Segmentation temporelle et estimation de la structure musicale par utilisation de probabilité a priori de retard et DTW contraint

Victor Bisot

Rapport de fin de Master 2 ATIAM

Encadrantt : Geoffroy Peeters

Juillet 2014

IRCAM

1 Place Igor-Stravinsky

75004 Paris, FRANCE

Résumé

Ce rapport de fin de stage vise à présenter de nouvelles méthodes d'estimation automatique de la structure musicale. Elles commencent souvent par traiter un fichier audio en le séparant en segments avant de les regrouper en classes. Ces deux étapes appelées de segmentation et de regroupement peuvent se faire sous plusieurs hypothèses sur le contenu du signal (répétition, homogénéité ou nouveauté). Nous utilisons dans ce rapport une approche par séquences qui suppose la présence de répétitions dans le signal. Cette approche nécessite d'identifier des diagonales dans des matrices d'auto-similarité afin de séparer le morceau en segments. Nous proposons une nouvelle méthode de segmentation s'inspirant de deux méthodes de l'état de l'art utilisant des matrices d'auto-similarité "temps-retard". Nous utilisons des probabilités a priori de présence de segments sur les retards (issue des travaux de Goto en 2003) comme pondération pour une différence de trame à trame temporelle de "vecteurs de structures" (introduit par Serra en 2012). Les évaluations de cette méthode sur plusieurs bases de tests ont montré que cette combinaison permet une nette amélioration des résultats de la segmentation. Nous proposons également deux nouvelles méthodes de regroupement reprenant les résultats de la segmentation. Nous utilisons un algorithme de déformation temporelle dynamique contraint afin d'identifier les répétitions des segments délimités par la segmentation. Nous séparons ensuite les segments en classes (deux segments répétés ont la même classe) à l'aide de deux nouvelles méthodes. L'évaluation de ces deux méthodes sur plusieurs bases de tests donne de bons résultats. Nous étudions également l'introduction de nouvelles contraintes sur l'algorithme de déformation temporelle dynamique afin d'améliorer nos méthodes de regroupement.

Mots clés : Structure musicale, matrice d'auto-similarité, segmentation temporelle, déformation temporelle dynamique

Abstract

This master's thesis aims at presenting new methods for automatic musical structure discovery. These methods often start by separating an audio file into segments before labelling them. These two steps are called segmentation and labelling, they can be based on different hypotheses on the musical content (homogeneity, repetition and novelty). A sequence approach is used for our methods wich is based on the repetitive aspect of the signal. The sequence approach works by detecting diagonals in a self-similarity matrix to seperate the signal into repetitive parts. We propose a new segmentation method inspired by two well known methods both using a "time-lag" self-similarity matrix. We use the possibility of having a segment on each lag (from Goto's work) as a prior knowledge for a "structure feature" frame to frame difference (from Serra's work). The evaluation of this method on 3 test sets shows that this combination improves the segmentation results. We will use a constrained dynamic time warping algorithm to find the repetitions of the estimated segments. We will then label the segments (two repeated segments will have the same label) using two new labelling methods. The results of the evaluation of these two new methods are quite promising. We also study the possibility of adding new constraints to improve the performances of our labelling methods.

Keywords : Musical structure, self-similarity matrix, temporal segmentation, dynamic time warping

Table des matières

1 Introduction

1.1 La structure musicale

La perception humaine de la musique se base sur l'organisation de la succession d'éléments sonores élémentaires en entités plus complexes. L'organisation des sons lors de l'écoute peut se faire sur plusieurs échelles temporelles qui peuvent aller de la simple note à la mélodie jusqu'à des parties plus longues de l'ordre de la dizaine de secondes. Les éléments musicaux à courtes échelles, comme le tempo ou la note, sont souvent quantifiables et sont définis par des règles musicales très strictes. En revanche il n'y a pas de règles pour la description de la structure musicale, il peut exister plusieurs constructions différentes de la structure pour un même morceau selon ce qu'elle traduit. Elle peut décrire la répartition du morceau en segments selon plusieurs critères comme l'instrumentation, la similarité acoustique ou encore le rôle du segment dans le morceau. Cette structure est perceptible par l'humain mais ne possède pas de définition stricte, elle est le résultat de relations et de répétitions entre des éléments musicaux plus courts.

Le manque de définition de la structure musicale pousse les chercheurs à poser différentes hypothèses afin de guider leurs méthodes d'estimation. Paulus et al. [1] décrivent trois hypothèses sur le contenu musical récurrentes dans les méthodes d'estimation de la structure : l'homogénéité, la répétition et la nouveauté. L'hypothèse d'homogénéité suppose que les segments de la structure sont composés de successions de temps aux propriétés acoustiques similaires, cette similarité est souvent liée à l'instrumentation ou à la dynamique. L'hypothèse de répétition suppose que le morceau est une succession de différents segments répétés au cours du temps. Pour les méthodes basées sur l'hypothèse de nouveauté il s'agit de repérer les positions temporelles des changements importants dans le morceau afin d'estimer la séparation entre les segments de la structure.

Les méthodes d'estimation de la structure fonctionnent pour la plupart en deux temps. Elles commencent par une étape de segmentation qui a pour rôle d'estimer la position des frontières entre les différentes parties de la structure. Une fois les frontières estimées, la description de la structure nécessite souvent d'identifier parmi les segments lesquels sont répétés entre eux. Ensuite une étape appelée de regroupement va attribuer une étiquette à chaque partie délimitée par la segmentation. Les méthodes de regroupement attribuent la même étiquette aux segments étant des répétitions les uns des autres. La structure musicale est un concept très accessible pour les amateurs de musique, il s'agit d'une information compréhensible par tous que l'on peut donner lors de l'analyse du contenu d'un morceau de musique. De plus l'estimation de la structure musicale possède de nombreuses applications innovantes : une navigation améliorée entre les parties d'un morceau dans un lecteur de musique [2], de la génération de résumé audio [3][4], du mélange de morceau (mash-up) [5] ou encore de la recherche de reprise [6][7].

1.2 État de l'art

Dans [4], les méthodes d'estimation de la structure sont classées en deux catégories : les approches par états et les approches par séquences. L'approche par état rejoint l'hypothèse d'homogénéité, elle suppose que le signal est découpé en segments homogènes au cours du temps, qui peuvent être répétés ou non. L'approche par état rejoint également l'hypothèse de nouveauté car on suppose qu'un passage entre deux segments homogènes très différents crée une forte "nouveauté". L'approche par séquence suppose que le signal est séparable en séquences d'évènements successifs répétées au

cours du temps. Elle rejoint l'hypothèse de répétition dans le cas où les segments répétés ne sont pas homogènes (sinon cela correspondrait à l'approche par état). L'apparition de la distinction entre ces deux approches est principalement due à l'introduction des matrices d'auto-similarité pour l'analyse musicale [8]. Ces matrices seront utilisées par la suite dans presque toutes les méthodes d'estimation de la structure. Elles permettent une bonne visualisation de la ressemblance entre le contenu des différentes trames d'analyse du signal et donnent une idée de l'approche à adopter pour la segmentation et le regroupement. L'hypothèse d'homogénéité et l'approche par état sont utilisées lors de la présence de blocs de forte similarité dans la matrice. L'approche par séquence est adaptée lors de la présence de diagonales qui correspondent à des répétitions de parties non homogènes. Peeters a proposé dans [9] une méthode pour estimer l'approche idéale à adopter pour une partie du morceau donnée selon la répartition en blocs ou en diagonales de la matrice d'auto-similarité.

TABLE 1 – Approche adoptée d'estimation de la structure selon les hypothèses d'homogénéité et de répétition

	Homogène	Non homogène
Répété	état	séquence
Non-répété	état	bruit

Méthodes basées sur l'hypothèse d'homogénéité ou de nouveauté

L'hypothèse d'homogénéité est associée à la présence de blocs dans la matrice d'auto-similarité correspondant à des segments homogènes, la présence successive de deux parties homogènes crée une impression visuelle d'échiquier dans la matrice. Ainsi Foote [8] propose en 2000 de détecter le passage entre deux segments homogènes différents à l'aide d'une convolution par une matrice à l'allure d'un échiquier (checkerboard kernel). La convolution de ce noyau se fait sur la diagonale principale de la matrice d'auto-similarité et donne des fortes valeurs lors du temps de passage d'un bloc de forte similarité à un autre. Elle permet de réaliser une segmentation en ajoutant une étape de sélection de pics afin d'identifier la position des frontières importantes entre les différents segments homogènes du morceau. La taille du noyau étant fixe, une hypothèse de durée a été posée sur la durée des évènements. Kaiser et Peeters [10] ont proposé, pour ne pas se limiter à une taille fixe, d'utiliser ces noyaux sur différentes échelles temporelles. Ils ont également introduit l'utilisation de deux nouveaux noyaux permettant de détecter le passage entre des parties non homogènes et des parties homogènes. La méthode de Foote, grâce à ses très bons résultats, reste très utilisée pour l'étape de segmentation dans la plupart des méthodes d'estimation de la structure utilisant une approche par état [11][12][13]. Certaines méthodes traitent les blocs en entier et non uniquement les positions des frontières entre deux blocs en utilisant des algorithmes de programmation dynamique pour la segmentation [14][15].

Différentes techniques d'apprentissage machine déjà présentes dans d'autres domaines sont utilisées pour l'étape de regroupement telles que les k-moyennes [16] ou des modèles de Markov cachés [17][18]. [1] présente une analyse plus complète de l'état de l'art en détaillant les contributions principales pour chaque approche.

Méthodes basées sur l'hypothèse de répétition

L'hypothèse de répétition est basée sur la présence de sous diagonales dans la matrice d'auto-similarité correspondant à des segments non homogènes répétés au moins une fois dans le morceau. Ainsi la plupart des méthodes de segmentation utilisant l'hypothèse de répétition fonctionnent en identifiant toutes les sous diagonales dans la matrice d'auto-similarité afin de connaître la position de tous les segments répétés. L'identification de ces diagonales passe souvent par l'utilisation de descripteurs adaptés à une approche par séquences comme les chromas, ils représentent l'aspect harmonique du signal. Beaucoup de variantes des chromas ont été développées, comme les CENS [19] les CRP [20] ou encore les HPCP [21], dans le but de mettre en valeur les segments non-homogènes répétés dans la matrice d'auto-similarité. En revanche l'apparition de variation de timbre, de tempo ou de dynamique ne permet parfois plus aux segments répétés d'être représentés par des diagonales continues et parallèles à la diagonale principale. Une étape importante des méthodes basées sur l'hypothèse de répétitions est de mettre en valeur ces diagonales dans la matrice d'auto-similarité. Cela peut se faire à l'aide de filtrage passe bas dans la direction diagonale et passe haut dans la direction orthogonale (afin de limiter les effets de blocs) [22], grâce à des procédés d'érosions dilatations sur des matrices binarisées [23] ou grâce à l'utilisation de descripteurs et de filtres invariants au tempo et à la dynamique [3].

Pour détecter les diagonales Goto [2] travaille sur matrice d'auto-similarité sur un axe temps-retard où les répétitions sont représentées par des segments parallèles à l'axe des temps (donc à retard constant). Il propose de détecter les diagonales en calculant sur quelles valeurs de retards il est plus probable de trouver des segments afin de réduire la recherche. D'autres travaux se basent sur une utilisation de cette matrice temps-retard comme ceux de Serra et al. [24] qui possèdent les meilleurs résultats de l'état de l'art en termes de segmentation sur plusieurs bases de tests. Ils se basent sur le fait que deux segments répétés commencent aux mêmes temps dans la matrice temps-retard. Ils estiment la position des frontières entre les segments du morceau en réalisant une différence trame à trame des vecteurs descripteurs particuliers qu'ils nomment "vecteurs de structures", cette différence sera importante sur les instants de début et fin de répétitions dans le morceau. Des méthodes de programmation dynamique sont également utilisées pour trouver les diagonales dans une matrice d'auto-similarité en effectuant une recherche de meilleur chemin [25] suivi d'un algorithme de Viterbi pour identifier les diagonales. Kaiser [26] propose en 2013 une nouvelle méthode de segmentation utilisant les deux approches : par états et par séquences. Il arrive en utilisant des méthodes connues pour les deux approches à une segmentation prenant en compte à la fois les répétitions non-homogènes et les parties homogènes.

1.3 Organisation du rapport

Nous présentons dans le chapitre 2 les bases de tests utilisées pour les évaluations de nos méthodes de regroupement et segmentation ainsi qu'une brève discussion sur les enjeux de l'annotation de la structure musicale. Le chapitre 3 introduit l'utilisation et la construction des matrices d'auto-similarité pour l'estimation de la structure musicale utilisant une approche par séquences. Nous décrivons le type de descripteurs utilisés pour la construction de la matrice ainsi que plusieurs traitements permettant de mettre en valeur les diagonales dans les matrices. Le chapitre 4 s'intéresse à la segmentation, après avoir introduit les matrices temps-retard nous commencerons par détailler le fonctionnement des deux méthodes de segmentation dont nous nous servirons pour notre sys-

tème : l'approche retard de Goto et l'approche temps de Serra. Nous proposons ensuite de nouvelles méthodes de segmentation que nous évaluons sur plusieurs bases de tests. Afin de compléter notre segmentation, nous ajoutons à nos méthodes de l'information issue de la segmentation se basant sur l'hypothèse d'homogénéité. Après la segmentation le chapitre 5 traite du regroupement, l'autre étape d'un système d'estimation de la structure. Nous détaillons le fonctionnement de l'algorithme DTW (dynamic time warping) contraint de Muller que nous utilisons afin de développer deux nouvelles méthodes de regroupement. Après avoir évalué et comparé nos méthodes de regroupement nous étudierons l'ajout de nouvelle contrainte au DTW afin d'améliorer les performances de nos méthodes.

2 Les bases de tests

Pour évaluer nos méthodes d'estimation de la structure nous utilisons des bases de tests connues de l'état de l'art annotées en segments et en labels. Les annotations en segments correspondent aux frontières entre les différentes régions musicales du morceau et sont utilisées pour évaluer les méthodes de segmentation. Les annotations en labels associent un label à chaque segment délimité par les annotations en segments, les segments répétés possèdent le même label. Afin de pouvoir comparer nos résultats à ceux de l'état de l'art, nous reprenons les bases de tests sur lesquelles des résultats ont déjà été publiés, nous présentons les bases utilisées ainsi que quelques statistiques sur les annotations (Table 2) :

BEATLES : Il s'agit de la base de test des Beatles issue des bases de tests Isophonics [27]. Elle est composée de 180 morceaux correspondant à tous les titres des 12 albums originaux des Beatles. Les annotations en structure sont fournies par Isophonics.

RWC POP-A : Il s'agit de la base de données RWC-Popular-Music [28] composée de 100 morceaux représentant la musique populaire japonaise et américaine. Ce sont les annotations originales de la base fournies par l'AIST [29].

RWC POP-B : Il s'agit de la même base de données que la précédente avec des annotations plus récentes fournies par l'INRIA (Institut national de recherche en informatique et automatique) [30].

L'évaluation des méthodes d'estimation de la structure pose des questions sur la pertinence des annotations des bases de tests. Les annotations en structure peuvent se faire sur différents niveaux, par exemple si un refrain est composé de deux répétitions successives, nous pourrions l'annoter en deux blocs ou en un seul bloc. De plus la position temporelle des frontières entre deux parties étant parfois peu contrastée, l'humain peut, selon sa perception de la musique, placer la frontière à des endroits très différents. [31] liste plusieurs manières d'annoter un morceau en structure. Nous pouvons séparer le morceau en différentes parties selon leur rôle (couplet, refrain...), selon leur similarité acoustique (des parties avec peu de variations), selon leur instrumentation ou encore en se basant sur des tests perceptifs.

La recherche en estimation de la structure se fait également sur la mise en place de modèles et de règles d'annotations en segment et en label dans le but de permettre à deux personnes différentes d'annoter un morceau de la même manière. Bimbot et al. [30] ont proposé un nouveau processus d'annotation inspiré et adapté de la sémiologie en analysant les propriétés de blocs dans le morceau. C'est sur la base de cette nouvelle méthode que les annotations de la base RWC POP ont été refaites pour donner celle de la base de tests RWC POP B.

TABLE 2 – Statistiques des bases de données : Le nombre de titres dans la base - le nombre moyen de frontières dans l'annotation - l'intervalle moyen en secondes entre deux frontières consécutives

	Nb Titres	Nb Frontières (écart type)	Intervalle (écart type)
BEATLES	180	9.2 (2.3)	16.0 (13.9)
RWC-POP-A	100	16.1 (4.0)	14.1 (6.8)
RWC-POP-B	100	16.8 (3.4)	13.7 (7.2)

3 Descripteurs audio et matrices d'auto-similarité

Le concept de matrice d'auto-similarité pour la musique a été introduit par Foote [8], elles sont souvent utilisées en estimation de la structure comme un moyen de représenter l'aspect répétitif d'un signal temporel. L'information que contiennent ces matrices est issue d'une séquence de vecteurs de descripteurs $V = (v_1, .., v_N)$ capturant certaines caractéristiques du signal à travers le temps. Ces descripteurs peuvent être de nature très différente, néanmoins en estimation de la structure, nous rencontrons majoritairement des MFCC (Mel Frenquency Cepstral Coefficients) et des chromas. Si nous posons $s(V_1, V_2)$ comme une mesure de la ressemblance entre deux observations V_1 et V_2 aux temps d'analyse t_1 et t_2, nous construisons alors la matrice d'auto-similarité en rangeant les valeurs de ressemblance entre chaque vecteur d'observations $S(1, 2) = s(V_1, V_2)$. La matrice S est donc carrée et symétrique.

