SEGMENTATION AUTOMATIQUE DE CATÉGORIES MUSICALES: ÉTUDE EXPLORATOIRE SUR LE FREE JAZZ

Marie Tahon, Zaher Belghith, Jean-Marc Chouvel, Pierre Michel

SEGMENTATION AUTOMATIQUE DE CATÉGORIES MUSICALES: ÉTUDE EXPLORATOIRE SUR LE FREE JAZZ

Marie Tahon
LIUM - Le Mans Université
marie.tahon@univ-lemans.fr

Jean-Marc Chouvel
IReMus - Paris Sorbonne Université
jeanmarc.chouvel@free.fr

Zaher Belghith
IReMus - Paris Sorbonne Université
GREAM - Université de Strasbourg
zaher.belghith@paris-sorbonne.fr

Pierre Michel
GREAM - Université de Strasbourg
spsm@unistra.fr

Résumé

Le travail présenté ici est le résultat d'un projet collaboratif pluridisciplinaire entre l'informatique et la musicologie. L'objectif est de développer une méthodologie permettant d'assister l'analyse musicale en faisant émerger des catégories. La recherche de catégories musicales dans un signal relève d'un exercice de segmentation et de regroupement en catégories. Pour cela, l'outil se base principalement sur la technique du *clustering spectral*, mais apporte également de nouvelles fonctionnalités telles que la transformée différentielle, l'utilisation de différents ensembles de descripteurs acoustiques, ou la détermination automatique du nombre de catégories présentes dans la structure. L'article précise les choix d'implémentation réalisés lors du développement de l'outil de segmentation et propose une analyse des résultats obtenus sur une œuvre du répertoire *Free Jazz*. Ces travaux ont permis de valider les nouvelles fonctionnalités apportées par les auteurs dans un contexte d'analyse musicale.

1. INTRODUCTION

Au cours des années 1950-1960, le jazz, tout comme les autres musiques de l'époque, se trouvait face à un besoin de renouvellement du langage et des dimensions structurelles musicales (comme la tonalité, le rythme), lui permettant de survivre face aux enjeux socio-culturels ou stylistiques imposés par cette ère moderne. Ce nouveau genre musical est apparu sous le nom de *Free Jazz*. Ekkehard Jost dans son livre référence [12] affirme : "Avec le *Free Jazz*, cependant, plusieurs styles personnels divergents se développent pour un seul point commun : le refus des codes conventionnels". Il ajoute : "Les conventions qui surgirent dans le *Free Jazz* (...) ne purent jamais l'unifier au sens ou pouvaient l'être les styles précédents" en parlant des styles qui s'inscrivent dans le courant jazz comme le bebop, le hard bop ou le cool jazz. Ceci nous a poussé à nous interroger sur le processus de création musicale du *Free Jazz* et la possibilité d'en déduire une structure, manuelle dans un premier temps et assistée par la machine dans un deuxième temps. L'objectif est de développer un outil apportant à l'analyste des éléments qui l'aideront à faire émerger une structure musicale, et ainsi lui permettre d'appréhender le corpus de *Free Jazz* et ses enjeux.

Selon plusieurs témoignages de musiciens tels que Cecil Taylor, l'énergie [1] du groupe, viendrait remplacer le langage traditionnel et conventionnel du jazz avec toutes ses composantes : harmonie, swing, forme, etc. [9]. Dans ce sens, de nombreuses recherches ont porté sur des critères de segmentation en fonction de l'énergie du groupe, ou d'élément saillant appelé par C. Canonne "marqueur formel", et défini comme un moment d'accélération, un changement brusque de registre, l'apparition d'un motif thématique ou d'un nouvel élément de texture sonore [4].

Dans l'hypothèse où ce processus implique une structure rythmique complexe, intégrant le swing sous un nouvel aspect continu et fluctuant, il permet à chaque musicien d'avoir plus de liberté en lui offrant un espace temporel pour faire appel à d'autres idées rythmiques et mélodiques, éléments dynamiques ou de texture, et par la suite, de participer à cette nouvelle construction temporelle [9].

Nous nous sommes appuyés sur ces différents témoignages, pour mettre en place notre méthodologie qui consiste à définir un niveau de structure clair en proposant des catégories précises suivant une logique temporelle. Ces catégories correspondent à des motifs (ou *patterns*) qui se répètent, éventuellement se superposent au cours de l'œuvre musicale. Suivant ce qui a été énoncé précédemment, les motifs peuvent être de type mélodique, rythmique, dynamique ou textural.

1 . Énergie étant à prendre dans le sens "variable du temps" et non "intensité" [12]

Le travail exposé dans cet article est le résultat d'une collaboration pluridisciplinaire entre l'informatique et la musicologie. Ce projet vise à déterminer une structure temporelle dans un style musical qui refuse *a priori* "les codes conventionnels". Ainsi, la recherche de catégories musicales dans une œuvre de *Free Jazz* relève d'un exercice de segmentation et de regroupement (*clustering*) en catégories. Cet article présente une extension [2] de l'outil de partitionnement par *clustering spectral* développé par McFee and Ellis [17] et précise les nouvelles fonctionalités ajoutées.

Du point de vue informatique, la segmentation d'un signal audio est un domaine de recherche très actif. En traitement automatique de la parole, les techniques se rapprochant le plus de l'application visée sont la segmentation et le regroupement en locuteurs [2]. Cependant, les techniques développées pour la parole (et/ou la musique de variété) sont généralement supervisées, c'est-à-dire qu'elles se basent sur l'apprentissage de modèles à partir de grandes quantités de données. Or le corpus que nous avons à disposition est réduit à quelques extraits musicaux. Nous avons donc orienté nos recherches vers des techniques de segmentation non-supervisée, et en particulier le *clustering spectral*. Cette approche est également dans la lignée directe de précédents travaux sur l'analyse computationnelle à l'aide de graphes [1] pour la musique contemporaine. Étant donné que les catégories représentent des concepts cognitifs et gestuels, leur extraction à partir d'une simple représentation acoustique relève du défi. De plus, le travail collaboratif consiste à apporter suffisamment de souplesse à l'outil afin de laisser à l'expert musical toute latitude d'observer les phénomènes étudiés.

Dans un premier temps, nous détaillerons les choix d'implémentation de l'outil de *clustering spectral* proposé pour la segmentation automatique. Dans un deuxième temps, nous présenterons les résultats obtenus sur un extrait de *Free Jazz*. Enfin, nous proposerons quelques pistes de réflexion autour de cette thématique.

