and evaluation of a bayesian music structure extractor" *Proceedings of the ISMIR Conference*, London, UK, pp. 420–425, Sep. 2005.

[2] M. A. Bartsch and G. Wakefield, "To catch a chorus : Chroma-based representations for audio thumbnailing" *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, pp. 15–18, October 2001.

[3] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, Second-Order Statistical Measures for Text-Independant Speaker Identification. In *Speech Communication*, Vol.17, No. 1-2, pp. 177-192, August 1995.

[4] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval : Current directions and future challenges" *Proceedings of the IEEE*, vol. 96, no. 4, April 2008.

[5] M. Cooper and J. Foote, "Media segmentation using self-similarity decom- position" *Proceedings of the SPIE Storage and Retrieval for Multimedia Databases*, San Jose, California, USA, pp. 167–175, January 2003.

[6] J. Foote, "Automatic audio segmentation using a measure of audio novelty" *Proceedings of the IEEE International Conference on Multimedia and Expo*, New York, New York, USA, pp. 452–455, August 2000.

[7] M. Goto, "A chorus-section detecting method for musical audio signals" *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, pp. 437–440, April 2003.

[8] M. Goto, H. Hashiguchi, T. Nishimura et R. Oka, "RWC Music Database : Music Genre Database and Musical Instrument Sound Database" *Proceedings of the ISMIR Conference*, USA, October 2003.

[9] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Data- base : RWC music database : Popular, Classical, and Jazz Music Databases" *Proceedings of the ISMIR Conference*, USA, pp. 287–288, October 2002.

[10] T. Jehan, "Hierarchical multi-class self similarities" *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, USA, October 2005.

[11] K. Jensen, “Multiple scale music segmentation using rhythm, timbre and harmony” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[12] M. Levy and M. Sandler, “Structural segmentation of musical audio by constrained clustering” *IEEE transactions on audio, speech and language processing*, vol. 16, no. 2, pp. 318–326, February 2008.

[13] M. Levy, M. Sandler, and M. A. Casey, “Extraction of high-level musical structure from audio data and its application to thumbnail generation,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, pp. 13–16, May 2006.

[14] B. Logan and S. Chu, “Music summarization using key phrases” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, pp. 749–752, June 2000.

[15] L. Lu, M. Wang, and H.-J. Zhang, “Repeating pattern discovery and struc- ture analysis from acoustic music data” *Proceedings of the Multimedia Information Retrieval Workshop*, New York, New York, USA, pp. 275–282, October 2004.

[16] J. Paulus and A. Klapuri, “Music structure analysis by finding repeated parts,” *Proceedings of AMCMM*, Santa Barbara, California, USA, p. 59-68, October 2006.

[17] G. Peeters, “Sequence representation of music structure using higher-order similarity matrix and maximum likelihood approach” *Proceedings of the ISMIR Conference*, Vienna, Austria, September 2007.

[18] G. Peeters, A. La Burthe, and X. Rodet, “Toward automatic music audio summary generation from signal analysis” *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, pp. 94–100, October 2002.

[19] E. Peiszer, T. Lidy, and A. Rauber, “Automatic audio segmentation : Segment boundary and structure detection in popular music” *Proceedings of the Workshop on Learning Semantics of Audio Signals*, Istanbul, Turkey, 2008.

[20] G. Peeters and E. Vincent, “QUAERO, Task 6.5 : Evaluation of music structuring and summarization, Evaluation Guidelines,” Internal Report, April 2009.

[21] L. R. Rabiner, “A tutorial on hmm and selected applications in speech recognition,” Proceedings of the IEEE Conference, vol. 77, no. 2, pp. 257– 286, February 1989.

[22] C. Rhodes, M. A. Casey, S. Abdallah, and M. Sandler, “A markov-chain monte-carlo approach to musical audio segmentation,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, pp. 797–800, May 2006.

[23] M. Seck, R. Blouet, and F. Bimbot, “The IRISA/ELISA Speaker Detection and Tracking Systems for the NIST’99 Evaluation Campaign”. *Digital Signal Processing*, Volume 10, Issues 1-3, January 2000, Pages 154-171.

[24] Y. Shiu, H. Jeong, and C. C. Jay-Kuo, “Similarity matrix processing for music structure analysis” *Proceedings of AMCMM*, Santa Barbara, Cali- fornia, USA, pp. 69–76, October 2006.