Le choix de l'approche par état ou par séquence motive grandement le choix des descripteurs utilisés. Pour l'approche par état, la recherche de parties homogènes dans le signal, l'utilisation de descripteurs liés au timbre est souvent favorisée, les plus connus étant les MFCC. Une partie homogène en timbre entre t_i et t_j dans un morceau de musique sera représentée par un bloc homogène dans la matrice d'auto-similarité, nous aurons des fortes valeurs de $S(x, y)$ pour $(x, y) \in [i : j]\times[i : j]$. Pour l'approche par séquence, la recherche de séquences répétées au cours du temps, nous préférons utiliser des descripteurs liés à l'aspect harmonique du signal comme les chromas. Une séquence répétée au cours du temps sera représentée par des diagonales parallèles à la diagonale principale de la matrice. Si une partie de t_1 à t_2 est répétée entre t_3 et t_4 nous aurons alors deux diagonales dans la matrice entre $s(1, 3)$ et $s(2, 4)$ et entre $s(3, 1)$ et $s(4, 2)$.

La matrice de similarité est le point de départ de notre système, c'est sur cette représentation de l'information que se basent les méthodes de segmentation et de regroupement présentées par la suite. Le choix des descripteurs utilisés, la construction et les traitements de la matrice ont une grande influence sur les performances de notre méthode d'estimation de la structure. Notre recherche s'axant principalement sur l'approche par séquence, nous utilisons essentiellement des chromas pour la construction de la matrice d'auto-similarité, que nous présentons dans la partie suivante. Nous détaillerons également les différents traitements effectués sur les matrices et les descripteurs de manière à mettre en valeur les diagonales.

3.1 Les chromas CENS

3.1.1 Les chromas

Notre approche par séquence de l'estimation de la structure nous amène à chercher des descripteurs capables de représenter uniquement les séquences répétées dans le signal. Nous souhaitons nous détacher de l'information liée au timbre pour éviter au maximum l'apparition de blocs dans nos matrices d'auto-similarité. Toutes nos méthodes de segmentation et de regroupement se basent sur la détection de diagonales (de répétitions) dans la matrice d'auto-similarité, tout ce qui ne correspond pas à des séquences répétées n'est pas exploité par nos méthodes de segmentation. Nous considérons donc tout ce qui ne s'apparente pas à une diagonale dans la matrice d'auto-similarité comme du bruit. Ce bruit correspond à de l'information pouvant être utile pour d'autres méthodes d'estimation de la structure mais n'apporte rien à la nôtre. Comme nous le verrons par la suite il

est possible de traiter ces matrices afin de mettre en valeur ces diagonales et de limiter la présence du bruit. Pour faciliter la tâche nous cherchons des descripteurs encourageant le plus possible l'apparition de diagonales tout en évitant de créer des blocs. Les chromas sont réputés être très efficaces pour ce genre d'exigences et sont la base de beaucoup de méthodes d'estimation de la structure. Les chromas dépendent tout de même indirectement du timbre mais favorisent l'apparition de séquences répétées par rapport aux parties homogènes.

Les vecteurs de chromas sont des vecteurs à 12 dimensions $V_t = (v_{t,1}, ...v_{t,12})^T$ où chaque dimension correspond à une des 12 notes de la gamme tempérée, $v_{t,1}$ est associé au DO et $v_{t,3}$ au RE et ainsi de suite. Ils reprennent l'idée que la perception humaine de la hauteur est périodique, c'est à dire que deux hauteurs paraissent semblables si elles diffèrent d'une octave. Une note correspond donc à un chroma plus une octave [32]. Chaque vecteur de chroma extrait d'un signal musical traduit la répartition de l'énergie à court terme entre les 12 bandes de chromas. La relative indépendance au timbre des chromas s'explique par leur cumul de la contribution de chaque harmonique d'un son donnée, habituellement caractéristique du timbre, pour exprimer uniquement la répartition de l'énergie en 12 bandes. Ils sont souvent utilisés dans des méthodes de détection d'accord [33], de synchronisation musicale [34] et pour l'estimation de la structure musicale.

Il existe une multitude de versions de la décomposition d'un signal en vecteurs de chromas qui modifient leurs propriétés par des pré ou post-traitements agissant sur leur comportement spectral, temporel ou dynamique. Ces changements peuvent apporter des différences intéressantes dans le comportement et l'information contenue dans les vecteurs de chromas, ils sont guidés par les exigences des différentes applications d'analyse musicale faisant appel à l'utilisation des chromas. Notre objectif n'étant pas de comparer leur efficacité nous justifierons uniquement l'arrêt de notre choix sur les CENS (Chroma Energy distribution Normalized Statistics) [19] utilisés dans [3]. Nous présentons brièvement les étapes du calcul des CENS, l'implémentation utilisée est celle de la Chroma ToolBox [35].

3.1.2 Les CENS

Pour calculer les CENS nous partons d'une décomposition du signal en vecteurs de pitchs à 88 dimensions, pour cela nous séparons l'énergie du signal en 88 bandes de fréquences à l'aide d'un banc de filtre où chaque filtre est centré sur une note entre LA0 et DO8 (note midi de 21 à 108). Nous convoluons chaque bande avec une fenêtre rectangulaire de 200ms et un recouvrement de 50% pour ressortir l'énergie moyenne à court terme dans chaque bande. Nous réduisons ensuite l'information à 12 bandes en additionnant les bandes correspondant à la même note à l'octave près pour former les vecteurs de chromas. Enfin pour tous les vecteurs de chroma nous normalisons chaque coordonnée par l'énergie totale de la fenêtre pour que le résultat soit invariant a la dynamique du signal. Nous pouvons trouver plus de détails sur la méthode dans [19].

Pour calculer les CENS nous prenons le vecteur de chromas normalisé à la trame d'analyse t $V_t = (v_{t,1}, ..., v_{t,12})$ que nous allons ensuite quantifier sur 5 valeurs possibles. Nous assignons la valeur 4 à $v_{t,i}$ si $v_{t,i} > 0.4$ (si il contient plus de 40% de l'information), 3 si $0.4 < v_{t,i} < 0.2$, 2 si $0.1 < v_{t,i} < 0.2$, 1 si $0.05 < v_{t,i} < 0.1$ et 0 sinon. Cette quantification introduit une sorte de compression logarithmique de l'information. Nous filtrons ensuite le résultat avec une fenêtre de

Hann de longueur 4 secondes (40 trames) avant de sous-échantillonner d'un facteur 5 pour obtenir un vecteur de CENS toute les 0.5 secondes. Nous travaillerons par la suite toujours avec un vecteur de descripteurs toutes les 0.5 secondes donc une fréquence de 2Hz, ainsi un segment de longueur 10 trames dans la matrice d'auto-similarité correspond à un segment de 5s dans le morceau. Pour les matrices d'auto-similarité et les courbes présentées dans les parties suivantes une unité de temps correspond donc à une demi-seconde

Le choix de descripteurs CENS pour nos méthodes d'estimation de la structure est motivé par leur robustesse à plusieurs paramètres comme le timbre, la dynamique et les variations de tempo locales. Les CENS sont les descripteurs utilisés dans [3] et s'avèrent être très efficaces pour l'application du DTW contraint que nous utiliserons par la suite. Nous avons également testé les différentes méthodes de segmentation et regroupement avec d'autres versions des chromas comme par exemple les CRP (Chroma DCT-Reduced log-Pitch)[20] qui ont significativement dégradé les résultats par rapport aux CENS.

3.2 Les deux constructions et traitements de la matrice d'auto-similarité

3.2.1 Construction par distance cosinusoïdale

La première construction de la matrice d'auto-similarité, ainsi que la plus utilisée dans nos méthodes, utilise une distance cosinusoïdale. Nous posons $S(i,j)$ comme la distance cosinusoïdale entre V_i et V_j. Les vecteurs de chroma ont été normalisés par norme l_2 lors de leur construction, $S(i,j)$ est donc directement égal au produit scalaire entre les vecteurs de CENS aux instants t_i et t_j.

$$S(i,j) =< V_i, V_j > \tag{1}$$

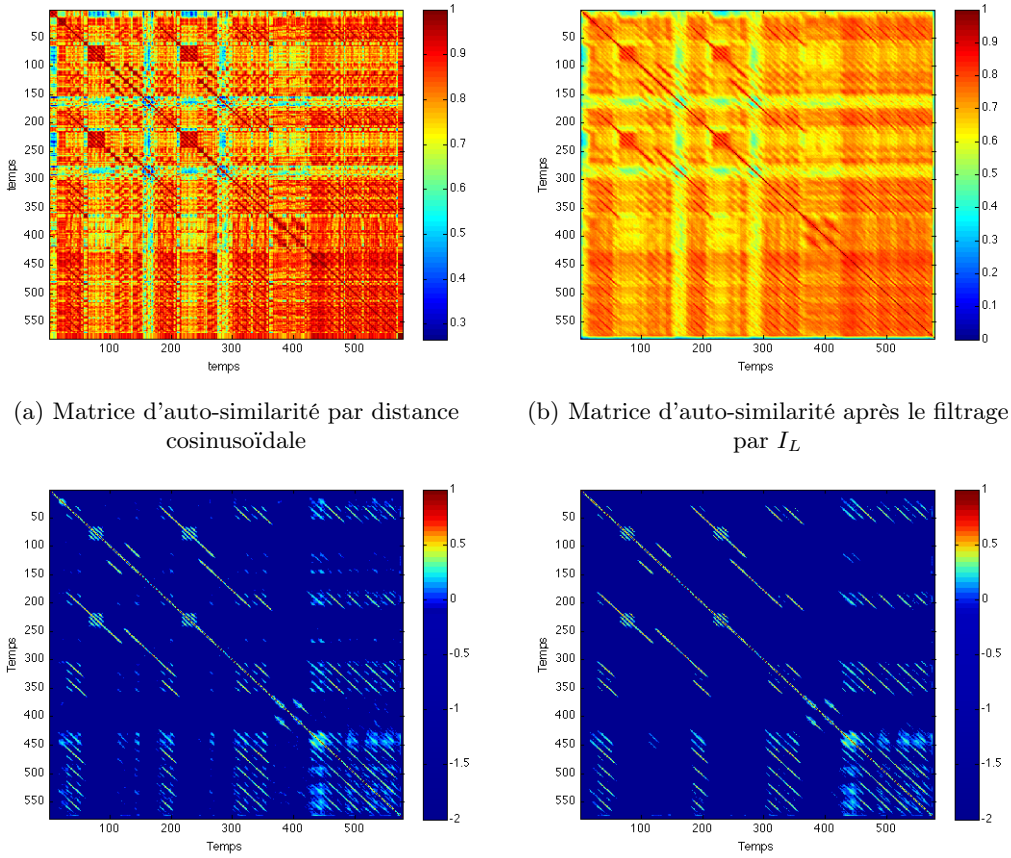
Les méthodes utilisées pour l'estimation de la structure ne peuvent fonctionner qu'avec une représentation particulière de la matrice d'auto-similarité. Il faut trouver des traitements qui permettent au maximum d'éliminer dans S tout ce qui ne correspond pas à une diagonale pour ne garder uniquement que l'information nécessaire à la détection des séquences répétées du morceau. Nous commençons par filtrer S par un filtre passe bas dans la direction de la diagonale principale et passe haut dans la direction orthogonale. Nous convoluons simplement la matrice S par la matrice unité I_L de taille L à laquelle nous ajoutons des valeurs négatives (souvent à -0.3) sur les deux diagonales adjacentes à la diagonale principale. Ce filtrage permet de faire ressortir les diagonales importantes de la matrice tout en réduisant les effets de blocs dans S. Nous prendrons souvent L=12, donc un filtrage passe bas sur 6 secondes.

$$S_{filtre} = S \star I_L \tag{2}$$

Grâce au filtrage, les points de S étant sur des diagonales possèdent les valeurs les plus importantes parmi toutes celles de S. Pour ne garder que les diagonales, nous n'allons garder qu'un pourcentage ρ des valeurs de la matrice filtrée. Nous cherchons ainsi un seuil τ qui permet de garder $\rho\%$ des points. Les points en dessous de ce seuil sont mis à $\delta = -2$ pour introduire une pénalisation utile pour l'algorithme DTW. Les points dans $[\tau : 1]$ sont remis linéairement à l'échelle dans $[0 : 1]$

$$S_{seuil}(i,j) = \begin{cases} -2, & S_{filtre}(i,j) < \tau \\ \frac{S_{filtre}(i,j)-\tau}{1-\tau}, & S_{filtre}(i,j) > \tau \end{cases} \tag{3}$$

Enfin nous espérons pouvoir enlever toute l'information non diagonale restante après l'opération de seuillage qui pourrait perturber la suite des calculs. Pour cela nous effectuons un filtrage médian diagonal. Si dans le voisinage diagonal de $S_{seuil}(i,j)$, il y a plus de la moitié des points à -2, alors $S_{seuil}(i,j) = -2$. Cette opération supprime le bruit : les points et les lignes non diagonales et les bords des blocs restants dans la matrice après le seuillage. Dans les parties suivantes, nous appellerons S la matrice d'auto-similarité obtenue après avoir effectué tous les traitements de mise en valeur des diagonales. Donc pour la construction par distance cosinusoïdale S correspondra à la matrice après le filtrage médian diagonal. Nous présentons Figure 1 l'allure d'une matrice d'auto-similarité après chaque étape de sa construction.



(a) Matrice d'auto-similarité par distance cosinusoïdale

(b) Matrice d'auto-similarité après le filtrage par I_L

(c) Matrice d'auto-similarité après le seuillage

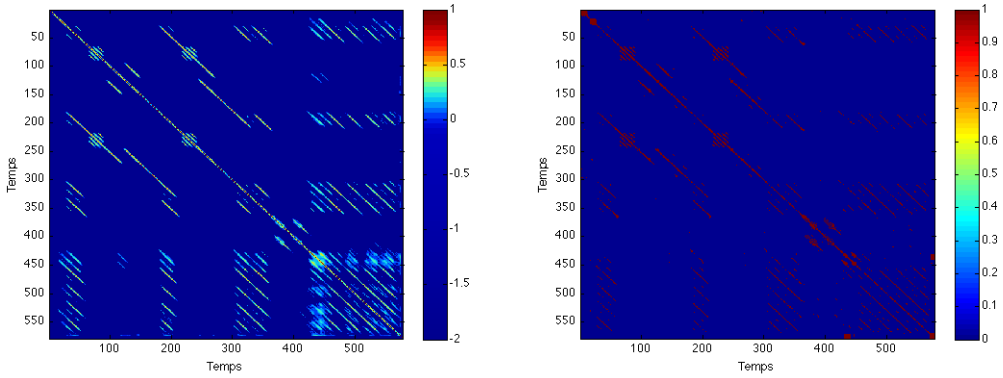
(d) Matrice d'auto-similarité finale après le débruitage diagonal

FIGURE 1 – Illustrations des 4 étapes de la construction et des traitements de la matrice d'auto-similarité pour le morceau 19 de la base RWC POP

3.2.2 Construction par K plus proches voisins

Nous introduisons ici une deuxième construction possible de la matrice d'auto-similarité basée sur la construction de Serra et al. [24] que nous utiliserons lors de l'étape de segmentation partie 4.1. Serra et .al possédant les meilleurs résultats de l'état de l'art en matière de segmentation, nous reprenons leur construction de matrice d'auto-similarité afin de reproduire le plus fidèlement possible leur méthode de segmentation. Cette reproduction nous permettra d'appliquer notre nouvelle pondération à une bonne méthode de segmentation en espérant obtenir de meilleurs résultats.

Les vecteurs de chromas utilisés pour cette construction prennent en compte l'information sur le passé récent du signal en concaténant aux vecteurs de chromas V_j (à la trame j au temps d'analyse t_j) les $m \in \mathbb{N}$ vecteurs précédents. Nous formons ainsi des nouveaux vecteurs à $m \times 12$ dimensions $\hat{V}_j = [V_j, V_{j-1}, ..., V_{j-m}]$, nous avons donc la nouvelle décomposition du signal en chromas $\hat{V} = (\hat{V}_1, .., \hat{V}_N)$. Dans [24] les auteurs utilisent une autre version plus complexe des chromas, les HPCP (harmonic pitch class profiles) [21], nous utiliserons les CENS tout en se rappelant que nous ne pourrons donc pas reproduire exactement les résultats. Nous construisons ensuite S sans utiliser directement de distances mais en utilisant une méthode de K plus proche voisins. Pour chaque vecteur $\hat{V}_i$, nous cherchons ses κ plus proches voisins parmi les autres vecteurs de chromas concaténés à l'aide d'une distance cosinusoïdale. Nous posons $S(i, j) = 1$ si $\hat{V}_i$ fait partie des κ plus proches voisins de $\hat{V}_j$ et réciproquement. Les valeurs de m que nous utilisons se situent entre 2 et 10, nous pouvons donc prendre en compte le passé jusqu'à 5 secondes pour la concaténation de nos descripteurs. Nous choisissons κ de sorte à ne garder que 2 à 5% des valeurs de $\hat{V}$ pour la recherche des plus proches voisins. Nous comparons Figure 2 l'allure d'une matrice d'auto-similarité pour les deux constructions. Contrairement à la matrice construite par distance cosinusoïdale après traitements, cette construction par K plus proches voisins nous donne un matrice binarisée (avec 0 et 1 comme valeurs possibles) et sans valeurs négatives. Les étapes de la construction des deux matrices d'auto-similarité sont résumées sur la Figure 3.