2. CLUSTERING SPECTRAL AVEC LAPLACIEN

Pour répondre aux objectifs de ce projet collaboratif musicologie et informatique, nous proposons un outil de segmentation et de regroupement en catégories musicales non supervisé et modulable par l'utilisateur. Cette modularité est donnée par la possibilité de représenter le signal à l'aide de plusieurs types de descripteurs acoustiques, mais aussi de laisser la possibilité soit à la machine, soit à l'analyste, le choix de certains paramètres. Le *clustering spectral* semble approprié à cette tâche de part son appartenance à la théorie des graphes et les résultats que cette technique permet d'obtenir pour l'analyse musicale.

2.1. Principes du clustering spectral

Le clustering est une technique non supervisée largement utilisée pour analyser des données. Cette approche consiste à regrouper des données de telle sorte que des points appartenant à un même groupe partagent des caractéristiques similaires. Le clustering spectral est devenu très populaire du fait que c'est une technique simple à implémenter, mais aussi qu'elle permet d'obtenir de très bonnes performances notamment sur du clustering de graphe. Les résultats obtenus dépassent généralement ceux d'algorithmes plus classiques tels que les k-means.

Le clustering spectral peut se décomposer en trois étapes principales. Dans un premier temps, il faut créer un graphe (ou une matrice) de similarité à partir des données. Ensuite, le calcul des K premiers vecteurs propres de la matrice de Laplacien permet de définir un vecteur acoustique pour chaque groupe. Enfin, l'application d'un algorithme de clustering (par exemple K-means) permet de regrouper nos données dans chaque groupe. L'utilisation de matrices de similarité comme étape intermédiaire au regroupement permet de représenter les données dans un espace spectral adéquat. Le fait que l'espace de description soit optimisé donne un large avantage à cette méthode.

2.2. Analyse musicale et clustering spectral

La plupart des méthodes utilisées pour l'analyse automatique de structures musicales sont basées sur le calcul d'une matrice d'auto-similarité obtenue à partir de descripteurs temporels du signal audio. Ce type de méthode permet de visualiser la structure musicale à partir de similarités extraites localement, c'est-à-dire entre deux trames consécutives [10, 16]. Les descripteurs temporels audio peuvent être soit des descripteurs de timbre (coefficients cepstraux, spectraux, etc.) soit des descripteurs harmoniques (hauteurs) [18].

Le choix de l'espace de représentation acoustique, qu'il soit spectral ou harmonique est un choix *a priori* qui dépendra du type de musique à analyser. L'idée de créer un espace de représentation qui soit adapté au contenu du signal à analyser est tout à fait intéressante. Une première manière de créer un espace de représentation adapté est de projeter les descripteurs temporels avec une analyse en composantes principales (PCA). Outre le fait d'optimiser l'espace de représentation acoustique, ce type d'approche à également l'avantage de réduire sa dimension [14]. Une seconde manière consiste à créer un nouvel espace de représentation à partir du Laplacien de la matrice de similarité obtenues à partir des descripteurs temporels [17].

2 . Disponible en libre accès sur
`git-lium.univ-lemans.fr/tahon/spectral-clustering-music`

2.3. Récurrences locales et répétitions au long-terme

Une des difficultés principales lors de l'analyse d'une structure musicale est le choix d'une temporalité adaptée. En effet, une structure musicale est souvent complexe et combine des éléments apparaissant sur des échelles temporelles différentes. Ainsi, le choix de l'échelle sera déterminant sur le partitionnement final. La plupart des techniques actuelles d'extraction automatique de structure sont basées sur une segmentation bas-niveau en trames d'une dizaine de millisecondes. La segmentation devient un problème de calcul de nouveauté, comme par exemple le calcul de distance [11] ou d'entropie entre deux trames consécutives [15].

La segmentation bas-niveau permet d'extraire les récurrences locales, mais ne prend pas en compte une structure hiérarchique de plus haut niveau. Notamment, les répétitions au long-terme de motifs rythmiques, timbraux ou harmoniques ne sont pas capturés par ce type de segmentation. C'est pourquoi McFee and Ellis [17] ont proposé une approche nouvelle qui combine les récurrences locales (ou *local consistency*) et les répétitions au long-terme. Cette approche est fondée sur l'analyse des vecteurs propres du Laplacien et permet d'obtenir une représentation compacte et multi-échelle.

2.4. Description de la méthode originale

La segmentation par clustering à partir du Laplacien d'une matrice de similarité, proposée par McFee and Ellis [17], vise à extraire automatiquement la structure musicale d'un enregistrement. Alors que la détection de répétitions sur une échelle de temps courte (par exemple la répétition d'accords) est une tâche relativement simple, la combinaison de ces multiples répétitions sur une échelle temporelle plus large est une tâche beaucoup plus complexe. Les auteurs proposent une combinaison pondérée des éléments consistants au niveau local avec une représentation des répétitions à une échelle plus large.

En premier lieu, on considère $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ une matrice temporelle de descripteurs de dimension d, par exemple un spectrogramme ou une séquence de MFCCs. Une matrice de récurrence binaire R est extraite de X telle que :

$$R_{ij} = \begin{cases} 1 \text{ si } x_i, x_j \text{ sont des k plus proches voisins} \\ 0 \text{ sinon} \end{cases} \tag{1}$$

On définit la matrice R^{loc} comme étant une matrice de similarité R (eq. 1) obtenue directement à partir de X, suivi d'une opération de filtrage et de vote majoritaire. Cette matrice capture les récurrences locales de la matrice X.

$$S_{ij} = \exp\left(-\frac{1}{2\sigma^2}|y_i - y_{i+1}|^2\right) \tag{2}$$

On définit également la matrice S^{rec} obtenue à partir d'une matrice S (eq. 2) de similarité entre deux segments temporels qui peuvent être éloignés dans le temps. Cette segmentation est donnée par une détection de segments temporels que nous aborderons plus tard. On combine alors deux matrices de similarités (eq. 3) : l'une caractérise les récurrences locales S^{rep} et l'autre les répétitions à plus long-terme R^{rec} de telle sorte que :

$$A = \mu S^{rec} + (1 - \mu) R^{loc} \tag{3}$$

Avec $0 < \mu < 1$ étant la pondération appliquée aux deux matrices. Pour plus de détail sur l'obtention des différentes matrices de similarité et le calcul de μ, le lecteur se reportera à la publication [17] [3] . Une fois la matrice A créée, on peut calculer L le Laplacien normalisé de A suivant le principe du clustering spectral.

