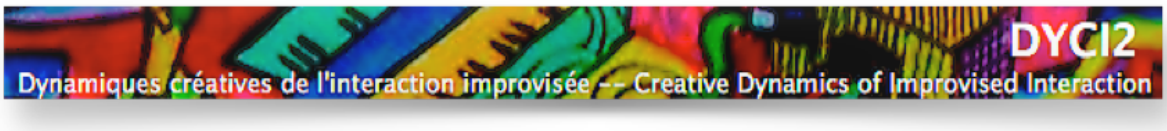

ANR-14-CE24-0002-01

Projet DYCI2,
WP1 Écoute informée créative,
SP1.2 et 1.3 Écoute structurante + Décomposition / recomposition

Rapport de livrable :

L1.3 Reconnaissance de structure de musiques répétitives polyphoniques *via* la « méthode Shazam »

Livrable	Date	Contributeurs	Rédacteur	Contenu
L1.3 Version 01	Septembre 2018	S. Fargeot, S. Marchand (Université de La Rochelle)	S. Fargeot	Maquette logicielle et base de données pour reproduire les expériences sur la méthode Shazam

Adresse du livrable logiciel

DYCI2_WP1_L1.3.zip
sur
https://forge.ircam.fr/p/Dyci2/

Université de La Rochelle

Compte-rendu Projet DYCI2

Reconnaissance de structure de musiques répétitives polyphoniques via la "méthode Shazam"

Auteur :
Simon FARGEOT

Superviseur :
Sylvain MARCHAND

21 septembre 2018

Introduction

Ce travail s'inscrit dans le cadre du projet ANR DYCI2 qui traite des interactions hommes-machines dans un contexte d'improvisation musicale. Il s'agit de proposer une méthode d'analyse d'un signal musical permettant d'en extraire des informations structurelles.

Aujourd'hui, l'industrie de la musique a couramment recours à des techniques et des outils de compositions qui mettent en jeu des structures répétitives et à base de samples. Ces techniques sont connues depuis l'apparition de la bande magnétique et se sont grandement démocratisées avec l'essor des technologies numériques, notamment à travers la musique électronique et le hip-hop. Notre travail consiste à développer un outil permettant de retrouver des informations sur la structure de musiques répétitives polyphoniques. Elle se base sur la technique d'audio-fingerprint (empreinte acoustique) rendue célèbre par Shazam, application de reconnaissance de contenu musical exact développée par Avery Wang.

1 Audio-Fingerprint

Pour mener à bien l'étude, nous nous appuyons sur les travaux effectués par Sébastien Fenet dans le cadre du projet Quaero. Plus particulièrement, il s'agit d'adapter la technique d'audio-fingerprint à la tâche de détection de structure.

1.1 Principe

L'audio-fingerprint est une technique couramment utilisée pour l'identification d'un contenu musical. Il s'agit de comparer l'empreinte acoustique d'un signal inconnu à celles de titres référencés dans une base de données, afin de retrouver les méta-données (titre, artiste, etc.) qui lui sont associées. Le calcul des empreintes acoustiques est effectué par l'analyse du signal sous sa forme spectro-temporelle. Il consiste plus particulièrement à binariser le spectrogramme en ne retenant que les coordonnées des points du spectrogramme ayant une énergie localement maximale. Pour obtenir une représentation binaire homogène sur le plan temps-fréquence, Fenet propose de diviser celui-ci en un quadrillage régulier et de retenir le maximum de chacune des cases du quadrillage, ce qui à pour principal avantage sa simplicité de mise en oeuvre.

La représentation binaire du spectrogramme présente une grande résistance à certains types de distorsions tels que l'ajout de bruit résiduel ou la compression dynamique. En effet, ces distorsions ont un impact très limité sur la répartition des pics spectraux du signal analysé.

1.2 Indexation

L'identification par empreinte acoustique d'un extrait musical inconnu ne peut être effectuée qu'au regard d'une base de données contenant les empreintes de morceaux de références. Il s'agit de retrouver dans la base, le titre ayant le plus de points actifs en commun avec l'extrait inconnu. La base de données doit alors être organisée de manière à rendre l'identification robuste et rapide. Avery Wang propose de mettre en place un système d'indexation à partir des coordonnées des points de la version binaire du spectrogramme et plus particulièrement à travailler avec des paires de points, répertorier les points un à un n'étant pas suffisamment informatif. Prenons deux points de l'empreinte acoustique de coordonnées (t_1, f_1) et (t_2, f_2). Il est possible de générer des clés d'indexations utilisant les informations croisées de ces points. Avery Wang suggère de construire les clés de la façon suivante : [f_1, f_2, t_2-t_1]. Chaque clé est associé aux méta-données de son morceau de référence : identifiant du morceau et date d'occurrence t1 de la paire. Notons qu'une indexation par paires implique un nombre de clé très conséquent puisque pour une empreinte de référence constituée de N points, N^2 clés sont calculées et inscrites dans la base. En pratique on limite le nombre de clé en ne conservant que celles de paires proches dans l'espace spectro-temporel. Ainsi, pour un point donné de l'empreinte acoustique, une zone cible relative au point est définie et seul les points inscrits dans cette zones seront utilisés pour la formation de paires avec le premier point.