(a) Matrice d'auto-similarité par distance cosinusoïdale après traitements

(b) Matrice d'auto-similarité par K plus proches voisins

FIGURE 2 – Comparaison entre les deux constructions de la matrice d'auto-similarité

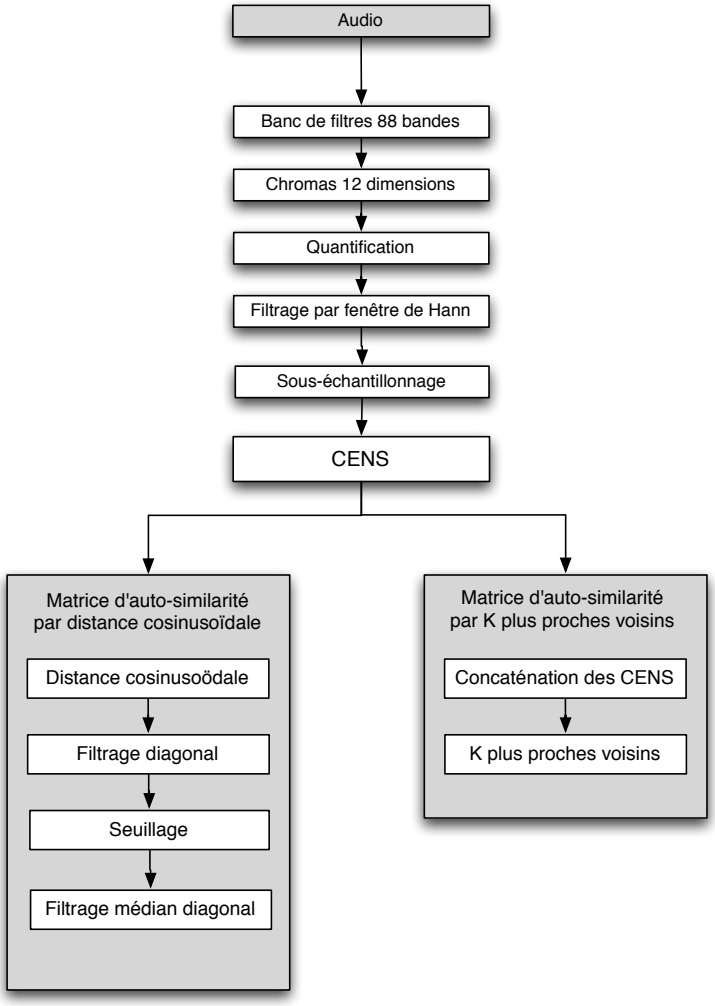


FIGURE 3 – Les étapes de la construction des deux matrices d'auto-similarité

4 Segmentation

La première étape d'un système d'estimation de la structure est souvent la segmentation, c'est-à-dire la recherche de frontières temporelles entre différentes parties d'un morceau (introduction, couplet, refrain...). La segmentation permet de délimiter les différentes parties du morceau afin de permettre à l'étape de regroupement de les classer en cherchant des répétitions parmi les segments trouvés. L'étape de segmentation se base souvent sur un calcul de nouveauté dans le signal temporel, il s'agit de repérer les temps où l'information contenue dans le signal est fortement modifiée. Pour l'approche par état les temps de nouveauté correspondent au passage d'une suite de temps homogènes vers une autre (le passage entre deux blocs dans la matrice d'auto-similarité). Pour l'approche par séquences un temps de nouveauté correspond au passage d'un segment répété à un autre, plus le segment est répété dans le morceau plus la nouveauté sera importante sur les instants de début et de fin de ce segment. Notre méthode de segmentation s'inspire de deux travaux présentés brièvement dans l'état de l'art, ceux de Serra et al. [24] et ceux de Goto [2] se basant sur l'aspect répétitif du contenu musical. Les deux méthodes utilisent une représentation en temps-retard de la matrice d'auto-similarité appelée "matrice temps-retard" (lag-matrix). Après avoir présenté plus en détails le fonctionnement de ces deux méthodes nous expliquerons comment se servir de la courbe de cumul des retards de la première comme information a priori pour les calculs de nouveauté de la deuxième.

La nouvelle pondération sera appliquée sur deux matrices de retard différentes issues des deux constructions de la matrice d'auto-similarité présentée partie 3.2.1. Nous présenterons le fonctionnement et les résultats de la méthode pour les deux constructions des matrices de retard. Nous présenterons également une modification qui ajoute l'utilisation de techniques de calcul de nouveauté issues de l'approche par état que nous activons à l'aide d'une détection d'approche état/séquence présentée dans [9].

4.1 Fonctionnement des systèmes de segmentation utilisés

4.1.1 Les matrices temps-retard

Nous avons parlé jusque-là uniquement de matrice d'auto-similarité sur des axes temps-temps où $S(i,j)$ traduit la ressemblance entre le temps d'analyse t_i et le temps d'analyse t_j. Les matrices de retard notées L sont sur un axe temps-retard, c'est à dire que la valeur $L(i,j)$ correspond à la ressemblance entre le signal au temps d'analyse t_i et le signal au temps d'analyse t_{i-j} pour $(i-j) > 0$. Les matrices de temps-retard ne possèdent de l'information que sur la partie triangulaire inférieure. Elles sont construites directement à partir de la matrice d'auto-similarité S :

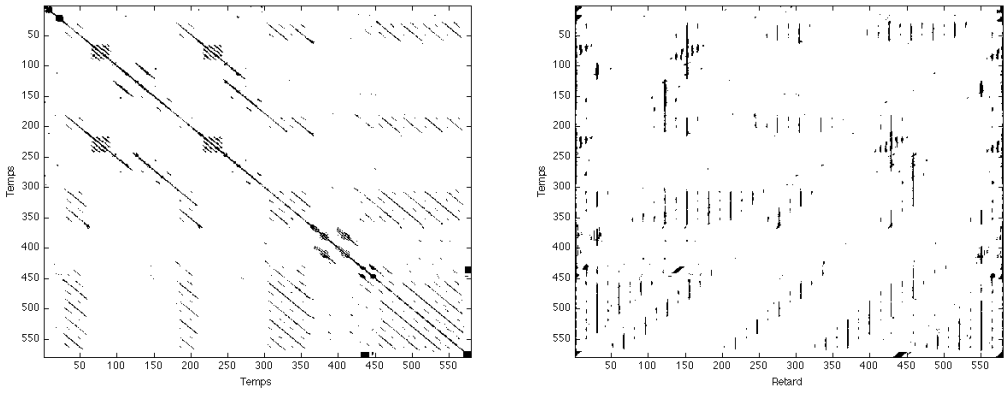
$$L(i,l) = L(t_i, l = t_j - t_i) = S(i, j = i + l) \text{ pour } i + l \leq N \text{ , } (i,l) \in \mathbb{N}^* \times \mathbb{N}^* \tag{4}$$

Afin de compléter la moitié manquante de L nous construisons une matrice de temps-retard circulaire L^* qui prend également en compte les retards négatifs (en prenant en compte le futur et le passé des observations), cela se fait par permutation circulaire des colonnes de S (Figure 4).

$$L^*(i,j) = S(i, k+1) \text{ où } k = i + j - 2 \tag{5}$$

4.1.2 L'approche retard de Goto

Dans [2] Goto propose de détecter automatiquement la position des refrains dans un morceau de musique en faisant l'hypothèse qu'ils correspondent aux parties les plus représentées et donc les



(a) Matrice d'auto-similarité par K plus proches voisins

(b) Matrice temps-retard circulaire

FIGURE 4 – Transformation de la matrice d'auto-similarité en matrice temps-retard circulaire - Les points noirs ont une valeur de 1 et les points blancs une valeur de 0

plus répétées d'un morceau. Sa méthode se base sur une détection de lignes dans la matrice temps-retard, l'objectif étant d'avoir une liste de toutes les répétitions existantes afin de sectionner, selon d'autres hypothèses, celles qui ont le plus de chances de correspondre à un refrain. Nous reprendrons uniquement sa méthode de détection des parties répétées afin de l'utiliser dans nos méthodes de segmentation.

Goto travaille avec une matrice temps-retard L calculée elle aussi à partir de vecteurs de chromas. Les répétitions apparaissent comme des lignes parallèles à l'axe des temps (retard constant) dans une matrice temps-retard. L'idée de Goto pour détecter les répétitions est d'estimer les temps de début et de fin des lignes présentes dans la matrice L. Pour faire cela il commence par calculer sur quelles valeurs de retards il est possible d'avoir des répétitions. Si nous n'avons pas de variation de tempo, les répétitions, les lignes parallèles à l'axe des temps, sont constantes en retard. Son idée est donc de cumuler les valeurs des colonnes de L pour calculer la fonction de retard f :

$$f(l) = \sum_{t_i=1}^{N-l} \frac{1}{N-l} L(t_i, l) \tag{6}$$

A l'aide d'un seuil et d'une méthode de sélection de pics il cherche des maximums locaux de la fonction f. Pour un pic $f(l_k)$ retenue pour le retard l_k nous allons analyser la colonne $L(:, l_k)$ afin d'estimer la position de répétitions existantes à ce retard. Nous nous intéressons seulement à la construction de la fonction f qui nous permettra d'affiner la recherche de répétitions de Serra et al.

4.1.3 L'approche temps de Serra et al.

L'approche de Serra se sert de la matrice temps-retard circulaire L^*. Nous rappelons que cette méthode possède les meilleurs résultats de l'état de l'art en segmentation pour l'estimation de la

structure musicale. Elle vise à estimer la position des frontières entre des parties homogènes ou répétées dans une série temporelle. L'objectif est d'avoir accès à une courbe de nouveauté temporelle qui indique, par la présence de maximums locaux, un changement de région dans la série temporelle. Malgré la volonté des auteurs de créer une méthode de segmentation polyvalente, son efficacité a surtout été prouvée pour l'analyse de structure musicale. De plus les descripteurs utilisés ainsi que les traitements effectués sur les matrices d'auto-similarité favorisent une approche par séquence. Il s'agit donc d'une méthode adaptée à notre recherche que nous espérons améliorer en utilisant l'information issue de la courbe de retard de Goto.

Comme nous avons pu le voir précédemment la présence de répétitions dans le signal fait apparaître des lignes dans la matrice temps-retard. Si nous posons N comme la taille de la matrice carrée L^*, Serra définit un "vecteur de structure" (structure feature) $g(i) = L(i,:)$ qui correspond à la i-ème ligne de L^* au temps d'analyse t_i. L'objectif est de repérer de grosses variations des descripteurs de structures sur des temps de début de répétitions. Pour cela nous calculons une courbe de nouveauté c comme la différence trame à trame à travers tout le signal des vecteurs de structures.

$$c(i) = ||g(i+1) - g(i)||^2 \tag{7}$$

Le calcul de la courbe de nouveauté revient à effectuer la norme de la différence ligne à ligne de la matrice de temps-retard circulaire. Une grande valeur de $c(i)$ indique un changement de région dans le morceau à l'instant d'analyse t_i ce qui correspond le plus souvent à un début ou une fin de répétition. La méthode de Serra utilise le fait que les segments répétés entre eux créent des lignes commençant toutes sur le même temps dans L^* et augmentent ainsi la différence trame à trame une fois arrivée sur ce temps. Le fonctionnement des deux méthodes sur une matrice temps-retard circulaire est présenté Figure 5.

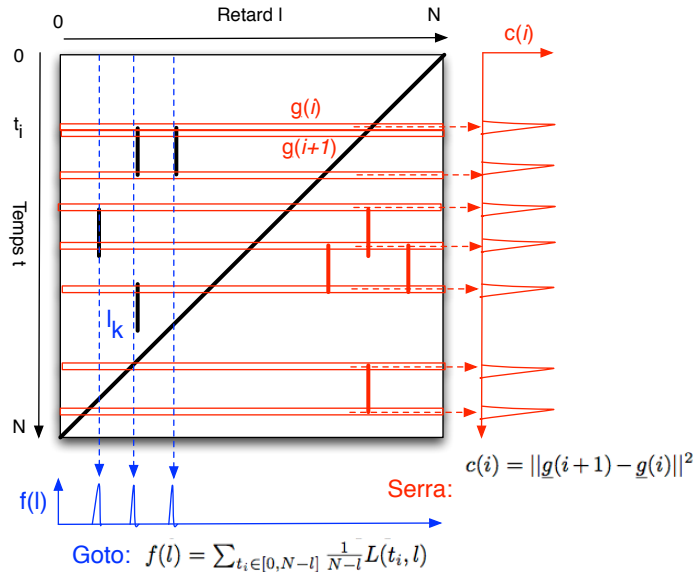


FIGURE 5 – Fonctionnement des deux approches sur une matrice temps-retard circulaire - **Bas** Courbe de retard de Goto calculée à partir de la somme cumulée des valeurs de chaque colonnes de la matrice - **Droite** Courbe de nouveauté de Serra calculée à partir de la différence ligne à ligne de la matrice

4.2 La méthode de segmentation proposée

Nous souhaitons maintenant fusionner ces deux méthodes pour créer une nouvelle méthode de segmentation plus robuste au bruit. Le bon fonctionnement de la méthode de Serra et al. suppose que les vecteurs de structures $g(i)$ ne sont pas bruités, c'est à dire qu'ils ont des valeurs élevées quand ils croisent une répétition et sont nulles sinon. Ceci est rarement le cas car dans les matrices d'auto-similarité, malgré les efforts de filtrage et de seuillage, il reste toujours de l'information ne correspondant pas à des répétitions de segments. Cette information se traduit par l'apparition de points ou de petites lignes dans L^* qui viennent perturber les vecteurs de structures en ajoutant des grandes valeurs à $g(i)$ à des temps ne faisant pas partie de segments répétés.

Afin de modéliser l'impact du bruit sur la courbe de nouveauté de Serra et al. nous posons $\hat{g}(i)$ comme étant la seule contribution des segments à laquelle nous ajoutons la contribution du bruit (tout ce qui n'est pas un segment) que nous modélisons par un bruit Gaussien centré $\mathcal{N}_{\mu=0,\sigma}$ nous avons

$$g(i) = \hat{g}(i) + \mathcal{N}_{\mu=0,\sigma} \tag{8}$$

Nous pouvons alors en déduire que

$$c(i) = ||g(i+1) - g(i)||^2 = \begin{cases} K + 2\sigma^2 \text{ si } K \text{ segments débutent/terminent à l'instant } t_i \\ 2\sigma^2 \text{ sinon} \end{cases} \tag{9}$$

Nous remarquons que la facilité à distinguer dans quel cas nous nous situons dépend grandement du taux de bruit de fond dans la matrice d'auto-similarité. Le calcul de la courbe de nouveauté de Serra et al. est donc particulièrement sensible au bruit. En pratique il suffit que quelques points isolés apparaissent dans la matrice de temps-retard sur la ligne $L^*(i,:) = g(i)$ pour que la valeur de $c(i)$ augmente. Les points isolés ne correspondent pas à des débuts et fins de segments et peuvent entraîner des erreurs dans l'étape de détection de maximums locaux de c. Nous espérons régler ce problème en reprenant la courbe de retard de Goto. Dans le cas d'un bruit blanc Gaussien centré, l'espérance des valeurs de $f(l) = \sum_{t_i=1}^{N-l} \frac{1}{N-l} L(t_i, l)$ est indépendante de σ. Nous nous en servirons comme information a priori sur la possibilité d'avoir un segment au retard l dans le calcul de Serra et al.

Pour intégrer cette pondération, nous considérons la matrice de temps-retard circulaire comme une distribution de probabilité jointe. Nous posons ainsi $p(t,l) = L^*(t,l)$ comme une estimation de la possibilité d'être sur la répétition d'un segment au temps t_i avec un retard de l par rapport au segment. Nous nommons c_1 la courbe de nouveauté de Serra et al.

$$c_1(t) = \int_l |\frac{\partial}{\partial t} p(t,l)|^2 dl \tag{10}$$

En ajoutant l'information a priori apportée par $f(l)$, nous obtenons une courbe de nouveauté c_2 qui favorise la détection de début/fin de segments lorsque qu'on est sur une forte valeur de f.

$$c_2(t) = \int_l p(l) |\frac{\partial}{\partial t} p(t,l)|^2 dl \tag{11}$$

Ici $p(l)$ correspond à la courbe de retard de Goto mais calculée cette fois sur L^* et non sur L. Ainsi $p(l)$ correspond à la marginale de $p(t,l)$ sur le temps

$$p(l) = \int_{t=1}^{t=N} p(t,l) dt \tag{12}$$

Ainsi la courbe de retard f se comporte comme une pondération de chaque dimension de $g(i)$, nous obtenons des vecteurs de structures de type $g_2(i,l) = f(l) \times g(i,l)$. Nous accordons une importance relative à la l-ème dimension $g(i)$, plus nous avons de répétitions sur le retard l (des lignes dans L^*) plus la l-ème dimension de $g(i)$ aura de l'importance dans la différence trame à trame.

L'utilisation de la courbe de Goto ne règle pas complètement le problème de la présence de bruit (points isolés) car si nous trouvons un point isolé en $L^*(t,l)$ et que $f(l)$ est proche de 1 alors nous aurons tout de même une forte valeur $c_2(t)$ lors de la différence trame à trame. Nous supprimons donc l'impact du bruit uniquement sur les colonnes de la matrice de temps-retard circulaire ne contenant pas de répétitions. Nous proposons pour cela une solution réutilisant la courbe de Goto mais calculée localement. L'objectif est de limiter l'impact des répétitions éloignées dans le temps, créant des fortes valeurs dans f, d'avoir une influence sur toutes les différences trame à trame à travers le temps.

Nous introduisons une troisième courbe de nouveauté c_3 qui correspond l'ajout de l'information a priori locale apportée par f. Nous définissons une courbe de retard locale $p_t(l)$ que nous allons utiliser comme a priori pour le calcul de c_3 de manière équivalente à c_2 :

$$p_t(l) = \int_{\tau=t-\Delta}^{\tau=t-\Delta} p(\tau, l) d\tau \tag{13}$$

Ce qui nous donne la courbe de nouveauté c_3

$$c_3(t) = \int_l p_t(l) |\frac{\partial}{\partial t} p(t, l)|^2 dl \tag{14}$$

Nous prenons en compte la présence de répétitions pour le calcul de $c_3(t)$ 10 secondes avant et après la trame t, nous avons donc $\Delta = 20$. Ainsi $p_t(l)$ est calculée de la même manière que la courbe de retard de Goto mis à part que nous la calculons sur une fenêtre centrée en t et non sur tout le signal. Cela permet de pondérer par la possibilité d'avoir un segment sur le retard l uniquement dans le voisinage temporel de $g(i)$ et donc de ne pas prendre en compte ce qui se passe dans le passé ou le futur lointain de la différence trame à trame en t_i.