Les K premiers vecteurs propres normalisés de L forment la nouvelle base de représentation acoustique du signal musical. Chacun de ces vecteurs représente une composante structurelle du signal, par exemple, une zone de silence, ou bien l'apparition d'un motif mélodique (voir section 6). Le choix de K influence directement le résultat de la segmentation obtenue. Afin de regrouper les différents segments temporels suivant leur proximité en terme de vecteurs propres, un algorithme classique de clustering est utilisé, par exemple, K-means. On obtient donc une segmentation temporelle où chaque segment est identifié par son appartenance à un des K clusters.

3. ECHELLES TEMPORELLES POUR LES RÉCURRENCES LOCALES.

Dans la méthode de McFee and Ellis les structures de répétitions sont basées sur les pulsations (*beats*). Il s'agit alors d'adapter ces structures pour un répertoire où il n'y a pas de beat. Les segments temporels utilisés pour construire la matrice de similarité S^{rep}, sont définis entre deux pulsations consécutifs. Chaque segment est un vecteur qui représente la moyenne des descripteurs sur la durée totale du segment. En effet, étant donné que l'ensemble d'évaluation des auteurs est un corpus de musique des Beatles, les pulsations sont facilement détectables, et il n'y a pas (ou peu) de changements de tempo. Cette approche doit évidemment être adaptée pour les musiques non mesurées, ou bien des musiques contenant des variations de tempo, comme le *Free Jazz*. C'est pourquoi un ajout fondamental pour notre travail a été de proposer plusieurs types d'échelles temporelles.

3 . librosa.github.io/librosa/auto_examples/plot_segmentation.html

3.1. Extraction des pulsations

L'extraction des pulsations (*beat detection*) est largement utilisée dans le domaine du *Music Information Retrieval* où les corpus musicaux contiennent une large quantité de musique à temps forts marqués (rock, variété, etc.). L'extraction de pulsations sur ce type de musique est fiable et robuste. Ainsi les récurrences locales sont-elles étudiées entre deux temps consécutifs (noires consécutives) a priori tous de même durée.

3.2. Extraction d'onsets

Une détection d'*onset* qui repère automatiquement le début de chaque note, peut également être utilisée pour déterminer les segments temporels. Ce type d'approche est pertinente dans le cas où il y a peu d'instruments qui jouent en même temps. Dans le cas contraire, le nombre d'onsets détectés risque d'être très important et cela entraine une explosion de la dimension des matrices de similarité, et donc des temps de calcul. Cette approche implique que les segments temporels sont de tailles variables. Cela n'a pas d'impact sur les calculs de similarité puisque l'ensemble des matrices sont synchronisées sur les onsets.

3.3. Temporalité fixe

Une solution envisagée pour les musiques non mesurées est de ne pas se baser sur une structure rythmique *a priori* mais d'utiliser une fenêtre de taille fixe. Ainsi tous les segments ont la même durée. Un risque majeur de cette méthode est que si la durée fixée est trop grande, on risque de passer à côté de changements importants dans la structure (car les segments seront moyennés quoi qu'il arrive), si au contraire, elle est trop petite, on risque de capturer une structure musicale très fine sans avoir de vision à plus long terme. De plus, avec des fenêtres temporelles trop faibles, on augmente fortement la dimension des matrices à traiter et donc du temps de traitement.

3.4. Temporalité manuelle

Finalement, la dernière possibilité envisagée, est celle de laisser à un expert le soin de segmenter manuellement le signal musical à analyser. En effet, une des difficultés majeures dans la segmentation automatique de la musique est de déterminer le niveau de structure que l'on cherche à faire émerger. A l'heure actuelle, il semble difficile de développer une solution complètement automatique qui fonctionnerait sur tout type de musique.

4. CHOIX DU NOMBRE DE CLUSTERS

Le choix du nombre de clusters est déterminant sur la structure finale obtenue. Selon la durée du signal et le niveau de structure souhaitée, le nombre de clusters optimal diffère. Deux approches ont été retenues pour définir le nombre de clusters K : manuelle ou automatique.

4.1. K fixé par l'utilisateur

Le nombre de clusters peut être fixé par l'utilisateur. Afin de permettre une meilleure analyse des résultats obtenus, une liste de n_K valeurs possibles pour K peut être entrée manuellement par l'utilisateur. En ce cas, n_k opérations de clustering spectral seront réalisées et l'utilisateur pourra visualiser les n_k segmentations associées.

4.2. K déterminé automatiquement

Il existe plusieurs méthodes non-supervisées permettant d'évaluer si un regroupement est bon ou pas. En première approche, on peut analyser les valeurs propres ou les vecteurs propres [22] afin de déterminer le nombre optimal de groupes.

Un autre approche consiste à tester plusieurs valeurs de K et déterminer un critère qui optimise K. Les critères à l'état de l'art sont tous basés sur la même idée : un regroupement est correct si tous les groupes sont compacts et qu'ils sont suffisamment éloignés les uns des autres. L'indice de *silhouette* [20] (eq. 4) cherche à minimiser la distance moyenne entre l'échantillon et tous les autres points appartenant au même groupe (a) en maximisant la distance moyenne entre l'échantillon et tous les autres points du groupe le plus proche (b).

$$s = \frac{b - a}{\max(a, b)} \tag{4}$$

On remarque que s est borné par -1 pour un mauvais regroupement et 1 pour un regroupement parfait. Un score proche de 0 indique des clusters qui se recoupent.

Pour K clusters, l'indice de Calinski-Harabaz [3] est donné par le rapport entre la dispersion moyenne intra-clusters et la dispersion inter-clusters. Un score sera élevé si les clusters sont denses et correctement séparés. Cette approche a l'avantage d'être rapide à calculer.

On définit également l'indice de Davies-Bouldin [7] DB qui correspond à la similarité moyenne entre chaque cluster C_i et son plus proche cluster C_j (eq. 5). La similarité R_{ij} est obtenue à l'aide de s_i la distance moyenne entre chaque point du cluster C_i et son barycentre et de d_{ij} la distance entre les barycentres des

clusters C_i et C_j.

$$DB = \frac{1}{K}\sum_{i=1}^{K} \max_{i\neq j} R_{ij} \quad \text{où } R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{5}$$

Bien que l'ensemble de ces méthodes soient classiquement utilisées pour optimiser le nombre de cluster K, la validité de ces mesures a toujours été réalisée sur des jeux de données simulées. Dans un cas de données réelles, ces indices n'apportent souvent pas de réponse satisfaisante [13].