1.3 Requête, analyse et identification

Pour la phase d'identification, l'empreinte acoustique du titre à identifier est calculée et une requête vers la base de données est faite pour l'ensemble des clés déduites de cette empreinte. Pour chaque paire de la requête, l'algorithme renvoie les méta-données de tous les titres de référence possédant cette paire. Il s'agit alors de trouver la référence ayant le plus grand nombre de clés cohérentes avec l'extrait inconnu. En effet, pour un extrait inconnu u de la référence r_0, commençant au temps d, toutes les clés contenues dans

u doivent être retrouvées dans r_0. Si l'on prend une clé k dont la date d'occurrence dans u est $t_{k,u}$ alors cette clé doit être retrouvée dans r_0 au temps $t_{k,r_0} = t_{k,u} + d$. Ainsi, si l'on étudie l'ensemble des valeurs $\{t_{k,r_0} - t_{k,u}\}$ pour toutes les clés k extraites de u, on devrait retrouver une accumulation maximum des valeurs $\{t_{k,r_0} - t_{k,u}\}$ autour de d. Avery Wang propose de traiter ces valeurs sous forme d'histogrammes avec un histogramme par référence. L'histogramme avec le plus haut maximum est alors celui de la référence correspondant le plus probablement à l'extrait inconnu. Toutefois, il se peut que l'extrait recherché ne soit pas dans la base de données, au quel cas le résultat de la requête retourne une mauvaise référence. Pour déterminer lorsqu'on a à faire à ce cas de figure, il est possible de calculer le rapport η entre le nombre de paires calculées pour l'extrait inconnu $n_{p,u}$ et le nombre de paires renvoyées par la meilleure référence n_{p,r_0} : $\eta = n_{p,r_0}/n_{p,u}$. Si cette valeur est inférieure à un certain seuil (en pratique $\eta < 1/2$) , alors on considère que l'extrait recherché n'est pas contenu dans la base.

1.4 Application à la détection de structure de musiques répétitives polyphoniques

La méthode Shazam est une méthode de reconnaissance de contenu exact. De ce fait, pour pouvoir l'adapter à une tâche de détection de structure musicale, nous avons choisi de nous focaliser dans un premier temps sur des musiques constituées d'un certain nombre d'échantillons musicaux (samples) répétés à l'identique à différents instants du morceau. Notons que cette technique de production est largement répandue dans l'industrie musicale actuelle. Notre objectif est donc de pouvoir retrouver, à l'aide d'une base de données constituée d'échantillons sonores élémentaires, la structure musicale d'une morceau composé d'un certain nombre de ces échantillons. Cela peut être effectué par l'analyse des histogrammes obtenus grâce à la "méthode Shazam". Il ne s'agit plus de trouver un titre unique dans la base de données, correspondant le mieux à l'extrait inconnu mais plutôt de retrouver les histogrammes de tous les échantillons contenus dans cet extrait d'une part et à partir de ces histogrammes de retrouver les occurrences de chaque échantillon dans l'extrait. En effet, si un échantillon de la base de donné est joué à plusieurs instants t_1, t_2, ..., t_N au cours de l'extrait à analyser, l'histogramme lié à cet échantillon devrait présenter un maximum pour chaque valeur $\{t_{k,r_0} - t_{k,u}\} = t_n$ avec n entier $\in [1; N]$.

Dans le domaine de la recherche de structures musicale, cette technique n'a, à notre connaissance, jamais été implémentée ni testée. Il est donc très difficile de trouver une base de test adapté à la méthode. Nous avons donc du construire de toute pièce une base de test adaptée à notre cas d'étude. Ce travail est détaillé dans la section 3.1.

2 Implémentation

L'implémentation de la méthode détaillée précédemment a été effectuée sous le logiciel Matlab. Dans un premier temps, nous nous sommes basés sur les travaux effectués par Sébastien Fenet pour le projet Quaero. Dans sa thèse, il donne une description détaillée de son implémentation notamment en ce qui concerne les valeurs types des différents paramètres permettant d'obtenir une méthode effective.

2.1 Spectrogramme binaire

La première étape consiste à analyser le signal sous sa forme spectro-temporelle. Le spectrogramme est obtenue par transformée de Fourrier à court terme fenêtré par une fenêtre de Hamming glissante de longueur $l = 64$ ms avec un recouvrement temporel de $t_{hop} = 32$ ms. Il s'agit ensuite d'en extraire les maxima locaux. Pour ce faire, Fenet propose de segmenter le spectrogramme par un quadrillage reparti de façon homogène sur le plan temps fréquence. Chaque case du quadrillage est de largeur $\Delta T_{tile} = 0.4$ s et de hauteur $\Delta F_{tile} = 400$ Hz. Ainsi, dans chaque case, la valeur maximale est mise à 1 tandis que toutes les autres valeurs sont mises à 0.