Pour extraire la position des frontières nous appliquons la même stratégie d'extraction de pics sur les trois courbes de nouveauté c_1,c_2 et c_3. Avant la recherche de pics nous commençons par contraindre l'étendue des valeurs des courbes c dans $[0 : 1]$ en posant $c = \frac{c}{\max_{1 \leq i \leq N} c(i)}$. Nous cherchons ensuite toutes les positions des maximums locaux des courbes qui sont au-dessus d'un certain seuil θ fixé à 0.1. Nous imposons également une distance minimale de 10 secondes entre la détection de deux maximums pour limiter le nombre de pics détectés lors de la présence de maximums locaux trop rapprochés dans les courbes de nouveautés. Les statistiques des bases de tests utilisées montrent que l'intervalle moyen entre deux frontières va de 13.7 secondes à 16.0 secondes. Un exemple de sélection de pics sur une courbe de nouveauté est présenté Figure 6. Nous résumons également toutes les étapes de notre méthode de segmentation Figure 8.

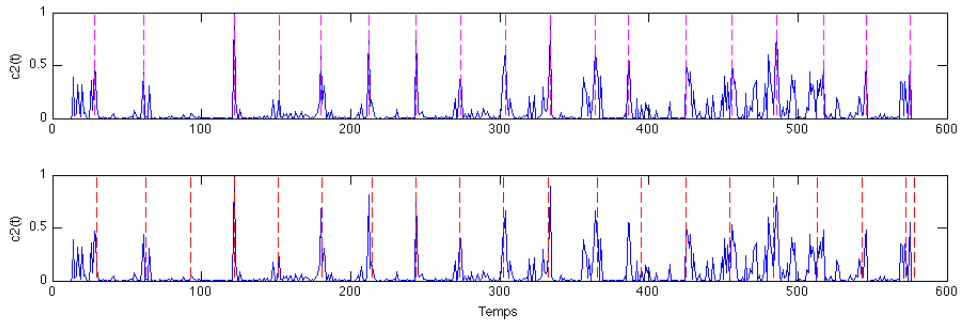


FIGURE 6 – Illustration de la sélection de pics sur la courbe c_2 (avec pondération locale) sur le morceau 19 de RWC POP - **Haut** Courbe c_2 avec les lignes pointillées aux positions des frontières estimées - **Bas** Courbe c_2 avec les lignes pointillées aux positions des frontières de l'annotation

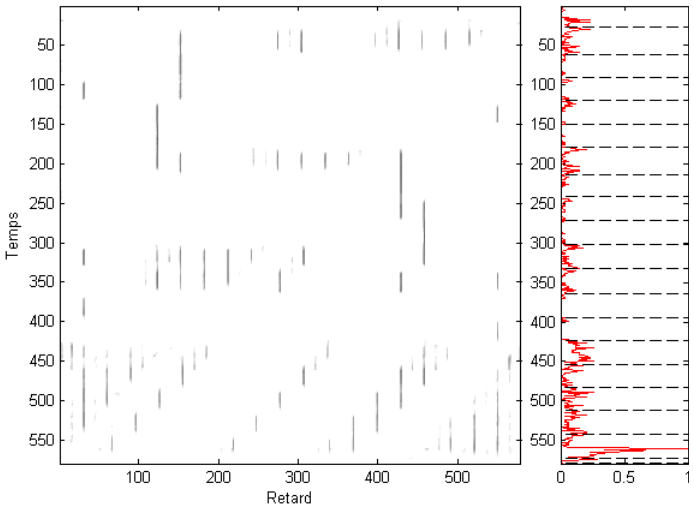
Nous obtenons finalement une liste de temps d'analyse correspondant aux frontières entres les segments de la structure. Ce sont ces positions de frontières que nous utiliserons pour évaluer nos méthodes de segmentation.

Sur les figures de la Figure 7 nous illustrons le calcul des trois courbes c_1,c_2 et c_3 pour le morceau 19 de la base RWC POP.

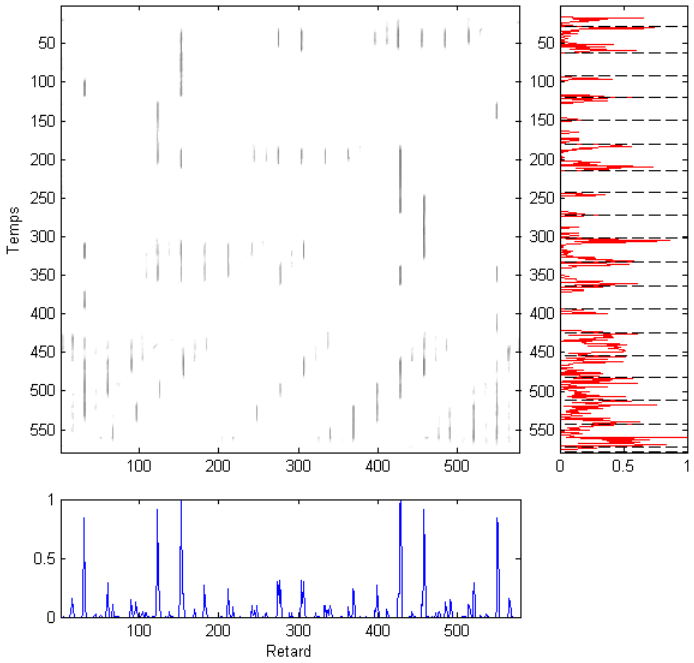
Sur la Figure 7 (a) nous présentons la méthode de Serra et al. [24]. Nous avons à gauche la matrice temps-retard circulaire et à droite la courbe c_1 (courbe rouge) à laquelle nous avons superposé la positon des frontières de l'annotation (lignes pointillées noires).

Sur la Figure 7 (b) nous présentons le calcul de la courbe c_2 (avec la pondération globale). En dessous de la matrice temps-retard circulaire nous représentons la courbe de retard de Goto $p(l)$ (courbe bleu). À droite la courbe c_2 calculée avec la pondération globale par la courbe de Goto. Nous remarquons que la pondération permet une meilleure discrimination des pics correspondant aux débuts et fin de segments.

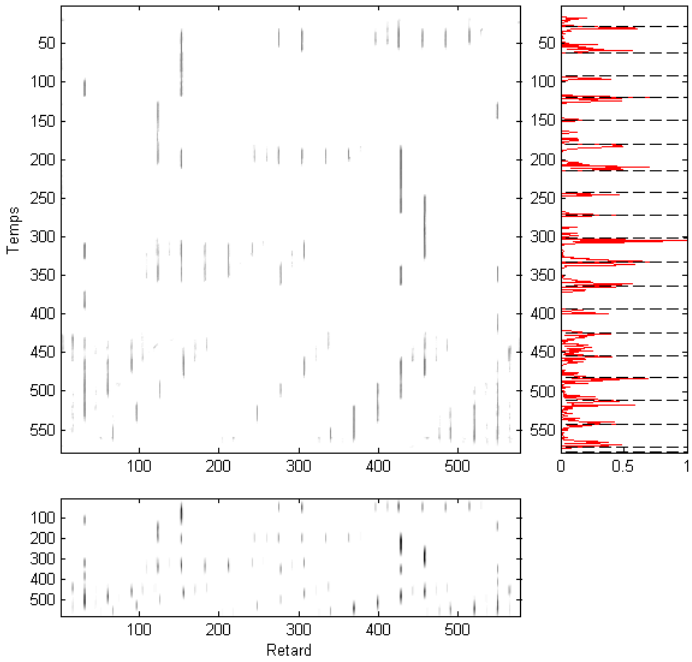
Sur la Figure 7 (c) nous présentons le calcul de la courbe c_3 (avec la pondération locale). En dessous de la matrice temps-retard circulaire nous représentons la courbe de retard de Goto locale $p_t(l)$ sous forme d'une matrice A où $A(t,l) = p_t(l)$. À droite la courbe c_3 calculée avec la pondération par la courbe de Goto calculée localement. Nous remarquons que la pondération permet une encore meilleure discrimination des pics par rapport à la pondération globale (courbe c_2).



(a) Calcul de $c_1(t)$



(b) Calcul de $c_2(t)$



(c) Calcul de $c_3(t)$

FIGURE 7 – Illustration du calcul des trois courbes c_1,c_2 et c_3 pour le morceau 19 de la base RWC POP - Voir partie 4.2 pour les explications

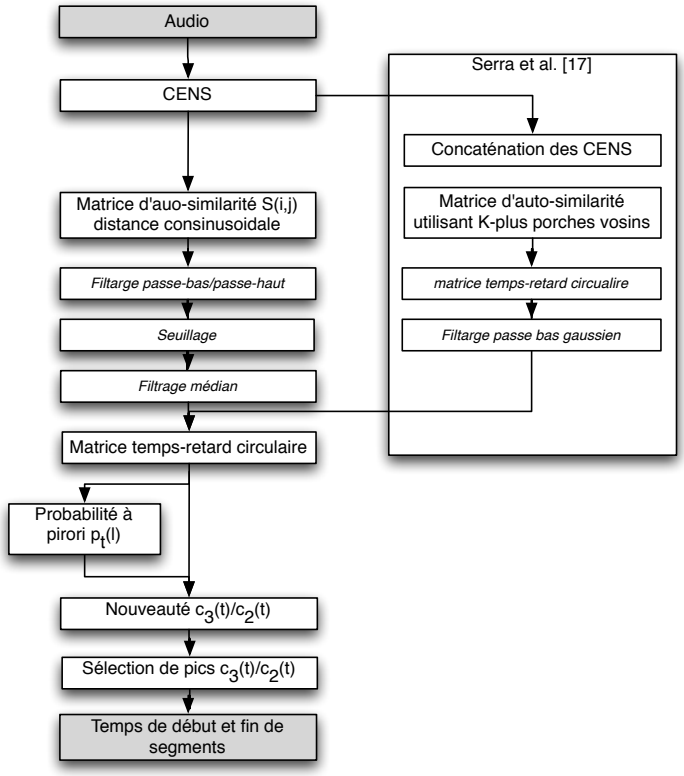


FIGURE 8 – Les étapes de la nouvelle méthode de segmentation proposée

4.3 Évaluation de la segmentation

4.3.1 Mesures pour l'évaluation

Afin de comparer nos résultats à ceux de l'état de l'art nous reprenons les mêmes mesures d'évaluation de la segmentation que celles utilisées dans la campagne d'évaluation annuelle MIREX (Music Information Retrieval Evaluation eXchange) ainsi que dans la plupart des évaluations de méthodes de segmentation présentent dans l'état de l'art. Nous commençons par définir deux ensembles :

— $\mathcal{A} = \{\tau_{a1}, ..., \tau_{an_a}\}$ l'ensemble des n_a temps correspondant aux frontières de segments dans l'annotation

— $\mathcal{E} = \{\tau_{e1}, ..., \tau_{en_e}\}$ l'ensemble des n_e temps correspondant aux frontières de segments dans l'estimation

Nous calculons une précision P qui mesure la quantité de frontières détectées faisant partie de l'annotation et un rappel R qui mesure la quantité de frontières de l'annotation présentes dans l'estimation. Les mesures de précision et de rappel autorisent deux frontières à être supposées identiques si elles sont à moins de λ secondes l'une de l'autre. La longueur λ de la fenêtre de tolérance est fixée à 0.5 ou 3 secondes par l'état de l'art, nous effectuerons les tests avec les deux

longueurs. Nous considérons donc que $\tau_{ai} = \tau_{ej}$ si $|\tau_{ai} - \tau_{ej}| > \lambda$.

$$P = \frac{|\mathcal{A} \cap \mathcal{E}|}{|\mathcal{E}|} \tag{15}$$

$$R = \frac{|\mathcal{A} \cap \mathcal{E}|}{|\mathcal{A}|} \tag{16}$$

$$F = \frac{2.P.R}{(P + R)} \tag{17}$$

4.3.2 Résultats avec la matrice d'auto-similarité par distance cosinusoidale

Les premiers tests que nous réalisons pour les trois courbes de segmentation se font avec la matrice d'auto-similarité construite par distance cosinusoidale telle que nous l'avons présentée partie 3.2.1 . Nous la transformons, après les traitements de mises en valeurs des diagonales, en matrice temps-retard circulaire. Le tableau Table 3 contient les scores de précision, rappel et f-mesure pour les deux fenêtres de tolérance, à 3 secondes et à 0.5 secondes. Nous comparons nos résultats aux meilleurs résultats des deux dernières années de la campagne d'évaluation annuelle MIREX (2012 et 2013) [1] ainsi que ceux publiés par Serra et al. [24] dont s'inspire notre méthode.

Nous pouvons remarquer que pour les trois bases de tests l'ajout d'information a priori sur les retards améliore la segmentation, $c_3 > c_2 > c_1$.

BEATLES : c_3 :76,1% c_2 :69,8% c_1 :65,7%
RWC-POP-A : c_3 :76,9% c_2 :72,9% c_1 :66,0%
RWC-POP-B : c_3 :78,2% c_2 :72,6% c_1 :67,3%

Les meilleurs scores que nous obtenons en terme de f-mesure, pour la tolérance de 3 et 0.5 secondes, sont avec la courbe c_3 qui utilise l'information sur les retards localement. Les résultats en f-mesure, avec la pondération locale, sont meilleurs que ceux obtenus dans les deux dernières années de MIREX. En revanche ils sont légèrement inférieurs à ceux obtenus par Serra et al. Nous pouvons également remarquer que la courbe c_1, qui reprend leurs calculs est largement en dessous de leurs résultats. Nous avons calculé la courbe c_1 en utilisant la matrice d'auto-similarité par distance cosinusoïdale (ils utilisent celle construite par K plus proches voisins avec un filtrage particulier). Nous reprendrons donc cette matrice de temps-retard circulaire en espérant s'approcher des résultats de Serra et al. sur la courbe c_1. Comme nos meilleurs résultats sur c_3 sont légèrement en dessous de ceux de [24] nous espérons qu'en reprenant leur matrice de temps-retard, nous pourrons nous approcher de leurs résultats pour c_1 et ainsi les améliorer à l'aide de l'ajout de l'information a priori sur les retards.

4.3.3 Résultats avec la matrice temps-retard par K plus proches voisins

Le deuxième test que nous effectuons est basé sur la matrice d'auto-similarité construite par K plus proches voisins tel que dans l'article de Serra et al. Après avoir construit la matrice de temps-retard circulaire, ils effectuent un dernier traitement permettant de mettre en valeur les segments

1. SMGA1 correspond à [Joan Serra, Meinard Mueller, Peter Grosche,JosepLluis Arcos]. FK2 correspond à [Florian Kaiser et Geoffroy Peeters]. RBH1 correspond à [Bruno Rocha, Niels Bogaards, Aline Honingh].

TABLE 3 – Résultats de l'évaluation avec notre matrice

RWC-Pop-A						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.791	0.0.817	0.783			
MIREX-2012 (SMGA1)	0.710	0.741	0.701	0.236	0.247	0.232
MIREX-2013 (FK2)	0.657	0.816	0.56	0.301	0.376	0.256
$c_1(t)$ (sans pondération)	0.660	0.700	0.648	0.315	0.338	0.308
$c_2(t)$ (avec pondération globale)	0.729	0.739	0.737	0.349	0.354	0.353
$c_3(t)$ (avec pondération locale)	**0.769**	0.770	0.78	**0.386**	0.392	0.390
RWC-Pop-B						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.8	0.81	0.805			
MIREX-2012 (SMGA1)	0.766	0.816	0.735	0.268	0.287	0.256
MIREX-2013 (RBH1)	0.673	0.701	0.664	0.375	0.392	0.368
$c_1(t)$ (sans pondération)	0.673	0.6745	0.689	0.238	0.223	0.263
$c_2(t)$ (avec pondération globale)	0.726	0.704	0.766	0.250	0.231	0.281
$c_3(t)$ (avec pondération locale)	**0.782**	0.782	0.816	**0.281**	0.264	0.31
Beatles						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.774	0.76	0.807			
$c_1(t)$ (sans pondération)	0.657	0.674	0.658	0.232	0.240	0.238
$c_2(t)$ (avec pondération globale)	0.698	0.696	0.718	0.254	0.258	0.265
$c_3(t)$ (avec pondération locale)	**0.761**	0.745	0.795	**0.262**	0.259	0.278

à l'aide d'un filtrage passe-bas. Nous construisons pour cela un noyau gaussien unidimensionnel de longueur 30 secondes et de variance $\sigma^2 = 0.16$ de sorte que le noyau soit maximal au centre et s'approche de zéro sur les bords. Nous utilisons ce noyau gaussien comme filtre passe-bas que nous appliquons, par une convolution, sur chaque colonne de L^* afin de mettre en valeur les segments tout en réduisant le bruit. La longueur du filtre est choisie telle que la moitié de la longueur corresponde à peu près à la distance moyenne entre deux frontières de l'annotation. Comme nous l'avons présenté dans les statistiques des bases de test (Tableau 2), l'intervalle moyen entre deux annotations se situe autour de 15 secondes ce qui justifie le choix de Serra et al. d'utiliser un filtre de longueur totale 30 secondes. Il est important de noter que contrairement à la première matrice temps-retard celle-ci ne possède pas de valeurs négatives mais des valeurs nulles sur les points ne correspondant pas à des répétitions.

Nous appliquons maintenant le calcul des trois courbes de nouveauté à la matrice L^* filtrée. Les résultats exposés en comparaison sont les mêmes que ceux du tableau précédent. Nous illustrons Figure 9 la différence d'allure entre une courbe de nouveauté calculé à partir des deux matrices. La courbe issue de la construction de Serra et al. (celle du bas) est moins bruitée. Elle permet une détection de pics plus facile par un écartement moyen plus grand entre deux maximums locaux et donc permet une segmentation plus précise.

Nous remarquons encore que l'ajout d'information sur le retard améliore les résultats mais en

TABLE 4 – Résultats de l'évaluation avec la matrice de Serra et al.