5. REPRÉSENTATIONS ACOUSTIQUES

Afin de calculer le Laplacien, nous avons besoin de deux représentations acoustiques. La première permet de déterminer la matrice de similarité des récurrences locales R_{loc}, alors que la seconde permet de déterminer la matrice de similarité des répétitions à long-terme S^{rep}.

5.1. Récurrences locales

Les récurrences locales R^{loc} sont calculées directement à partir de $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ avec n le nombre de trames temporelles. Les trames sont généralement extraites toutes les 10 ms et ont une durée de 30 ms. Dans l'article original [17], X est un spectrogramme en échelle log, calculé avec une transformée à Q-constant [21]. La transformée à Q constant a la particularité de conserver un ratio entre la bande passante et la fréquence centrale constant. Ainsi la résolution spectrale est inversement proportionnelle à la fréquence. Cette particularité est adaptée aux signaux musicaux. Par la suite, la représentation acoustique des récurrences locales par un spectrogramme à Q-constant sera conservée.

5.2. Répétitions à long-terme

Dans le script original seul les coefficients cepstraux sont utilisés pour le regroupement en sections musicales. Ces descripteurs acoustiques ont la particularité d'être robustes au bruit, de séparer la source du filtre dans le cas de la parole et ainsi de proposer une représentation compacte et non redondante du contenu fréquentiel d'un signal.

Suivant le type de musique à analyser, les structures musicales peuvent être basées sur le timbre mais également sur les notes jouées. On ajoute donc naturellement les chroma qui correspondent à l'énergie par bandes de fréquences, ces fréquences étant centrées sur les notes de la gamme tempérée. Afin d'observer le signal musical sous différents angles, une batterie de descripteurs spectraux (barycentre, rolloff, contraste et régularité) qui décrivent l'enveloppe spectrale, a également été envisagée.

Enfin, la parole comme la musique sont des phénomènes dynamiques alors que les descripteurs acoustiques (cepstre, chroma ou spectre) sont statiques. Il est donc pertinent de prendre en compte les variations des descripteurs acoustiques entre deux trames consécutives (les Δ). De même qu'en apprentissage automatique, on agglomère plusieurs descripteurs acoustiques dans un même vecteur, on laissera la possibilité à l'utilisateur de traiter un ou plusieurs ensembles simultanément.

Les descripteurs acoustiques sont là pour représenter le contenu fréquentiel du signal musical. La particularité du clustering spectral est d'utiliser une représentation intermédiaire du signal, c'est-à-dire utiliser de nouveaux "descripteurs" (les vecteurs propres) qui permettent de représenter au mieux les aspects saillants contenus dans le signal. Une PCA permet également d'obtenir ce type de représentation intermédiaire.

De part la méthode, l'utilisateur ne pourra donc pas faire simplement le lien entre un résultat d'analyse et un descripteur en particulier. Il pourra faire le lien entre le résultat et un ensemble de descripteurs (spectral, cepstral ou chroma).

5.3. La transformée différentielle

La perception humaine est particulièrement sensible aux changements fréquentiels [6]. Alors que la plupart des représentations musicales sont basées sur des descripteurs statiques, la visualisation de la transformée différentielle permet d'observer d'émergence des changements au niveau mélodique, rythmique ou timbral.

Nous proposons donc de représenter uniquement les valeurs Δ (différentiel entre un segment et son précédent) des descripteurs acoustiques utilisés pour les répétitions à long-terme. Si les segments sont définis par les onsets de note, le différentiel entre deux notes similaires jouées par un même instrument n'aura aucun intérêt. Par contre, si les segments sont définis comme des plages temporelles à instrumentarium constant, le différentiel entre deux plages consécutives représentera le changement de timbre lié à l'apparition de nouveaux instruments.

6. MÉTHODES DE VISUALISATION

L'outil détaillé dans le présent article permet de visualiser la segmentation musicale à différentes étapes du processus d'extraction. Afin de confronter le résultat de la segmentation avec un cas pratique déjà annoté, nous avons repris l'exemple proposé dans [5] à savoir le dernier mouvement *Rondo - Allegretto* de la sonate K545 de Mozart.