RWC-Pop-A						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.791	0.0.817	0.783			
MIREX-2012 (SMGA1)	0.710	0.741	0.701	0.236	0.247	0.232
MIREX-2013 (FK2)	0.657	0.816	0.56	0.301	0.375	0.256
$c_1(t)$ (sans pondération)	0.780	0.846	0.742	0.254	0.271	0.246
$c_2(t)$ (avec pondération globale)	**0.784**	0.843	0.750	0.289	0.316	0.275
$c_3(t)$ (avec pondération locale)	0.735	0.827	0.682	0.245	0.300	0.215
RWC-Pop-B						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.8	0.81	0.805			
MIREX-2012 (SMGA1)	0.766	0.816	0.735	0.268	0.287	0.256
MIREX-2013 (RBH1)	0.673	0.700	0.664	0.375	0.392	0.368
$c_1(t)$ (sans pondération)	0.799	0.795	0.818	0.338	0.326	0.359
$c_2(t)$ (avec pondération globale)	**0.823**	0.846	0.820	0.389	0.408	0.381
$c_3(t)$ (avec pondération locale)	0.797	0.856	0.765	0.336	0.369	0.318
Beatles-B						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.774	0.76	0.807			
$c_1(t)$ (sans pondération)	0.772	0.792	0.773	0.371	0.365	0.394
$c_2(t)$ (avec pondération globale)	**0.805**	0.813	0.817	0.439	0.430	0.450
$c_3(t)$ (avec pondération locale)	0.799	0.790	0.827	0.422	0.416	0.442

revanche cette fois il s'agit de c_2 avec la pondération globale qui prend le dessus.

BEATLES : c_3 :79,9% c_2 :80,5% c_1 :77,2%
RWC-POP-A : c_3 :73,5% c_2 :78,4% c_1 :78,0%
RWC-POP-B : c_3 :79,7% c_2 :82,3% c_1 :79,9%

Dans la méthode Serra et al., la matrice est filtrée par un filtre passe-bas de 30 secondes alors que le calcul de la courbe de Goto locale se fait dans une fenêtre de 20 secondes. A cause du filtrage passe bas, les valeurs maximales de $p_t(l)$ se situent au milieu des segments alors que nous souhaitons des fortes valeurs sur les bords des segments. Les pics de nouveauté peuvent donc être déplacés vers le centre des segments et peuvent dégrader les scores de la segmentation sur la courbe c_3. Ce phénomène est moins présent lors de l'utilisation de la première matrice temps-retard car la présence de valeurs négatives dans la matrice temps-retard augmente fortement la différence trame à trame sur les instants de début et fin de segments. Ainsi les valeurs négatives empêchent le décalage des pics vers le centre des segments d'avoir lieu.

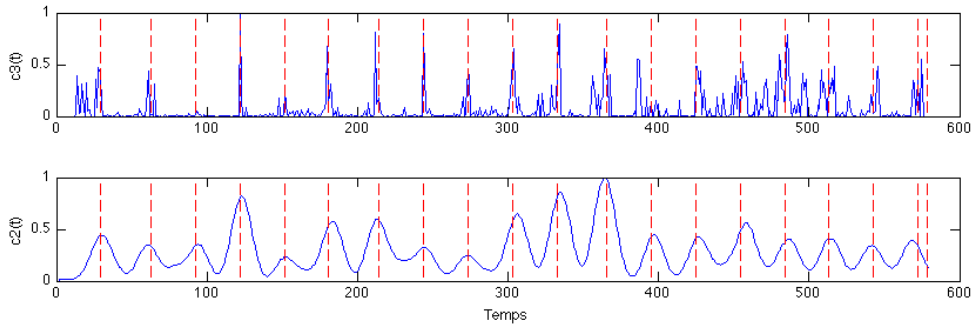


FIGURE 9 – Comparaison entre les deux allures de la courbe c_2 (avec pondération globale) calculées à partir les deux constructions de la matrice d'auto-similarité sur le morceau 19 de RWC POP- **Haut** Courbe c_2 calculée à partir de la matrice d'auto-similarité par distance cosinusoïdale avec les lignes pointillées aux positions des frontières de l'annotation - **Bas** Courbe c_2 calculée à partir de la matrice d'auto-similarité par K plus proches voisins avec les lignes pointillées aux positions des frontières de l'annotation

L'ajout de l'information sur les retards améliore les résultats lors de l'application sur les deux matrices différentes. Nous arrivons même, en reprenant les matrices de Serra et al. , à dépasser les meilleurs résultats de l'état de l'art en segmentation pour les deux fenêtres de tolérance. Nous espérons pouvoir encore l'améliorer avec une méthode de détection d'état présentée dans la partie suivante.

4.4 Amélioration par détection d'état

Nos méthodes de segmentation se basent essentiellement sur l'approche par séquences pour la détection de segments répétés dans les matrices d'auto-similarité. Nous espérons dans cette partie améliorer encore la segmentation en intégrant de l'information issue de méthodes de segmentation adaptées à l'approche par état. Peeters a proposé dans [9] d'estimer automatiquement l'approche (séquence ou état) optimale à adopter à travers le morceau. En effet nous pouvons avoir des régions d'un morceau composées en majorité de blocs homogènes favorisant une approche par état pour l'estimation de la structure du morceau. Il existe des méthodes efficaces pour la détection de blocs (partie homogène dans la matrice d'auto-similarité) que nous reprendrons pour améliorer notre segmentation.

4.4.1 Segmentation pour une approche par état

Les méthodes de segmentation pour l'approche par état se basent sur un calcul de nouveauté qui nous informe sur les changements entre parties homogènes dans le morceau. Il s'agit de repérer les temps auxquels nous quittons un bloc homogène dans la matrice d'auto-similarité pour aller vers un nouveau bloc. Pour cela nous utilisons la "novelty measure" (courbe de nouveauté) de Foote [8], calculée en convoluant un noyau gaussien bivariant sur la diagonale principale de la matrice d'auto-similarité.

Nous posons G le noyau gaussien, une matrice carrée de taille M et S la matrice d'auto-similarité par distance cosinusoïdale sans traitements (pour ne pas supprimer les parties homogènes). Le calcul de la nouveauté n de longueur N se fait le long de la diagonale de S. La courbe de nouveauté correspond à la diagonale principale du résultat du filtrage de S par G. Nous avons donc $n(k)$ la nouveauté à la trame k :

$$n(t) = \sum_{i=-\frac{m}{2}}^{i=+\frac{m}{2}} \sum_{j=-\frac{m}{2}}^{j=+\frac{m}{2}} G(\frac{m}{2}+i, \frac{m}{2}+j)S(t+i, t+j) \tag{18}$$

La matrice G à une structure d'échiquier de taille 2×2.

$$G = \begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix} \tag{19}$$

Où

$$Q_1 = -JQ_3 = -Q_2J = JQ_4J \tag{20}$$

Avec J une matrice $\frac{m}{2} \times \frac{m}{2}$ avec des 1 sur l'anti-diagonale principale et des 0 ailleurs. Dans une approche simplifiée nous pourrions nous contenter de remplir la matrice Q_4 avec des -1 mais nous préférons pondérer les valeurs par une gaussienne radiale accordant moins d'importance aux valeurs éloignées du centre de la matrice. Pour cela nous posons

$$Q_4(x, y) = -\exp(-\frac{r^2}{2\sigma^2}) \tag{21}$$

Où le rayon est défini comme

$$r^2 = \frac{4}{m^2}((x-1)^2 + (y-1)^2) \tag{22}$$

La convolution de G sur la diagonale principale crée une forte valeur de nouveauté $n(k)$ lorsqu'un bloc commence ou se termine à la trame k. Nous présentons l'allure du noyau de Foote Figure 10 ainsi qu'un exemple de calcul de nouveauté par convolution de ce noyau sur une matrice d'auto-similarité Figure 11. La courbe de nouveauté (en bas de la figure 11) possède des maximums locaux lors des temps de début et fin de blocs homogènes de forte similarité dans la matrice.

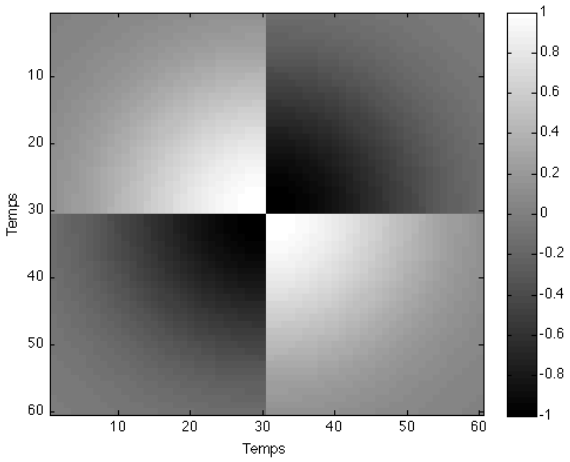


FIGURE 10 – Allure du noyau de Foote

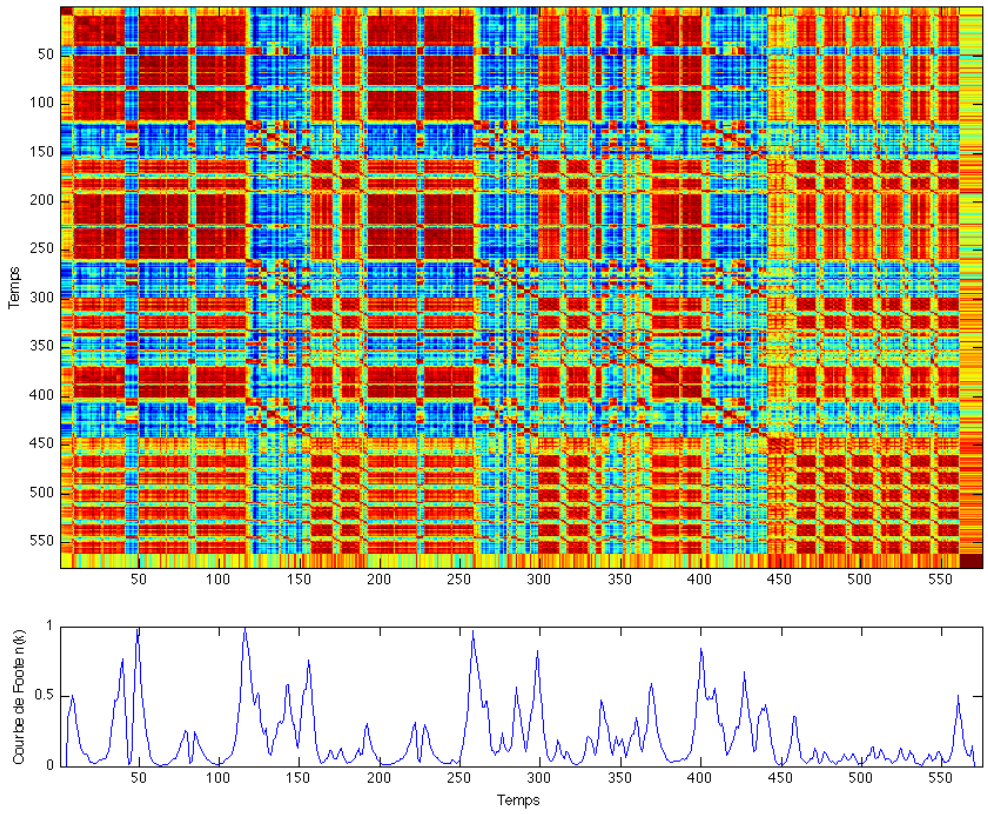


FIGURE 11 – Calcul de la courbe de nouveauté de Foote sur le morceau 33 de la base RWC POP - **Haut** Matrice d'auto-similarité par distance cosinusoïdale - **Bas** Courbe de nouveauté de Foote

4.4.2 La courbe de détection d'états

[9] propose de détecter la présence d'états dans la matrice d'auto-similarité en calculant une courbe s indiquant pour chaque trame d'analyse la probabilité d'être sur une partie homogène. Le calcul de s se base sur une détection de blocs dans la matrice d'auto-similarité. Nous extrayons pour la trame d'analyse k une sous matrice E_τ de taille L issue de la diagonale principale.

$$E_t(i,j) = S(t+i-1, t+j-1) \text{ avec } (i,j) \in [t:t+L] \times [t:t+L] \tag{23}$$

L'objectif est d'estimer l'homogénéité des matrices E_t pour chaque temps d'analyse t. Nous regardons pour cela si l'information est majoritairement présente sur la diagonale ou si elle est répartie dans toute la sous-matrice. Il s'agit donc calculer la moyenne des valeurs de la matrice sur la moyenne des valeurs de la diagonale. Nous posons $\mu(E_t)$ la moyenne de des valeurs de E_t, $\mu(diag(E_t))$ la

moyenne des valeurs de la diagonale et $\sigma(E_t)$ la variance des valeurs de E_t

$$s(t) = \frac{\mu(E_t) - \sigma(E_t)}{\mu(diag(E_t))} \tag{24}$$

Les valeurs de s si situent entre 0 et 1, nous fixons un seuil de 0.9 pour le détection d'un état, si $s(t) > 0.9$ alors nous pouvons supposer que la trame t est dans une partie homogène du signal. Nous pouvons grâce à cette méthode savoir s'il est préférable d'utiliser l'information sur la segmentation de Foote (par état) pour améliorer notre segmentation. Nous avons Figure 12 un exemple de courbe de détection d'états dans une matrice d'auto-similarité. Nous remarquons que les blocs homogènes de la matrice créent des valeurs de la courbe d'état qui dépassent le seuil de détection de 0.9.

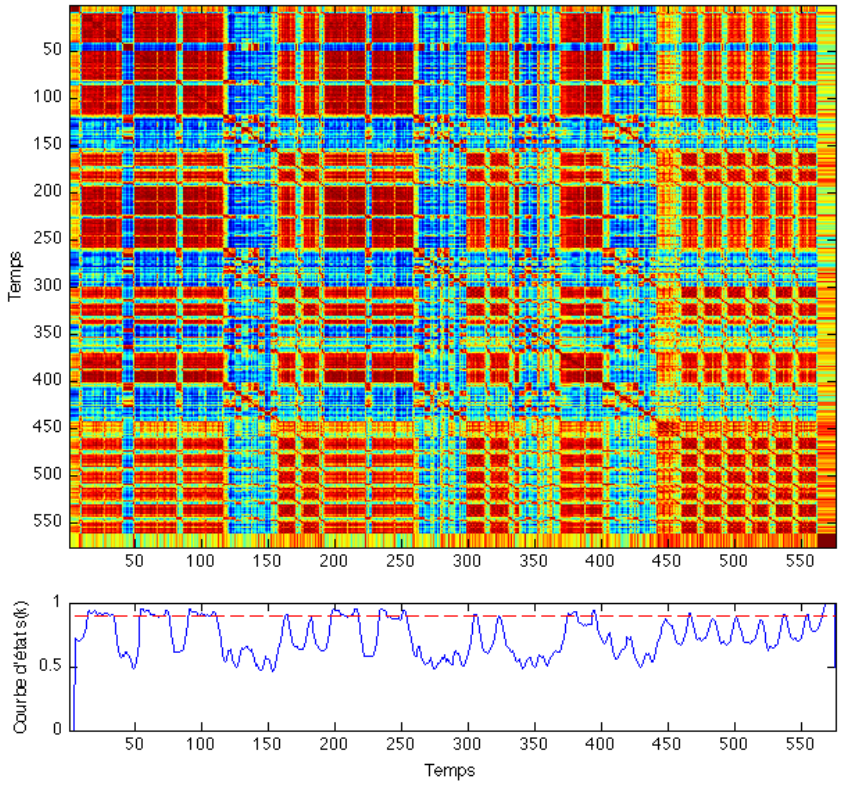


FIGURE 12 – Calcul de la courbe de détection d'état sur le morceau 33 de RWC POP - **Haut** Matrice d'auto-similarité par distance cosinusoïdale - **Bas** Courbe de détection d'état, la ligne horizontale correspond au seuil de détection d'état fixé à 0.9

4.4.3 Amélioration de la segmentation par détection d'états

L'idée est de se servir de s (la courbe de détection d'états de Peeters) pour sélectionner la courbe de nouveauté d'une des deux approches la plus appropriée. Pour l'approche par séquences nous reprenons la courbe de nouveauté qui a donné les meilleurs résultats en termes de segmentation. Nous choisissons donc c_2 (pondération globale par la courbe de Goto) que nous calculons à partir de la matrice temps-retard circulaire de l'article de Serra et al. Pour l'approche par état nous reprenons n la courbe de nouveauté de Foote. Nous calculons maintenant une nouvelle courbe de nouveauté c_4 comme la somme pondérée par s des nouveautés pour les deux approches.

La pondération effectuée sur la somme des deux courbes dépend des valeurs de s, quand les valeurs de s sont élevés nous donnons plus d'importance à la courbe de nouveauté de Foote $n(t)$ et nous annulons presque sa contribution lorsque les valeurs de s sont basses. Nous posons β la valeur contrôlant la pondération, les meilleurs résultats obtenus lors de l'évaluation ont été trouvés avec une valeur de 0.5.

$$c_4(t) = (s(t) - \beta)n(t) + (1 - s(t) + \beta)c_3(t) \tag{25}$$

Les valeurs de s descendent rarement en dessous de 0.5, ainsi nous annulons quasiment l'impact de la courbe de nouveauté n lorsque nous sommes sur des parties non-homogènes. Nous préférons donner plus d'importance à l'approche par séquence car c'est celle qui possède les meilleurs résultats. L'approche par état vient compléter l'information manquante sur des parties homogènes non répétées pouvant tout de même indiquer un changement de partie dans le morceau. Nous appliquons ensuite la même stratégie de sélection de pics que pour c_1,c_2 et c_3 pour extraire les frontières entre les régions du morceau.