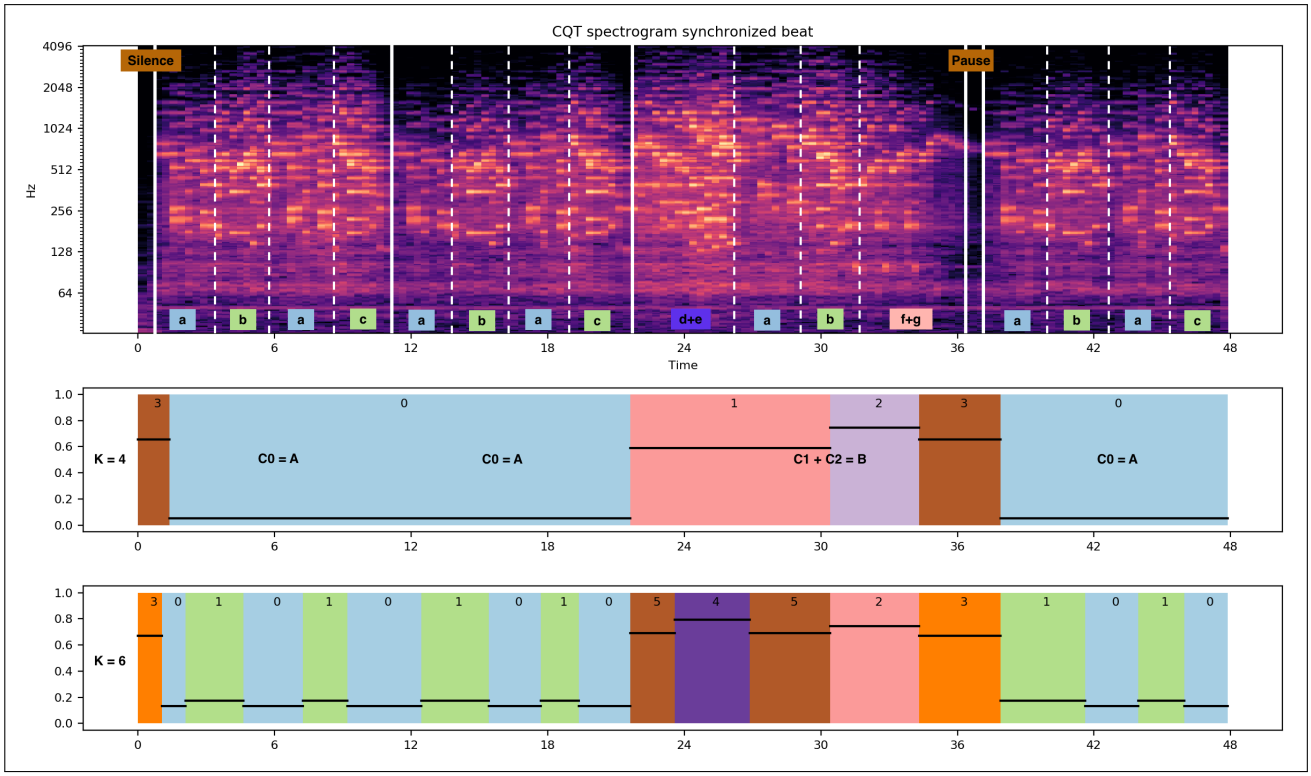

Figure 1. Partitionnement final obtenu sur le début du dernier mouvement *Rondo - Allegretto* de la sonate K545 de Mozart avec utilisation des chroma et une segmentation temporelle initiale basée sur une extraction de pulsation. En haut la représentation fréquentielle à Q-constant synchronisée sur les pulsations, au milieu, la structure obtenue avec $K = 4$ clusters, en bas avec $K = 6$ clusters. La courbe en escalier représente la distance cosinus entre le cluster courant et la référence.

6.1. Structure temporelle

La représentation de la structure temporelle consiste à tracer en fonction du temps 1) la représentation fréquentielle (ici transformée à Q-constant) ; 2) le partitionnement final pour différentes valeurs de K. Ce partitionnement final diffère de la segmentation temporelle initiale (onsets, pulsations ou manuelle). En effet, les différents clusters formant le partitionnement peuvent contenir un ou plusieurs segments temporels suivant leur similarité au regard des descripteurs acoustiques utilisés.

La figure 1 montre un résultat typique obtenu en appliquant l'algorithme de clustering spectral sur un extrait de Mozart avec une extraction automatique des beats. Tout d'abord, on remarque que la segmentation automatique (définie par les frontières de clusters) est légèrement décalée par rapport à la segmentation manuelle (barres blanches verticales). Une temporalité définie manuellement aurait sans doute permis de meilleurs résultats sur la segmentation finale, elle ne remet pas en cause pour autant les résultats sur le regroupement des clusters. En effet, en faisant le partitionnement avec $K = 4$ clusters c_i, on remarque que chaque cluster contient des informations structurellement différentes. En particulier, l'extrait étudié est de forme *AABA* classique. On retrouve complètement cette forme dans le partitionnement obtenu avec $K = 4$ clusters. Plus précisément, c_3 contient du silence ou une pause musicale, alors que c_0 contient la partie A et $c_1 \cup c_2$ contiennent la partie B.Si l'on se reporte sur une analyse experte [6], on peut noter que la partie A contient trois motifs a, b, c organisés tels qu'on ait $A = ab\,ac\,|\,ab\,ac$ avec a de type question, b, c de type réponse. On retrouve ce découpage plus fin également à l'aide d'un nombre de clusters plus élevé $K = 6$.

6.2. Evaluation des segments

Afin d'obtenir une mesure quantitative des différences entre les différents clusters obtenus, nous avons inclus une fonction qui détermine la distance entre un cluster et une référence. La référence peut être entrée manuellement, ou bien prise sur les premiers instants du signal audio. Nous avons fait le choix d'une similarité cosinus qui est largement utilisée pour comparer deux vecteurs, notamment dans la communauté parole. Soit deux vecteurs A et B (ici les valeurs propres du Laplacien, i.e. une représentation acoustique compacte d'un cluster), la distance cosinus est donnée par l'équation 6 où $X \cdot Y$ est le

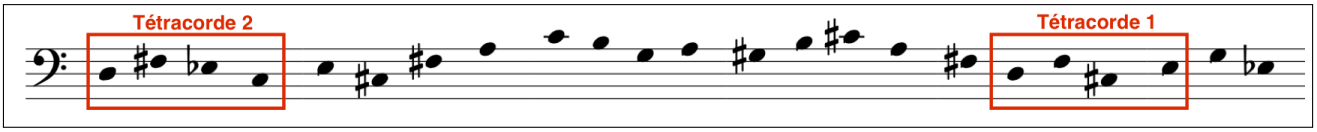

Figure 2. Série de 23 notes extraite de *Globe Unity*.

produit scalaire et $||X||$ est la norme de X.

$$\cos\theta = \frac{X \cdot Y}{||X|| \cdot ||Y||} \tag{6}$$

Plus les vecteurs sont orthogonaux, plus la similarité cosinus tend vers 0, plus les vecteurs ont des orientations proches, plus la similarité cosinus tend vers 1. Par exemple, sur la figure 4, on peut noter la courbe en noire représentant la similarité cosinus du cluster courant par rapport à la référence (début du morceau ici). On peut remarquer que les clusters c_0 et c_1 (correspondant aux motifs *a* et *b*) sont très proches, alors que les motifs centraux (c_2 à c_5) sont éloignés des clusters c_0 et c_1. Cela s'interprête facilement si l'on regarde les différences harmoniques entre les parties *A* et *B*.

7. APPLICATION AU FREE JAZZ

L'année 1966 a marqué la naissance du premier orchestre de Free Jazz Européen *Globe Unity*, faisant appel à des musiciens venant de plusieurs collectifs ayant contribué au développement de cette musique comme le quintet de Manfred Schoof et le quartet de Gunter Hampel. Le concert de cet orchestre en 1966, au festival "Donaueschingen Tage fur Neue Musik" a donné lieu à l'enregistrement du disque *Globe Unity 67/70* (édité en 1967). Nous avons fait le choix d'analyser ce disque suite à la présence d'un support musical (partitions) fourni par le compositeur. En se référant à la partition du morceau *Globe Unity*, les 2 premières minutes montrent une mise en place de l'orchestre sous forme de groupes instrumentaux homogènes interagissant en créant des textures différentes. Il est à noter que la structure de *Globe Unity*, est fondée sur une série de 23 notes, précisée figure 2. La série peut se décomposer en tétracordes structurés en tierces mineures, et formant, à leur tour, des patterns.

7.1. Travail préliminaire

Les premières recherches [4] réalisées sur le morceau *Globe Unity* du disque de même nom se sont d'abord focalisées autour des sources primaires concernant le processus de création collectif. Ces recherches ont permis de découvrir quelques partitions transcrites à la main sous forme de graphiques par le compositeur. Ces partitions apportent des informations précieuses sur l'orchestration ou l'interprétation, cependant elles n'apportent aucun détail sur le contenu musical (complexité rythmique, matériau mélodico-harmoique) généré par le processus d'interaction entre les musiciens, ce qui nous amène à mener des recherches poussées au niveau de la matière acoustique.

Ce travail de recherche inter-disciplinaire tente de mettre en lumière le processus de composition ainsi que le processus d'interaction entre les musiciens sur les deux premières minutes de l'extrait.

7.2. Segmentation initiale

Une première analyse des résultats avec l'approche originale de McFee and Ellis a montré que l'extraction de beats ou d'onsets n'était pas du tout pertinente pour ce type de musique improvisée, mais qu'une temporalité fixe apportait des résultats satisfaisants. Une approche manuelle a l'avantage 1) de permettre à l'analyste de définir lui-même implicitement le niveau de structure qu'il cherche ; 2) d'orienter fortement la recherche de similarité ; et donc apporte des résultats plus pertinents qu'un fenêtrage fixe. Ainsi, l'extraction du rythme en fonction du tactus suggéré par les phrases jouées collectivement par les musiciens, a été faite manuellement.

Nous constatons une certaine régularité au niveau des segments manuels. Ceci s'explique par la cohésion au sein du collectif et l'importance du tactus qui reste fixe et partagé par tous les musiciens. On notera cependant une augmentation de la durée des segments lors des transitions due aux notes tenues.

7.3. Analyse mélodico-harmonique (chroma et Δ)

En choisissant une représentation fréquentielle sur 12 chroma (correspondant aux 12 demi-tons de la gamme tempérée), nous pouvons tracer l'évolution temporelle de ces descripteurs ainsi que leur dynamique (o transformée différentielle). L'analyse conjointe de l'évolution des chroma et respectivement, des Δ chroma (fig 3 en haut, resp. en bas) permet de dégager une forme caractérisée par deux grands moments formels.

Exposition du matériau (1'00 - 1'10). L'analyse des chroma révéle la saillance des hauteurs du tétracorde 1 (do#, ré, mi, fa), première couleur harmonique issue de la série utilisée dans ce morceau ; puis des hauteurs du tétracorde 2 (do, ré, mib, fa#), seconde couleur harmonique.

Phase de transition (1'10 - 1'15. L'analyse des chroma montre une stabilité harmonique et dyna-

4 . Travaux musicologiques menés par Zaher Belgith

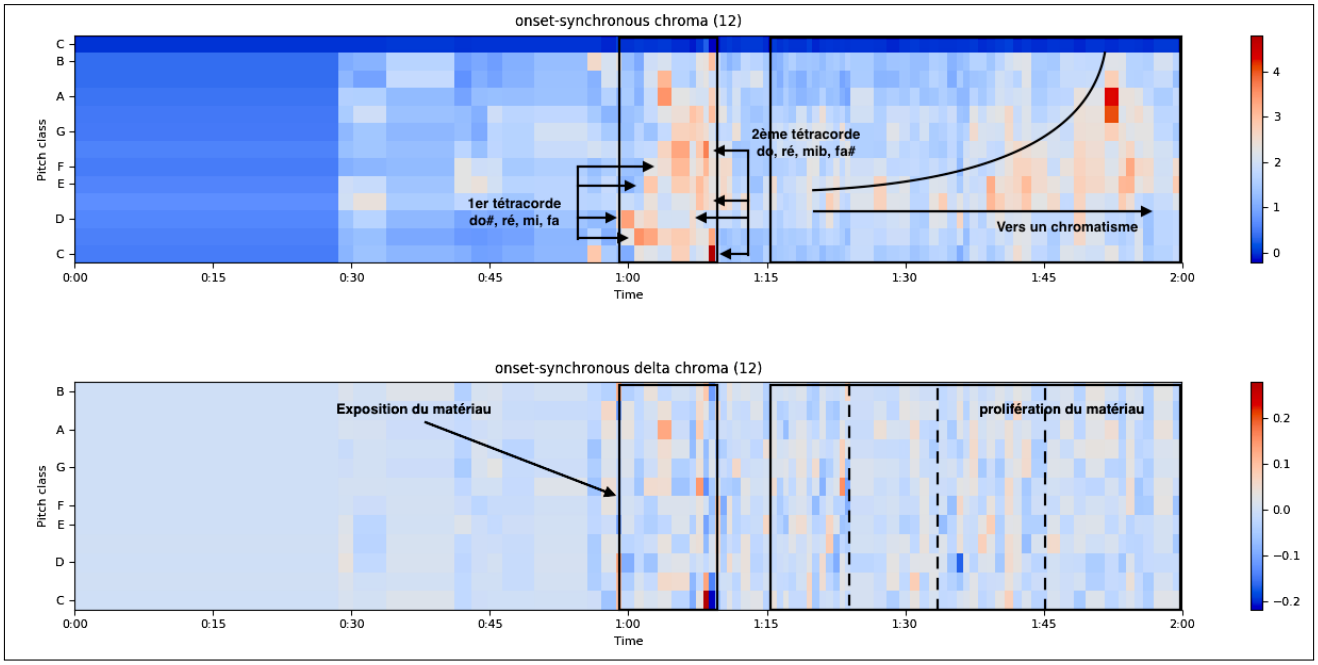

Figure 3. Représentation des chroma (haut) et des Δ chroma (bas) de l'extrait *Globe Unity*.

mique. Nous verrons plus loin qu'il s'agit d'un crescendo qui ne peut pas être visualisé sur ce type de représentation.

Prolifération du matériau (après 1'15). Le matériau musical est développé et retravaillé en variant les modes de jeu (trilles, glissandi, accelerando) et en accentuant certaines notes. L'ajout successif de nouvelles notes au pattern initial par un jeu de chevauchement des tétracordes permet d'augmenter la densité harmonique et donne l'illusion d'un chromatisme total (comme indiqué fig. 3 en haut). Ainsi, le compositeur crée un brouillage de la perception chez l'auditeur et renforce l'aspect atonal de la pièce.

Cette analyse montre que le morceau a été composé autour de tétracordes juxtaposés dans le temps qui se fondent ensuite en un chromatisme qui renforce l'aspect atonal de la pièce. D'après la transformée différentielle (fig. 3 en bas), le second moment formel – prolifération – peut se subdiviser en 4 sections : rappel du tétracorde 1 (1'15 - 1'24), ajout du tétracorde 2 (1'24 - 1'35), ajout d'un troisième tétracorde (fa, sol, lab, sib) (1'35 - 1'45) puis ajout du dernier tétracorde (fa#, sol#, la, si) (1'45 - 1'57). Ces derniers tétracordes s'inscrivent dans la continuité des premiers selon la même logique de tierces mineures.