4.4.4 Evaluation

Nous évaluons les performances de l'ajout de la détection d'état à nos méthodes de segmentation sur les mêmes bases de données que les précédentes évaluations. Nous reprenons la meilleure méthode de segmentation, la courbe c_2 avec la matrice de Serra et al. comme comparaison pour la nouvelle courbe c_4. Nous pouvons remarquer que l'ajout de la nouveauté par état améliore les résultats en termes de f-mesure pour la fenêtre de tolérance de 3 secondes.

BEATLES : c_4 :80.9% c_2 :80,5%
RWC-POP-A : c_4 :80,1% c_2 :78,4%
RWC-POP-B : c_4 :83.9 % c_2 :82,3%

L'ajout d'information issue de la courbe de nouveauté de Foote se fait uniquement sur la présence d'états, ainsi sur l'extraction de pics de la courbe de nouveauté c_4, nous ajoutons quasiment uniquement de nouvelles frontières par rapport à c_2. Il est rare de voir des frontières présentes après la détection de pics issues c_2 et ne pas l'être dans celle issue de c_4. L'intérêt de cette nouvelle méthode est qu'elle permet de garder les frontières issues de la segmentation par approche par séquence en y ajoutant la celles issue de l'approche par état lors de l'apparition de fortes parties homogènes. Nous pouvons également remarquer que l'écart avec les meilleurs résultats de l'état de l'art se creuse encore plus nous avons donc valider pour ces trois bases de tests une bonne méthode de segmentation prenant en compte l'information issue des deux approches.

TABLE 5 – Résultats de l'évaluation de la segmentation avec l'ajout de la détection d'état.

RWC-Pop-A						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.791	0.817	0.783			
MIREX-2012 (SMGA1)	0.710	0.741	0.701	0.236	0.247	0.232
MIREX-2013 (FK2)	0.657	0.816	0.560	0.301	0.375	0.256
$c_2(t)$ (avec pondération globale)	0.784	0.843	0.750	0.289	0.316	0.275
$c_4(t)$ (avec détection d'état)	**0.801**	0.854	0.77	-	-	-
RWC-Pop-B						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.8	0.81	0.805			
MIREX-2012 (SMGA1)	0.765	0.815	0.732	0.267	0.286	0.255
MIREX-2013 (RBH1)	0.672	0.700	0.664	0.374	0.392	0.368
$c_2(t)$ (avec pondération globale)	0.823	0.846	0.820	0.389	0.408	0.381
$c_4(t)$ (avec détection d'états)	**0.839**	0.86	0.832	-	-	-
Beatles-B						
Méthode	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [24]	0.774	0.76	0.807			
$c_2(t)$ (avec pondération globale)	0.805	0.813	0.817	0.439	0.430	0.450
$c_4(t)$ (avec détection d'états)	**0.809**	0.849	0.791	-	-	-

5 Regroupement

Notre estimation de la structure se basant sur une approche par séquences il nous faut trouver un moyen d'identifier les répétitions existantes dans le morceau. Grâce à une bonne méthode de segmentation présentée dans la partie précédente nous possédons une bonne estimation des frontières temporelles entre les différentes séquences. Il s'agit maintenant de regrouper les ensembles de séquences répétées au cours du temps en classes pour terminer l'estimation de la structure. Pour réaliser cette étape appelée de regroupement (ou labelling), nous nous basons sur un nouvel algorithme de déformation temporelle dynamique (DTW) proposé par [3]. Cette nouvelle méthode est utilisée pour trouver la partie la plus représentative du morceau en trouvant la partie la plus répétée, ceci dans le but de générer automatiquement des résumés audio. Nous souhaitons reprendre ce travail en ajoutant certaines contraintes au DTW pour estimer l'ensemble de la structure d'un morceau de musique. Après avoir détaillé le fonctionnement du DTW contraint nous présentons les deux méthodes de regroupements ainsi que les résultats obtenus par chacune d'elles sur les bases de tests. Nous introduisons également l'ajout de nouvelles contraintes sur le DTW et l'influence qu'elles ont sur les résultats du regroupement.

5.1 Détection des séquences répétées par DTW contraint

L'algorithme DTW à déjà été utilisé en estimation de la structure pour comparer la ressemblance entre deux segments d'un morceau en trouvant un chemin qui les aligne de façon optimale. Cette ressemblance peut être utilisée comme distance dans un algorithme de regroupement dans une approche par état [11]. Si les deux segments s'alignent bien et que la distance est faible alors ils sont considérés comme une répétition l'un de l'autre. Ce que propose de faire Muller dans sa variante du DTW est de pouvoir détecter d'un seul coup toutes les répétitions d'une partie du morceau en changeant simplement quelques conditions sur le DTW classique. Par l'introduction d'une mesure de "fitness" calculée à la sortie du DTW nous pouvons également quantifier la ressemblance entre les répétitions. Nous pouvons ainsi espérer appliquer cette méthode à chaque segment délimité par la segmentation pour identifier le nombre, la ressemblance et la position des répétitions dans le morceau.

L'objectif de l'algorithme de DTW est d'aligner de façon optimale deux séquence d'observations $\underline{X} = (x_1, ..., x_N)$ et $\underline{Y} = (y_1, ..., y_M)$ (correspondant ici aux vecteurs de chromas) en trouvant un chemin de coût minimum entre les points des observations. Dans notre cas la série d'observations $\underline{Y}$ est une partie du morceau et $\underline{X}$ correspond au morceau entier. Nous cherchons à trouver toutes les sous parties de $\underline{X}$ (du morceau) qui peuvent s'aligner avec $\underline{Y}$ (une sous-partie). Les sous-parties de $\underline{X}$ qui s'alignent avec $\underline{Y}$ correspondent à des répétitions de $\underline{Y}$ dans le morceau. Contrairement au DTW classique nous pouvons avoir des parties de $\underline{X}$ qui ne sont pas alignées avec $\underline{Y}$ et $\underline{Y}$ peut être alignée avec plusieurs sous parties de $\underline{X}$.

Nous partons d'une partie $\underline{Y} = (y_0, ..., y_{M-1}) = (x_t, ..., x_{t+M-1})$ de longueur M prise à partir de la trame t que nous souhaitons aligner avec tout le signal $\underline{X} = (x_1, ..., x_N)$ avec N la longueur de la matrice d'auto-similarité S. Nous nous servons des points de la matrice d'auto-similarité comme distance entre deux observations $d(y_i, x_j) = S(t + i, j)$. Nous définissons une sous matrice $\underline{\underline{S_\alpha}}$ de taille $N * M$ dont les colonnes sont celles de $\underline{\underline{S}}$ de t à $t + M - 1$. Nous définissons une matrice de

distance cumulée $\underline{\underline{D}}$ de taille $N * (M + 1)$, $\underline{\underline{D}}$ possède une colonne en plus de $\underline{\underline{S_\alpha}}$ (cette colonne d'indice 0 permet à l'algorithme de sauter certaines sous parties de $\underline{X}$ et de pouvoir reprendre plusieurs alignements entre $\underline{X}$ et $\underline{Y}$). Nous remplissons la matrice des distances cumulées comme pour le DTW classique sauf que nous cherchons à maximiser la distance entre les points et que la colonne additionnelle introduit des conditions supplémentaires. Nous commençons par construire la matrice des distances cumulées comme suit

Pour $n \in [2 : N]$ et $m \in [2 : M]$ nous avons

$$\underline{\underline{D}}(n, m) = \underline{\underline{S_\alpha}}(n, m) + \max\{\underline{\underline{D}}(n-1, m-1), \underline{\underline{D}}(n-1, m-2), \underline{\underline{D}}(n-2, m-1)\} \tag{26}$$

Cette condition favorise la progression diagonale du DTW, comme la matrice $\underline{\underline{S_\alpha}}$ possède des valeurs négative la progression est également encouragée sur les diagonales de la matrice d'auto-similarité qui correspondent effectivement aux répétitions de $\underline{Y}$.

Ensuite pour la colonne d'indice 0, nous avons pour $n \in [2 : N]$ avec $\underline{\underline{D}}(1, 0) = 0$

$$\underline{\underline{D}}(n, 0) = \max\{\underline{\underline{D}}(n-1, 0), \underline{\underline{D}}(n-1, M)\} \tag{27}$$

Cette condition permet au DTW de repartir du début de $\underline{Y}$ une fois qu'il a fini d'aligner une sous partie avec un score d'arrivée supérieur au point de départ (d'où la nécessité de la pénalisation). Ainsi nous revenons sur la colonne 0 lorsqu'une diagonale est détectée et nous pouvons progresser dans la colonne 0 sans pénalisation tant qu'une nouvelle diagonale ne commence pas.

Pour finir de remplir la matrice de distance cumulée il faut pouvoir passer de la première à la deuxième colonne pour commencer à progresser sur une nouvelle diagonale. Nous avons pour m=1

$$\underline{\underline{D}}(n, 1) = \underline{\underline{D}}(n, 0) + \underline{\underline{S_\alpha}}(n, 1) \text{ et } \underline{\underline{D}}(1, m) = -\infty \text{ pour } m \in [2 : M] \tag{28}$$

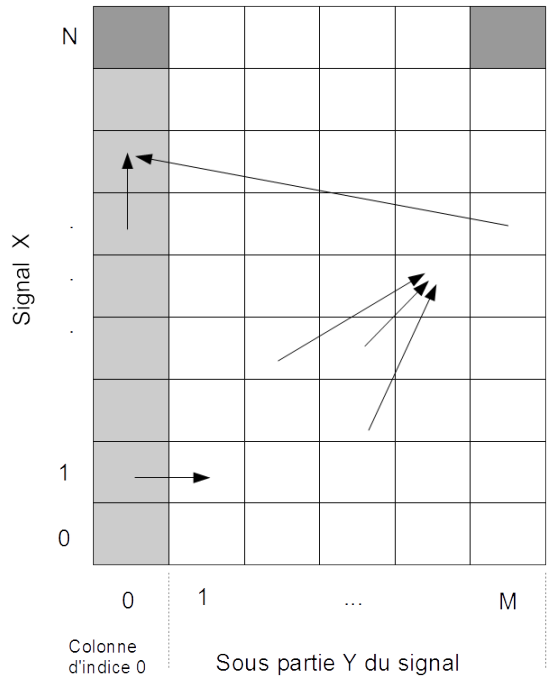


FIGURE 13 – Les prédécesseurs possibles pour la construction de la matrice des distances cumulées

Nous venons de terminer de détailler les étapes de la phase aller du DTW contraint. Pour terminer il faut récupérer la position de toute les répétitions ce qui se fait à l'aide d'une phase retour classique (backtracking). Nous prenons comme point de départ pour la phase retour le maximum entre $\underline{\underline{D}}(N,0)$ ou $\underline{\underline{D}}(N,M)$ (si une diagonale termine sur le coin supérieur de $\underline{\underline{S_\alpha}}$). Une fois la matrice des distance cumulées calculée nous pouvons facilement obtenir toutes les répétitions de $\underline{Y}$ en projetant les diagonales détectées sur $\underline{X}$.

Nous obtenons en sortie de l'algorithme une famille de segments $A = (\alpha_1, ..., \alpha_K)$ avec pour $k \neq j$ α_k est une répétition de α_j et $\alpha_k \cap \alpha_j = 0$. Nous fixons α_1 comme étant le segment mère, c'est à dire celui qui a joué le rôle de $\underline{Y}$ dans le DTW. Il est forcément présent dans la famille A car

$\underline{Y}$ est une partie de $\underline{X}$ et s'aligne parfaitement avec lui-même. Nous posons $|\alpha_j|$ comme la longueur du segment α_i.

Nous obtenons également une famille de chemins $P = (p_1, ..., p_K)$ où un chemin est une des diagonales détectées par le DTW. Un chemin p_i de longueur l est défini par un ensemble de couples $((n_1, m_1), ..., (n_l, m_l))$ correspondant aux positions des points du chemin dans $\underline{\underline{S}}$. Les segments de A correspondent donc aux projections sur l'axe des ordonnées des chemins de $\overline{P}$. Nous présentons Figure 14 l'allure d'une matrice des distances cumulées issue de l'application du DTW à un segment du morceau 19 de la base RWC Pop.

La dernière étape est d'attribuer une score à la famille A trouvée par DTW sur le segment mère α_1. Ce score permet de traduire à la fois si la famille est composée de forte répétitions et si elle couvre un grande partie du morceau. Pour cela nous reprenons le calcul de "fitness" de Muller [36]. Nous définissons ainsi le score de fitness $\phi(A)$ de la famille A comme la moyenne harmonique entre γ et σ où

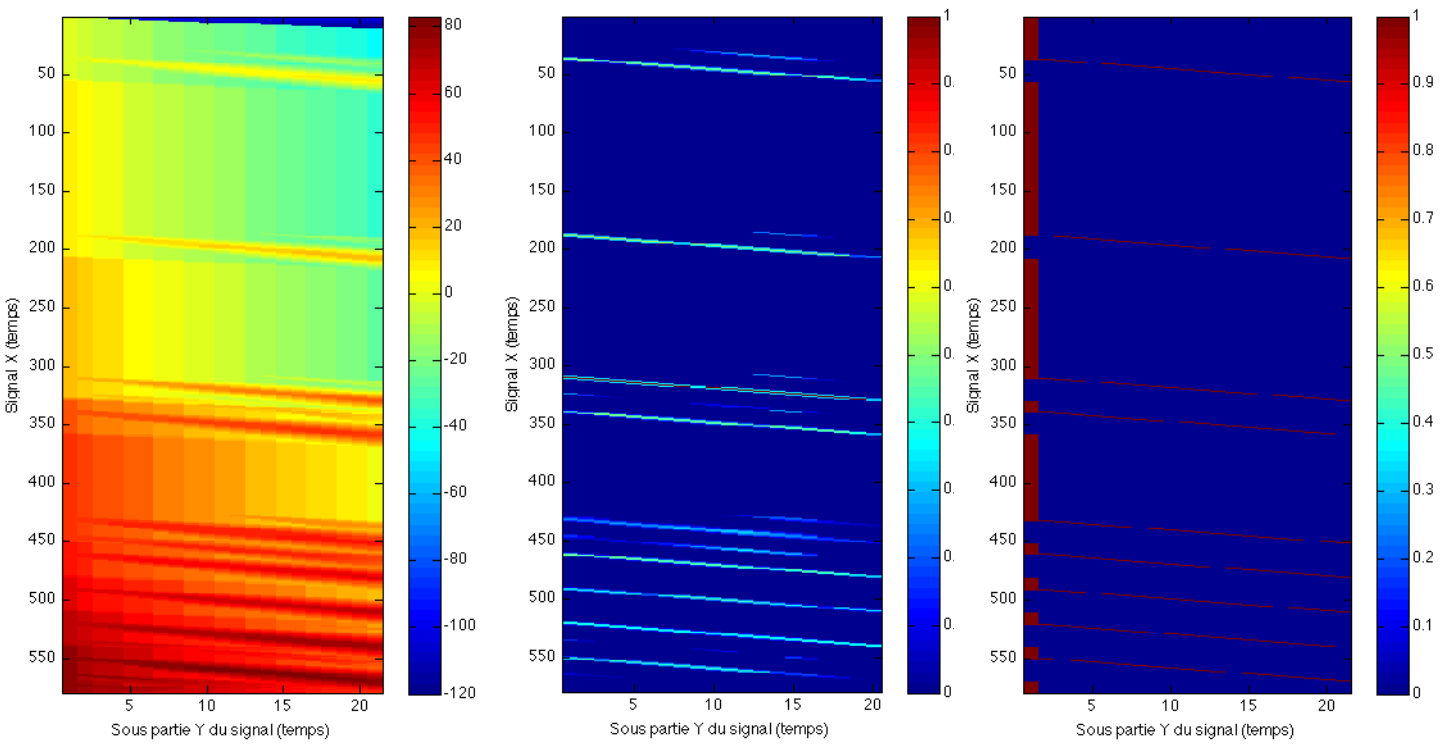
$$\sigma = \frac{\max\{\underline{\underline{D}}(N, 0), \underline{\underline{D}}(N, M)\} - |\alpha_1|}{\sum_{k=1}^{K} l_k} \text{ avec } l_k \text{ la longueur de } p_k \tag{29}$$

σ mesure la moyenne des valeurs des points de $\underline{\underline{S}}$ appartenant aux chemins de la famille P et donc la ressemblance moyenne des segments détectés autres que le segment mère.

$$\gamma = \frac{\sum_{k=2}^{K} |\alpha_k|}{N} \tag{30}$$

γ représente le pourcentage du morceau qu'explique la famille A. Enfin le score final ϕ de la famille A est calculé comme la moyenne harmonique des deux grandeurs précédentes.

$$\phi(A) = \frac{2 \cdot \sigma \cdot \gamma}{\sigma + \gamma} \tag{31}$$



(a) (b) (c)

FIGURE 14 – **(a)** Partie $\underline{\underline{S_\alpha}}$ de $\underline{\underline{S}}$, matrice de similarité entre $\underline{X}$ en ordonnées et $\underline{Y}$ en abscisses - **(b)** Matrice des distances cumulées $\underline{\underline{D}}$ - **(c)** Diagonales détectées à l'étape retour

5.2 Ajouts de contraintes à l'étape aller du DTW

Notre utilisation du DTW dans les méthodes de regroupement se fait seulement sur quelques segments dans l'objectif d'identifier les répétitions existantes parmi ceux de la famille de segments détectés issue de la segmentation. Dans le but d'améliorer les performances du regroupement il faudrait encourager la détection de répétitions à trouver celles qui correspondent au mieux aux segments délimités par la segmentation. Nous ajoutons pour cela de nouvelles contraintes à l'étape aller du DTW en se basant sur les courbes de nouveautés de la segmentation. Nous souhaitons nous servir de l'information qu'elles nous donnent sur la probabilité d'avoir un segment à un temps t pour favoriser la détection de diagonales commençant ou finissant sur des instants de pics de nouveauté. Pour mieux comprendre la mise en place de cette nouvelle contrainte nous détaillons les conditions nécessaires à la détection d'une diagonale dans le DTW utilisé.