Alors que la représentation des chroma permet une analyse mélodico-rythmique basée sur la saillance de certaines hauteurs, le transformée différentielle (i.e. les Δ) permet de confirmer l'évolution dynamique entre les différents patterns (ici des tétracordes). La dynamique nous indiques "les transferts d'énergie en jeu dans les transitions harmoniques" [6]. En particulier la transformée différentielle permet de confirmer la segmentation temporelle du moment de prolifération en quatre sous sections. La caractérisation mélodico-harmonique de ces sections se faisant grâce aux chroma directement.

7.4. Partitionnement final

Après avoir proposé un niveau de structuration à l'aide de la représentation des chroma, nous souhaitons mettre en évidence un autre niveau de structuration en mélangeant plusieurs descripteurs audio. Dans la mesure où les musiciens de ce répertoire cherchent à exprimer, pendant le processus de l'interaction, un contraste en modifiant le registre ou la texture sonore, le choix de combiner des descripteurs spectraux comme le centroid spectral ou le roll-off et des chroma, s'impose. Le nombre de clusters K déterminé automatiquement est très élevé et implique une segmentation beaucoup trop fine par rapport aux attendus de l'expert. Ainsi, après plusieurs essais, celui-ci a fixé manuellement à $K = 6, 8$. La segmentation obtenue est reportée figure 4 et les regroupements des clusters suivant les moments musicaux identifiés à la section précédente sont précisés dans le tableau 1.

Dans la configuration $K = 6$, l'algorithme de *spectral clustering* identifie correctement les moments d'introduction, d'exposition et de transition. Ce cluster se caractérise par un crescendo dynamique (succession d'instruments, gardant le principe de séparation de couleurs instrumentales où les timbres ne seront pas mélangés) suivi d'un petit crescendo d'intensité exprimé par les trompettes pour annoncer le début d'un autre moment. Par contre, les trois dernières sections du moment de prolifération ne correspondent qu'à

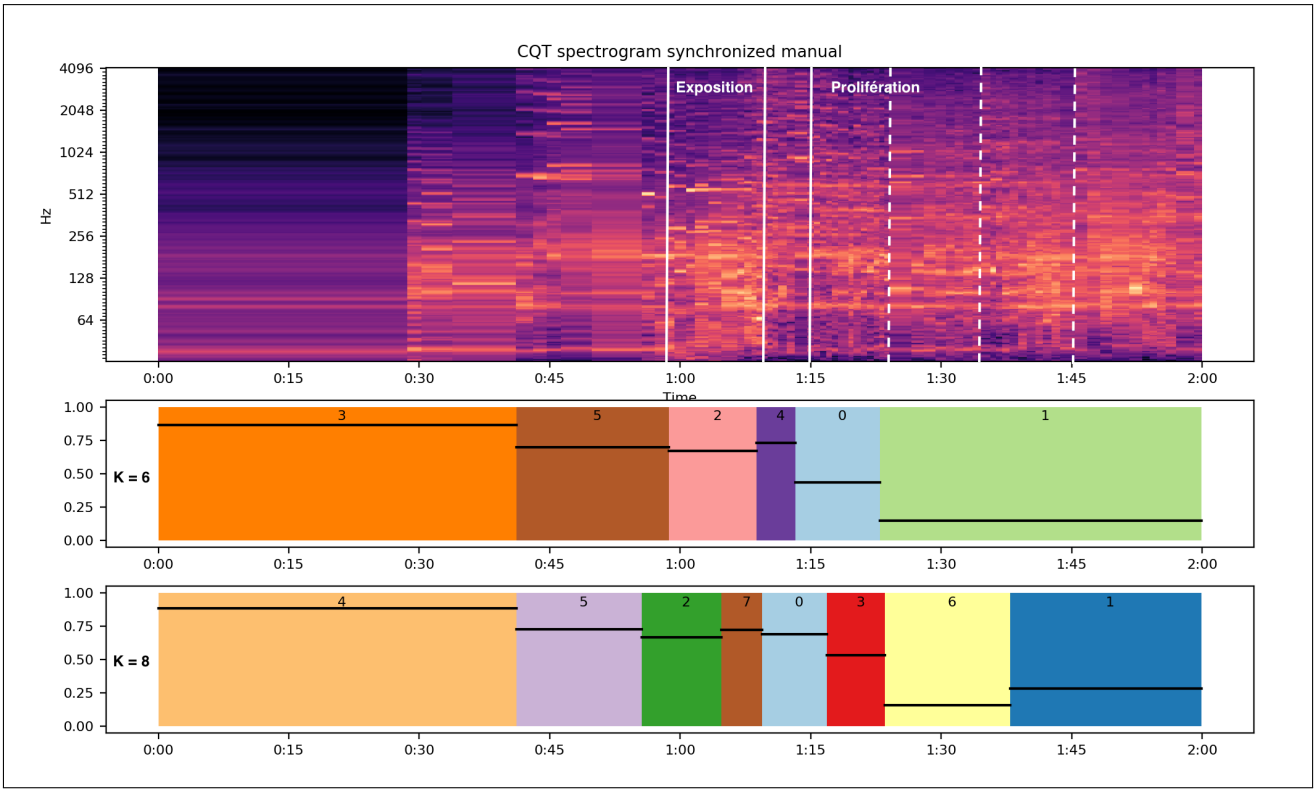

Figure 4. Partitionnement final obtenu sur l'extrait de *Globe Unity* avec utilisation des chroma+spectraux et une segmentation temporelle initiale manuelle. En haut la représentation fréquentielle à Q-constant synchronisée sur les segments manuels, au milieu, la structure obtenue avec $K = 6$ clusters, en bas avec $K = 8$ clusters. La courbe en escalier représente la distance cosinus entre le cluster courant et la référence. Les barres verticales blanches rappellent la segmentation obtenue avec la représentation des chroma.

Description		$K = 6$	$K = 8$
Introduction		$c_3\, c_5$	$c_4\, c_5$
Exposition		c_2	c_2
Transition		c_4	$c_7\, c_0$
Prolifération	Tétra 1	c_0	$c_0\, c_3$
	Tétra 1+2	c_1	c_6
	Tétra 1-3	c_1	c_1
	Tétra 1-4	c_1	c_1

Table 1. Descriptifs des clusters obtenus sur *Globe Unity*.

un seul cluster (c_1 à partir de 1'24).