Nous nous donnons un segment mère $\alpha_1 = [t : t + M - 1]$ de taille M dont nous allons chercher les répétitions à l'aide du DTW.
Nous posons $\underline{\underline{S_\alpha}} = \underline{\underline{S}}(N, (t+1) : (t+1+M)) \in \mathbb{R}^{NxM}$ l'extrait utile de la matrice d'auto-similarité.
Nous posons $\underline{\underline{D}} \in \mathbb{R}^{Nx(M+1)}$ la matrice des distances cumulées.
Nous nous donnons un chemin $p_c = ((n_1, m_1), ..., (n_l, m_l))$ respectant les contraintes de déplacement qui a comme projection sur l'axe des temps $\alpha_c = [n_1 : n_l]$ candidats à être une répétition du segment mère α_1

Nous avons p_c une diagonale si α_c est disjoint de toute répétition déjà détectée et si nous avons

$$\underline{\underline{D}}(n_l, M) > \underline{\underline{D}}(n_l, 0) \tag{32}$$

Or

$$\underline{\underline{D}}(n_l, M) = \sum_{i=1}^{l} \underline{\underline{S_\alpha}}(n_i, m_i) + \underline{\underline{D}}(n_1, 0) \tag{33}$$

Donc p_c est une diagonale si

$$\sum_{i=1}^{l} \underline{\underline{S_\alpha}}(n_i, m_i) > 0 \tag{34}$$

Tant que la somme des points d'un chemin est positive il peut être considéré comme une répétition. L'idée est de trouver un moyen d'encourager les diagonales à commencer et finir sur les temps de nouveautés. Cela supprimerait les diagonales incohérentes avec la segmentation et pourrait encourager certaines diagonales plus timides à être détectées si elles respectent bien les points de nouveautés.

Nous introduisons une pénalisation à l'étape aller en utilisant la courbe de nouveauté c_4 et un seuil de pénalisation θ. Lors de l'étape de sélection de pics pour les méthodes de segmentation (page 23) nous avions déjà introduit un seuil $\theta = 0.1$ à partir du quel nous pouvons considérer un maximum local comme correspondant à la présence d'une frontières entre deux segments. L'idée est de se servir du même seuil que dans la détection de pics de la segmentation pour savoir pour un temps donné si nous devons pénaliser ou encourager une progression diagonale. Cette pénalisation aura lieu en deux points, sur le temps de départ de la diagonale et le temps d'arrivée. Les valeurs

de la dernière colonne de la matrice des distances cumulées sont donc modifiées.

Nous introduisons également un coefficient α qui permet de contrôler l'influence de la contrainte en pondérant la pénalisation proportionnellement à la longueur du segment mère analysé. Cela permet d'accorder une importance équivalente à la pondération quelque soit la taille du segment mère. Nous remarquons que lorsque la courbe de nouveauté est au-dessus du seuil pour les instants de débuts et fin de diagonale la valeur de $\underline{\underline{D}}(n_l, M)$ augmente. Nous avons plus de chances de valider la condition équation (32) et donc de valider la projection de cette diagonale en tant que répétition du segment mère.

$$\underline{\underline{D}}(n_l, M) = \sum_{i=1}^{l} \underline{\underline{S_\alpha}}(n_i, m_i) + (c_4(n_1) + c_4(n_l) - 2 * \theta).(\alpha M) + \underline{\underline{D}}(n_1, 0) \tag{35}$$

5.3 Algorithmes de regroupement

Nous disposons maintenant d'un moyen d'identifier les répétitions d'un segment du morceau. Notre méthode segmentation nous permet d'estimer les frontières entre les régions du morceau, il ne reste plus qu'à trouver une méthode de regroupement efficace pour attribuer un label à chaque partie. Nous appellerons segments les parties délimitées par les frontières estimées lors de la segmentation. Le morceau est donc une succession de segments correspondant à la succession des différentes parties musicales. Ce sont ces segments que nous souhaitons regrouper en classes afin d'identifier les répétitions présentes dans le morceau. Nous présentons une première méthode de regroupement itérative qui traite les segments un par un en leur attribuant un label définitif. Nous étudierons pour cette méthode l'ajout de contraintes sur le DTW car les répétitions des segments ne correspondent pas forcément à d'autres segments candidats à un label. En effet nous pouvons avoir plusieurs labels possibles pour un même segment si ces frontières sont mal définies par la segmentation. Pour traiter ce genre de cas, d'autant plus fréquents si la segmentation est moins précise, nous disposons d'une deuxième méthode par détection groupée. Cette méthode fait une recherche exhaustive de tous les regroupements possibles, elle contraint uniquement les segments mères (le segment ayant le score de fitness le plus élevé de sa classe) à être délimités par les temps de nouveautés. Les répétitions elles peuvent se détacher des frontières en évitant au maximum le recouvrement entre différents segments.

5.3.1 Regroupement itératif

A l'issue de la segmentation nous obtenons un ensemble de n_e de temps de nouveautés correspondant aux débuts et fins de segments. Nous travaillons avec la famille F de $I = (n_e - 1)$ segments ayant pour bornes deux temps de nouveautés successifs. Nous cherchons à attribuer un label à chaque segment de F, les segments répétés auront le même. Nous ne traitons que les segments délimités par deux temps de nouveauté successifs (nous ne prenons pas $[\tau_1 : \tau_3]$ par exemple) car nous considérons que la segmentation délimite déjà correctement les différentes partie du morceau. Si $[\tau_1 : \tau_3]$ était une partie qui ne pouvait pas se séparer en deux sous parties alors τ_2 ne ferait pas partie des temps de nouveautés.

$$F = \{s_i \mid i \in [1, I]\} = \{[\tau_i : \tau_{i+1}] \mid i \in [1, I]\} \tag{36}$$

Nous effectuons les étapes suivantes jusqu'à ce que chaque segment ait un label

1. Nous appliquons le DTW pour chaque segment $s_i \in F$ afin de connaître la position de ses répétitions.
2. Nous classons les segments par ordre décroissant de leur score de fitness $\phi(s)$
3. Nous prenons le premier segment s_i qui ne possède pas encore de label
4. Si une des répétitions de s_i croise suffisamment un ou plusieurs autres segments $s_j \in F$ (si s_i couvre au moins 75% de s_j) qui n'ont pas de label alors nous leur attribuons le même label et nous retirons s_j et s_i de la liste des segments. Si aucune des répétitions de s_i ne croise suffisamment un autre segment de F nous attribuons un label unique à s_i et nous le retirons de la liste des segments.

Nous finissons donc par attribuer un label à chaque segment de famille F sans en modifier leur taille ou leur position ce qui nous permet de terminer l'estimation de la structure. Cette méthode a l'avantage de ne pas posséder de paramètres à régler et s'applique donc de la même manière quel que soit le morceau.

5.3.2 Regroupement par détection groupée

La méthode de regroupement par détection groupée se sert de la segmentation uniquement comme point de départ pour les recherches de répétitions, l'attribution des labels n'est pas contrainte à rester dans les segments de F. Nous imposons uniquement aux segments mères d'être parmi ceux de la famille F. En conséquence nous pouvons avoir des parties du morceau qui ne se voient pas encore attribuer de label, dans ce cas nous attribuons un label unique à chaque partie non marquée par l'algorithme. Ces parties non marquées correspondent aux parties non répétées ou ne pouvant être expliquées par une approche par séquence (ce que nous considérons comme du bruit lors de la construction des matrices d'auto-similarité).

L'algorithme de regroupement par détection groupée cherche à trouver une famille de P segments mères $G = \{s_p \mid p \in [1, P]\}$ avec $G \subset F$, qui expliquent au mieux le signal. Nous nous servons d'un score $\varphi(G)$ prenant en compte le score de "fitness" des segments mères (premier terme de l'équation (37)) ainsi que la quantité de signal expliquée par ces P segments et leurs répétitions (deuxième terme de l'équation (37)). Ceci afin d'encourager les familles de segments mères qui, avec leurs répétitions, recouvrent une grande partie du morceau.

$$\varphi(G) = (\sum_{p=1}^{P} \gamma(s_p)) \cdot (\sum_{p=1}^{P} \phi(s_p)) \tag{37}$$

Pour réaliser le meilleur regroupement au sens du score φ pour un ensemble de P segment, nous exécutons les étapes suivantes pour toutes les combinaisons possibles de P segments différents parmi F.

1. Nous exécutons le DTW sur chaque segment $s_p \in G$ pour avoir accès aux positions de leurs répétitions et leur score $\phi(s_p)$

2. Si un des P segments ainsi que toutes leurs répétitions se recouvrent en un seul point nous jetons cette combinaison
3. Sinon nous calculons le score $\varphi(G)$ de l'ensemble G de P segments mères
4. Nous gardons la famille des P segments G ayant le score $\varphi(G)$ le plus élevé

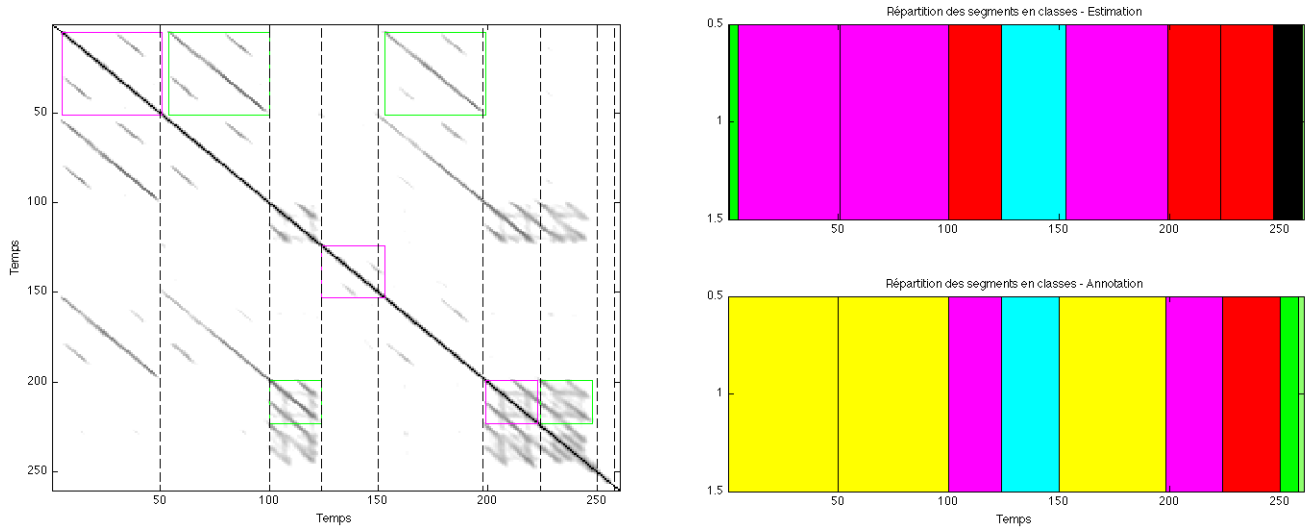
Nous obtenons finalement une segmentation où le signal est réparti en P classes, dans chaque classe tous les segments sont toutes les répétitions d'un même segment mère. C'est avec cet algorithme, en cherchant des familles de $P = 5$ segments, que nous avons les meilleurs résultats, c'est également la plus longue en temps de calcul comme la recherche est combinatoire et non itérative. Cependant il n'y a aucune raison de chercher uniquement des familles de 5 segments, il faudrait donc un moyen d'estimer la taille optimale de la famille pour chaque morceau traité.

Nous présentons Figure 15 et Figure 16 deux exemples d'application de la méthode de regroupement par détection groupée sur des morceaux de la base des Beatles. Pour les deux figures nous avons sur la gauche la matrice d'auto-similarité utilisée pour l'estimation de la structure. Nous avons ajouté sur cette matrice des lignes en pointillés noires sur les temps des frontières des segments de l'annotation. Les carrés magenta sur la diagonale principale correspondent à la position des segments mères à partir desquels nous avons appliqué le DTW, c'est-à-dire les segments possédant le score de fitness le plus élevé. Les carrés verts correspondent à la position des répétitions estimées par le DTW, ils correspondent aux répétitions du segment représenté par le carré magenta avec lequel ils sont alignés horizontalement.

Nous avons sur la droite une illustration de la structure où chaque couleur correspond à une classe et où chaque rectangle correspond à un segment de la structure. Nous avons sur la partie haute la représentation du résultat de l'estimation de la structure après le regroupement par détection groupée et sur la partie basse la structure issue de l'annotation.

La figure 15 présente un cas où la méthode de regroupement donne une bonne estimation de la structure du morceau par rapport à l'annotation. Nous pouvons remarquer sur la figure de droite que nous avons attribué un mauvais label à un seul des segments. Notre estimation considère que ce segment est une répétition du segment voisin car il est représenté par une diagonale détectée par le DTW dans le couloir temporel d'un des segments mères. Cet exemple montre que les descripteurs que nous utilisons peuvent nous amener à considérer deux segments comme étant similaires alors que l'annotation indique le contraire.

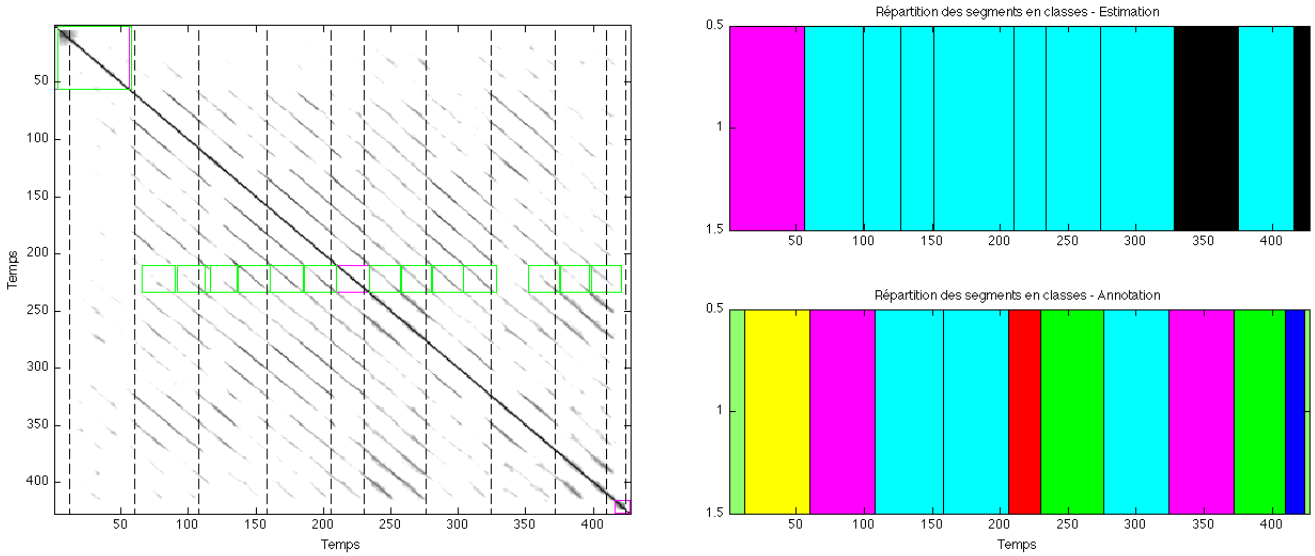
La figure 16 présente un cas où la méthode de regroupement donne un résultat éloigné de l'annotation. Dans ce cas la matrice d'auto-similarité possède un excès de diagonales par rapport au nombre de répétitions présentes dans l'annotation. Ainsi l'estimation à attribuer à tort le même label à la plupart des segments. Nous pouvons voir également que la segmentation de l'estimation est différente de celle de l'annotation ce qui peut réduire les performances de nos méthodes de regroupement.



(a) Matrice de d'auto-similarité

(b) Représentation de la structure

FIGURE 15 – Résultat de l'application du regroupement par détection groupée au morceau 17 de la base des Beatles - Voir partie 5.2.2 pour les explications



(a) Matrice d'auto-similarité

(b) Représentation de la structure

FIGURE 16 – Résultat de l'application du regroupement par détection groupée au morceau 134 de la base des Beatles - Voir partie 5.2.2 pour les explications

5.4 Évaluation des deux méthodes

5.4.1 Procédure d'évaluation du regroupement

L'évaluation du regroupement est l'une des tâches de la partie estimation de la structure de la campagne d'évaluation annuelle MIREX. Nous présentons deux scores correspondant à deux manières différentes d'évaluer l'estimation de la structure, la f-mesure par paire et les scores de sur et sous-segmentation. Ces scores, en particulier pour la f-mesure par paire, sont ceux utilisés en majorité dans l'état de l'art et dans MIREX. L'évaluation est faite sur les mêmes bases de données que la segmentation en prenant en compte cette fois le label associé à chaque segment de l'annotation.

Le premier score utilisé dans les évaluations est la f-mesure par paire introduite par [18]. Il faut dans un premier temps décomposer l'estimation et l'annotation en trames de 100ms possédant chacune le label correspondant à la partie de la structure dont elle a été extraite. NOus regroupons ensuite toutes les paires de trames, pour l'annotation et l'estimation, ayant le même label pour former les deux ensembles P_e (pour l'estimation) et P_a (pour l'annotation). Nous nous en servons pour calculer une précision P et un rappel R et enfin une f-mesure F comme la moyenne harmonique des deux grandeurs.

$$P = \frac{|P_e \cap P_a|}{|P_e|} \tag{38}$$

$$R = \frac{|P_e \cap P_a|}{|P_a|} \tag{39}$$

$$F = \frac{2.P.R}{(P+R)} \tag{40}$$

Le deuxième score est celui de sur-segmentation S_o et de sous-segmentation S_u introduit par Lukashevich [37]. Ces scores permettent une meilleure analyse du comportement et des performances des systèmes d'estimation de la structure musicale. Contrairement à la f-mesure par paire, les scores de sur-segmentation et sous-segmentation ne sont pas sensibles au nombre et à la distribution des classes ce qui permet de comparer les performances des méthodes de regroupement sur des morceaux ayant des structures complètement différentes. De plus ils peuvent nous indiquer sur quel niveau de structure nous travaillons, par exemple si le score S_o est élevé cela nous indique que notre méthode aura tendance à estimer une structure avec des segments plus longs. Cela peut arriver en regroupant deux segments successifs de l'annotation en un seul dans l'estimation, ce qu'on appelle la sous-segmentation. Ces scores sont particulièrement utiles pour l'évaluation des algorithmes de regroupement hiérarchiques pouvant, selon les paramètres, estimer différents niveau de structure. Ils donnent une bonne indication du comportement de nos méthodes même si le niveau de la structure est déjà fortement influencé par la segmentation.