La configuration $K = 8$ permet d'affiner notre structure formelle. En effet, la phase de transition peut être étendue aux clusters c_7 (crescendo dynamique) et c_0 (crescendo d'intensité), ce dernier termine la transition et commence la prolifération. La fin de la première section du moment de prolifération (Tétra 1) est représentée par c_3. Ce cluster se caractérise par une complexité rythmique où chaque musicien propose un geste rythmique différent de celui proposé par l'autre musicien tout en suggérant un tactus commun par différents gestes rythmiques sans pour autant l'exprimer [19]. Il s'avère qu'aucune des configurations ne permet de retrouver la subdivision en quatre du moment de prolifération. Seule la première section (Tétra 1) est obtenue. Notre hypothèse est que la segmentation proposée par l'analyste à partir de la représentation en chroma est très fortement liée à un modèle cognitif que l'outil informatique ne peut pas capturer.

8. CONCLUSION ET PERSPECTIVES

L'article présente une description détaillée de l'outil et une proposition d'interprétation des résultats obtenus sur une œuvre musicale. D'un point de vue informatique, l'implémentation de ce type de technique et son adaptation à un contexte aussi exigeant que le *Free Jazz* est un défi. Cette première étape a montré que l'algorithme est capable de dégager une structure formelle cohérente et pertinente avec les résultats de l'analyste. Cela s'avère de plus être une aide précieuse pour la découverte de nouvelles catégories musicales. Combiné avec la représentation de la transformée différentielle, nous présentons un outil complet d'analyse musicale qui a montré qualitativement ses preuves sur de la musique de variété, de la musique classique et du *Free Jazz*. Une évaluation quantitative sur de plus nombreux extraits musicaux est nécessaire pour étudier les limites d'utilisation de

cet outil.

La détermination d'un niveau de structure clair est fondamentale pour que l'analyste et l'auditeur appréhende les enjeux du *Free Jazz* et la manière dont la musique est mise en œuvre par les différents musiciens. Ceci permettra de proposer un modèle de forme bien défini et par la suite, de dégager une *gestalt* du *Free Jazz*. Ce modèle confirmait l'idée que le morceau peut être déterminé a priori dans sa totalité plutôt que de penser en terme d'une succession temporelle de simples accidents formels ou de moments collectifs d'appels/réponse partagés par des musiciens sans avoir une ligne directrice.

Ce type d'approche pluridisciplinaire est extrêmement riche et permet d'apporter des regards nouveaux dans chacune des disciplines. Les approches d'analyse computationnelle sont en plein développement d'un point de vue technique, cependant le lien entre informatique, musique et cognition reste encore à faire.

9. REFERENCES

[1] Ahn, Y.-K., "L'analyse musicale computationnelle : rapport avec la composition, la segmentation et la représentation à l'aide de graphes". *Thèse de doctorat*, Université Pierre et Marie-Curie, ParisVI, 2009.

[2] Broux, P.-A., Doukhan, D., Petitrenaud, S. , Meignier, S. and Carrive, J. "Segmentation et Regroupement en Locuteurs : comment évaluer les corrections humaines". In *Journées d'Études sur la parole (JEP)*, Aix-en-Provence, France, 2018.

[3] Calinìski, T., and Harabasz, J., "A dendrite method for cluster analysis". In *Communications in Statistics-theory and Methods* 3, pp. 1–27, 1974.

[4] Canonne, C. "L'improvisation collective libre : de l'exigence de coordination à la recherche de points focaux : cadre théorique, analyses, expérimentations." *Thèse de doctorat*. Université Jean Monnet - Saint-Etienne, 2010.

[5] Chouvel, J.-M. "Structural analysis and Cognitive activity : towards real-time methods in musical analysis". In *Structure and Cognition*, feb.,7, 2005.

[6] Chouvel, J.-M., Bresson, J. and Agon, C. "L'analyse musicale différentielle : principes, représentation et application à l'analyse de l'interprétation". In *Electroacoustic Music Studies Network Conference - EMS'07*, Leicester, Royaume-Uni, 2007.

[7] Davies, D., Bouldin, L. and Donald W. "A Cluster Separation Measure". In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1 (2), pp. 224–227, 1979.

[8] Deutsch, D. "Music perception", in *Frontiers of Bioscience*, vol. 12, pp. 4473–4482, 2007.

[9] Draksler, K. "Cecil Taylor : life as ... A structure within improvisation". *Non publié* , 2013. Disponible à `http://www.kajadraksler.com/Taylor.pdf`

[10] Foote, J. and Cooper,M. "Visualizing Musical Structure and Rhythm via Self-Similarity". In *proc. of International Computer Music Conference*, vol. 2001, 2001.

[11] Foote, M. "Automatic audio segmentation using a measure of audio novelty". In *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, pp. 424–455, 2000.

[12] Jost, E. "Free Jazz : Une étude critique et stylistique du jazz des années 1960". Cotro, V. (traducteur), *Outre mesure*, Paris (eds), coll. Contrepoints, 2002.

[13] Lamirel, J.-C., Dugué, N. and Cuxac, P. "New efficient clustering quality indexes". In *proc ; of International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, Canada, 2016.

[14] Levy, M., Sandler, M. and Casey, M. "Extraction of high-level musical structure from audio data and its application to thumbnail generation". In *proc. of ICASSP*, Toulouse, France, 2006.

[15] Liuni, M., Röbel, A., Romito, M., and Rodet, X. "Rényi information measures for spectral change detection". In *proc. of ICASSP*, Prague, Czech Republic, 2011.

[16] Martin, B., Robine, M. and Hanna, P. "Musical structure retrieval by aligning self-similarity matrices". In *proc. of International Society for Music Information Retrieval Conference*, 2009.

[17] McFee, B. and Ellis, D P.W. "Analyzing song structure with spectral clustering". In *proc. of International Society for Music Information Retrieval.*, 2014.

[18] Ong, B. S., Gómez, E. and Streich, S. "Automatic extraction of musical structure using pitch class distribution features". In *proc. Workshop Learning the Semantics of Audio Signals*, pp. 53–65, 2006.

[19] Porter, L. "John Coltrane : His life and Music". Cotro, V. (traducteur), *Outre mesure*, Paris (eds), coll. Contrepoints, 2007.

[20] Rousseeuw, P. J., "Silhouettes : a Graphical Aid to the Interpretation and Validation of Cluster Analysis". In *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[21] Schoerkhuber, C. and Klapuri, A. "Constant-Q transform toolbox for music processing". In *proc. of Sound and Music Computing Conference*, Barcelona, Spain. 2010.

[22] Zelnik-Manor, L. and Perona P. "Self-Tuning Spectral Clustering", in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1601–1608, 2004.