L'idée du calcul de S_o et S_u reprend celle d'utiliser des entropies conditionnelles $H(A/E)$ et $H(E/A)$ pour évaluer la structure [38]. Ces deux grandeurs atteignent zéro quand la segmentation est idéale mais leur maximum dépend du nombre et de la répartition des états. Il s'agit donc pour calculer S_o et S_u de trouver la valeur maximale que peut atteindre $H(A/E)$ et $H(E/A)$.

Nous définissons

— N - Le nombre total de trame de 100 ms, il est identique pour l'annotation et l'estimation
— N_a - Le nombre d'états de l'annotation
— N_e - Le nombre d'états de l'estimation
— $n_{i,j}$ - Le nombre de trames qui appartiennent à l'état i pour l'annotation et l'état j pour l'estimation
— n_i^a - Le nombre de trames appartenant à l'état i dans l'annotation
— n_j^e- Le nombre de trames appartenant à l'état j dans l'estimation

Nous posons $p_{i,j}$ la distribution jointe des labels ainsi que p_i^a et p_j^e les distributions marginales pour l'annotation et l'estimation

$$p_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{i,j}} \tag{41}$$

$$p_i^a = \frac{n_i^a}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{i,j}} \tag{42}$$

$$p_j^e = \frac{n_j^e}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{i,j}} \tag{43}$$

On a également les distributions conditionnelles

$$p_{i,j}^{a|e} = \frac{n_{i,j}}{n_j^e} \text{ et } p_{j,i}^{e|a} = \frac{n_{i,j}}{n_i^a} \tag{44}$$

On a alors les entropies conditionnelles

$$H(E|A) = -\sum_{i=1}^{N_a} p_i^a \sum_{j=1}^{N_e} p_{j,i}^{e|a} \log_2 p_{j,i}^{e|a} \tag{45}$$

$$H(A|E) = -\sum_{j=1}^{N_e} p_j^e \sum_{i=1}^{N_a} p_{i,j}^{a|e} \log_2 p_{i,j}^{a|e} \tag{46}$$

Pour normaliser les scores il faut calculer la valeur maximale des entropies conditionnelles, elle est atteinte quand les états de l'estimation sont distribués uniformément par rapport à ceux de l'annotation. Dans ce cas les distributions conditionnelles s'écrivent $p_{i,j}^{a|e} = \log_2 N_e$, ce qui nous permet d'exprimer le maximum des entropies.

$$H(E|A)_{max} = \log_2 N_e \text{et} H(A|E)_{max} = \log_2 N_a \tag{47}$$

On définit finalement S_o et S_u

$$S_o = 1 - \frac{H(E|A)}{\log_2 N_e} \text{ et } S_o = 1 - \frac{H(A|E)}{\log_2 N_a} \tag{48}$$

Nous reprenons donc le calcul des entropies conditionnelles en les normalisant pour obtenir des scores aux valeurs restant toujours dans [0 1]. Ils atteignent 1 lorsque l'estimation est identique à l'annotation et vont vers 0 lorsque les labels sont distribués aléatoirement.

5.4.2 Évaluation sans l'ajout de contraintes

Nous rassemblons dans le tableau suivant les résultats de l'évaluation du regroupement pour les deux méthodes sur les 3 bases de tests présentées précédemment. "Méthode 1" correspond au regroupement itératif et la ligne "Méthode 2" correspond au regroupement par détection groupée. Nous ne disposons pas de comparaison pour la base RWC Pop B car lors des campagnes d'évaluation MIREX les données n'étaient pas annotées en label, de plus les annotations datant de 2012 nous ne disposons pas encore de comparaison dans l'état de l'art. Pour RWC Pop A les comparaisons sont issues de MIREX et pour la base des Beatles les comparaisons à d'autres travaux de l'état de l'art sont celles rapportées par Serra et al. [24].

Table 6 – Résultats de l'évaluation du regroupement

RWC pop A					
Method	F	P	R	S0	Su
Méthode 1	0.615	0.751	0.543	0.637	0.752
Méthode 2	**0.630**	0.622	0.645	0.680	0.635
SMAG	0.688	0.746	0.666	0.717	0.719
FK2	0.635	0.706	0.612	0.674	0.745
RWC Pop B					
Méthode	F	P	R	S0	Su
Méthode 1	0.702	0.692	0.747	0.798	0.768
Méthode 2	**0.714**	0.713	0.743	0.809	0.783
Beatles					
Method	F	P	R	S0	Su
Méthode 1	0.663	0.722	0.673	0.718	0.745
Méthode 2	**0.690**	0.689	0.774	0.744	0.708
Chen [39] & Li	0.630	0.610	0.690	-	-
Mauch et al. [40]	0.660	0.770	0.610	-	-
Serra et al [24].	0.711	0.681	0.787	-	-

Les résultats des deux méthodes se situent en dessous des meilleurs scores de l'état de l'art mais sont néanmoins meilleurs que les méthodes moins récentes que celle de Serra et al [24]. Il semblerait que nos algorithmes de segmentation et de regroupement favorisent l'annotation B pour RWC Pop. En effet pour RWC Pop A nous remarquons une différence importante entre les scores S_o et S_u qui montre une tendance à la sur-segmentation par rapport à l'annotation. En revanche pour RWC Pop B les valeurs S_o et S_u sont plus élevées et plus rapprochées ce qui confirme un meilleur équilibre entre la sur-segmentation et la sous-segmentation.

L'évaluation sur les trois bases de données montre une légère supériorité de la méthode par détection groupée par rapport au regroupement itératif. Nous rappelons qu'il s'agit de la méthode qui accorde le plus d'importance aux résultats du DTW en lui laissant l'opportunité de modifier les frontières entre les régions du signal pour optimiser le regroupement. Nous espérons améliorer

les scores de la première méthode en ajoutant des contraintes sur l'étape aller du DTW. Il faudra également trouver un moyen d'estimer la taille optimale de la famille à chercher pour la méthode de regroupement par détection groupée.

5.4.3 Évaluation avec l'ajout de contraintes

Nous analysons l'influence de l'ajout de la contrainte à l'étape aller du DTW sur la première méthode de regroupement. Nous ne présentons pas l'impact de l'ajout de la contrainte sur la deuxième méthode car l'évaluation montre qu'elle dégrade fortement les résultats pour les trois bases de tests. Pour la méthode de regroupement groupé l'attribution des labels ne se fait pas forcément dans les bornes établies par la segmentation. Ainsi en ajoutant la contrainte nous perdons la possibilité d'adapter la segmentation pour optimiser le regroupement ce qui peut expliquer qu'elle s'applique mal à cette méthode.

Pour discuter de l'efficacité de l'ajout de la contrainte nous réalisons les mêmes évaluations que pour les deux méthodes de regroupement. Nous comparons cette fois uniquement la méthode de regroupement itérative avec et sans la nouvelle contrainte sur l'étape aller du DTW.

Table 7 – Résultats de l'évaluation du regroupement

RWC pop A					
Méthode	F	P	R	S0	Su
Méthode 1 sans contraintes	0.615	0.751	0.543	0.637	0.752
Méthode 1 avec contraintes	**0.619**	0.725	0.560	0.647	0.729
RWC Pop B					
Méthode	F	P	R	S0	Su
Méthode 1 sans contraintes	0.702	0.692	0.747	0.798	0.768
Méthode 1 avec contraintes	0.**705**	0.715	0.728	0.788	0.784
Beatles					
Méthode	F	P	R	S0	Su
Méthode 1 sans contraintes	**0.663**	0.722	0.673	0.718	0.745
Méthode 1 avec contraintes	0.659	0.703	0.758	0.729	0.724

En comparant les score en terme de f-mesure par paire nous remarquons que les variations sont très faibles, ils sont légèrement meilleurs pour les annotations RWC et un peu moins bons pour les Beatles. Nous pouvons cependant deviner l'impact de l'ajout de la contrainte en regardant les scores de sur et sous-segmentation. En effet dans chaque cas les valeurs de S_o et S_u se rapprochent lors de l'ajout de la nouvelle contrainte, nous perdons donc les tendances à sur ou sous-segmenter pour s'approcher d'une estimation de la structure plus stable par rapport aux annotations. Cet équilibre s'explique par la prise en compte de la courbe de nouveauté, qui a permis de valider une bonne méthode de segmentation ce qui permet donc également d'encourager le regroupement à estimer une structure adaptée aux annotations.

6 Conclusion

Durant ce stage nous avons étudié les deux principales étapes d'un système d'estimation de la structure musicale qui sont la segmentation et le regroupement. Les méthodes développées se basent principalement sur une approche par séquence. L'utilisation de cette approche nous a amenés à utiliser des descripteurs sensibles à l'aspect harmonique du signal comme les chromas CENS tout en restant relativement indépendant du timbre. Nous avons introduit l'utilisation de matrices d'auto-similarité ainsi qu'une succession de traitements visant à mettre en valeur les diagonales correspondant aux séquences dans un morceau de musique.

L'étude de la segmentation pour une approche par séquences nous a dirigés vers les travaux de Serra et al.[24], ils possèdent les meilleurs résultats de l'état de l'art pour cette tâche. Nous avons amélioré la méthode de Serra et al. en utilisant des probabilités a priori sur les retards en se basant sur la courbe de retard de Goto. En se servant de la courbe de retard comme pondération globale ou locale nous avons pu mettre au point une nouvelle méthode de segmentation performante. L'évaluation de cette nouvelle méthode sur des bases de tests connues de l'état de l'art a donné de très bons résultats. En particulier en utilisant une matrice temps-retard circulaire basée sur la construction de Serra et al. nous sommes passés au-dessus des meilleurs résultats publiés jusqu'à maintenant pour les trois bases de tests. Nous avons encore amélioré notre méthode de segmentation en combinant notre courbe de nouveauté basée sur une approche par séquence à celle de de Foote grâce à une détection d'état dans le morceau.

Nous avons proposé des nouvelles méthodes de regroupement en nous basant sur un algorithme de déformation temporelle dynamique proposé par Muller [3]. Cet algorithme permet d'identifier la position de toutes les répétitions d'un segment du morceau. En reprenant les résultats de la segmentation nous avons proposé deux méthodes pour regrouper les segments en classes. Une première méthode de regroupement itérative attribue un label à chaque segment et ses répétitions en les prenant par ordre décroissant de score de "fitness" [36]. Une deuxième méthode, plus performante, cherche l'ensemble de n segments qui, avec leurs répétitions, expliquent au mieux la structure selon certains critères. Nous avons également modifié le comportement de la première méthode en ajoutant des nouvelles contraintes prenant en compte les courbes de nouveauté de la segmentation. Ces nouvelles contraintes ont été appliquées à l'algorithme DTW pour encourager la détection de répétitions commençant ou finissant sur les frontières issues de la segmentation. L'évaluation de nos méthodes de regroupement sur les trois mêmes bases de tests a donné de bons résultats. Contrairement à la segmentation nous ne dépassons pas les meilleures méthodes de l'état de l'art mais nous avons validé deux bonnes méthodes de regroupement utilisant uniquement une approche par séquence.

L'utilisation de techniques adaptées à l'approche par état nous a permis d'améliorer nos méthodes de segmentation. La méthode de détection d'état introduite dans la partie 4.4.3 peut nous permettre de séparer le morceau en deux, une partie à traiter avec une approche par séquence et une partie à traiter avec une approche par état. Kaiser [10] a introduit deux nouveaux noyaux de convolution inspiré du noyau de Foote [8] (partie 4.4.1) permettant de détecter le passage entre une partie homogène et une partie inhomogène dans la matrice d'auto-similarité. Ces noyaux peuvent être ajoutés à nos méthodes de segmentation pour plus de précision lorsqu'on se sert des deux approches simultanément. Nous pourrions, comme pour la segmentation, nous servir de méthodes

de regroupement se basant sur les blocs dans la matrice d'auto-similarité et donc sur une approche par état pour compléter nos méthodes de regroupement. Grâce à la détection d'états nous pouvons savoir quelle partie il faut traiter à l'aide de méthodes de regroupement adaptées aux états ou aux séquences. Nous pourrions également modifier nos algorithmes de regroupement afin de trouver de meilleurs critères pour traiter l'ensemble des répétitions identifiées à l'issue du DTW. Il faudrait par exemple estimer la taille optimale de la famille de segments mères pour le regroupement par détection groupée tout en définissant un meilleur critère pour l'identification du regroupement optimal.

Références

[1] Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report : Audio-based music structure analysis. 2010.

[2] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5) :1783–1794, 2006.

[3] Meinard Muller, Nanzhu Jiang, and Peter Grosche. A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(3) :531–543, 2013.

[4] Geoffroy Peeters. Deriving musical structures from signal analysis for music audio summary generation :"sequence" and "state" approach. In *Computer Music Modeling and Retrieval*, pages 143–166. Springer, 2004.

[5] Tristan Jehan. *Creating music by listening*. PhD thesis, Massachusetts Institute of Technology, 2005.

[6] Emilia Gómez, Perfecto Herrera, and Beesuan Ong. Automatic tonal analysis from music summaries for version identification. In *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[7] Matija Marolt. A mid-level melody-based representation for calculating audio similarity. In *ISMIR*, pages 280–285, 2006.

[8] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 1, pages 452–455. IEEE, 2000.

[9] Geoffroy Peeters. Music structure discovery : measuring the" state-ness" of times. In *Late-Breaking News from the 12th International Symposium for Music Information Retrieval (ISMIR)*. Citeseer, 2011.

[10] Florian Kaiser and Geoffroy Peeters. Multiple hypotheses at multiple scales for audio novelty computation within music. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 231–235. IEEE, 2013.

[11] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6) :1159–1170, 2009.

[12] Matthew L Cooper and Jonathan Foote. Automatic music summarization via similarity analysis. In *ISMIR*, 2002.

[13] Florian Kaiser and Thomas Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *ISMIR*, pages 429–434, 2010.

[14] Kristoffer Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing*, 2007(1) :159–159, 2007.

[15] Michael M Goodwin and Jean Laroche. A dynamic programming approach to audio segmentation and speech/music discrimination. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 4, pages iv–309. IEEE, 2004.

[16] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *ISMIR*, volume 2, pages 94–100, 2002.

[17] Jean-Julien Aucouturier and Mark Sandler. Segmentation of musical signals using hidden markov models. *Preprints-Audio Engineering Society*, 2001.

[18] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2) :318–326, 2008.

[19] Meinard Müller, Frank Kurth, and Michael Clausen. Audio matching via chroma-based statistical features. In *ISMIR*, volume 2005, page 6th, 2005.

[20] Meinard Muller and Sebastian Ewert. Towards timbre-invariant audio features for harmony-based music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3) :649–662, 2010.

[21] Emilia Gómez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3) :294–304, 2006.

[22] Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *ISMIR*, pages 35–40, 2007.

[23] Lie Lu, Muyuan Wang, and Hong-Jiang Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 275–282. ACM, 2004.

[24] Joan Serra, Meinard Müller, Peter Grosche, and Josep Lluis Arcos. Unsupervised detection of music boundaries by time series structure features. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[25] Yu Shiu, Hong Jeong, and C-CJ Kuo. Similar segment detection for music structure analysis via viterbi algorithm. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 789–792. IEEE, 2006.

[26] Florian Kaiser and Geoffroy Peeters. A simple fusion method of state and sequence segmentation for music structure discovery. In *Proc. of the 14th International Society for Music Information Retrieval Conference, Curitiba, Brazil*, 2013.

[27] Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Christopher Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler. Omras2 metadata project 2009. In *Proc. of 10th International Conference on Music Information Retrieval*, 2009.

[28] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database : Popular, classical and jazz music databases.

[29] Masataka Goto. Aist annotation for the rwc music database. In *ISMIR*, pages 359–360, 2006.

[30] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, Emmanuel Vincent, et al. Methodology and conventions for the latent semiotic annotation of music structure. 2012.

[31] Geoffroy Peeters and Emmanuel Deruty. Is music structure annotation multi-dimensional ? a proposal for robust local music annotation. In *Proc. of 3rd Workshop on Learning the Semantics of Audio Signals*, pages 75–90, 2009.

[32] Roger N Shepard. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36(12) :2346–2353, 1964.

[33] Matthias Mauch and Simon Dixon. Simultaneous estimation of chords and musical context from audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6) :1280–1289, 2010.

[34] Cyril Joder, Slim Essid, and Gaël Richard. A comparative study of tonal acoustic features for a symbolic level music-to-score alignment. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 409–412. IEEE, 2010.

[35] Meinard Müller and Sebastian Ewert. Chroma toolbox : Matlab implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011. hal-00727791, version 2-22 Oct 2012*. Citeseer, 2011.

[36] Meinard Müller, Peter Grosche, and Nanzhu Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. In *ISMIR*, pages 615–620, 2011.

[37] Hanna M Lukashevich. Towards quantitative measures of evaluating song segmentation. In *ISMIR*, pages 375–380, 2008.

[38] Samer Abdallah, Katy Noland, Mark Sandler, Michael A Casey, Christophe Rhodes, et al. Theory and evaluation of a bayesian music structure extractor. 2005.

[39] Ruofeng Chen and Ming Li. Music structural segmentation by combining harmonic and timbral information. In *ISMIR*, pages 477–482, 2011.

[40] Matthias Mauch, Katy Noland, and Simon Dixon. Using musical structure to enhance automatic chord transcription. In *ISMIR*, pages 231–236, 2009.