2.2 Extraction des paires

La deuxième étape consiste à créer et encoder des paires de points du spectrogramme binarisé. En théorie toutes les combinaisons de deux pics peuvent être extraites, toutefois pour limiter la quantité des paires stockés, Wang propose de n'extraire que les paires de points dont les distances spectrale $\|f_2 - f_1\|$ et temporelle $\|t_2 - t_1\|$ sont inférieures à un certain seuil, respectivement $\Delta F_{max} = 350$ Hz et $\Delta T_{max} = 3$ s. En pratique chaque point de coordonnées $(t_1,\ f_1)$ du spectrogramme binaire est sélectionné on cherche tous les points de coordonnées (t_2, f_2) qui respectent les conditions suivantes :

$0 < t1 - t2 < \Delta T_{max}$ et $-\Delta F_{max}/2 < f_2 - f_1 < \Delta F_{max}/2$. Les paires sont ensuite encodées selon la clé suivante : $[f_1, \quad f_2 - f_1 + \Delta F_{max}/2, \quad t_1 - t_2]$. Par soucis de simplicité dans la construction de la base de données, la deuxième composante de la clé est différente de celle proposée par Wang et Fenet. Cette composante $f_2 - f_1 + \Delta F_{max}/2$ a été préférée à simplement f_2 pour limiter la taille de la dimension qui lui est associée dans la base de données.

2.3 Apprentissage, construction d'une base de données de référence

La phase d'apprentissage de cette méthode consiste à créer une base de données constituée d'extraits sonores de référence. Pour chaque morceau les clés sont calculées, il faut alors être capable de stocker l'information qui leur est associée dans une base de données de manière optimale. Dans un soucis de rapidité d'exécution Avery Wang propose d'avoir recours à une table de hachage pour stocker les données. Sous Matlab une solution efficace est d'utiliser une structure à 3 dimensions dont chaque dimension correspond à une composante de la clé. Ainsi la première dimension correspond à f_1, la deuxième à $f_2 - f_1 + \Delta F_{max}/2$ et la 3ème à $t_1 - t_2$. De cette manière il est possible de définir la taille de chaque dimension :

- la première dimension allant de 0 à $F_{max} = F_s/2$ avec une résolution fréquentielle $r_1 = 10$ Hz,
- la deuxième dimension allant de 0 à ΔF_{max}, avec une résolution fréquentielle $r_2 = r_1 = 10$ Hz,
- la troisième dimension allant de 0 à ΔT_{max}, avec une résolution temporelle $r_3 = t_{hop} = 32$ ms.

Pour une signal sonore échantillonné à 44,1 kHz, la taille de la structure sera de $F_s/2/r_1 \times \Delta F_{max}/r_2 \times \Delta T_{max}/r_3$, soit $2205 \times 35 \times 94$.

Pour créer une base de référence, on crée une structure vide de cette taille qu'il s'agira de remplir au fur est à mesure que les clés sont calculées. Par exemple pour un titre de référence identifié m_1, lorsqu'une paire de points (t_1, f_1), (t_2,f_2) est extraite, le couple de valeurs $[m_1, t_1]$ sont stockées dans la structure à l'adresse correspondant à la clé. Notons qu'une clé peut être retrouvée à différents instants et dans différents extraits, auquel cas celle-ci pointe vers une liste contenant l'ensemble des couples qui lui sont associés.

2.4 Requête et fabrication des histogrammes

Pour analyser le contenu d'un extrait inconnu, il s'agit d'extraire toutes les paires de l'extrait et d'en déduire les clés associées selon la méthode décrite précédemment. Des requêtes sont ensuite émise vers la base de données à partir des clés calculées pour l'extrait inconnu. Le résultat des requêtes se présente sous la forme d'un ensemble d'histogrammes, chaque histogramme étant associé à un échantillon de référence (cf : section 1.3). Dans notre cas d'étude, la résolution temporelle des histogrammes doit être la plus fine possible puisqu'il s'agit de retrouver précisément les temps d'occurrence de chaque échantillon de référence contenu dans l'extrait inconnu. Le résolution temporelle minimum correspond donc au recouvrement temporelle choisi pour la STFT, soit $r_{hist} = t_{hop} = 32$ ms. Pour chaque requête les histogrammes sont mis à jour en accumulant les valeurs obtenue précédemment avec celles calculées à partir de la requête actuelle.

2.5 Analyse des histogrammes et détection de structure

Une fois l'ensemble des histogrammes calculées, la détection de structure se présente en une tâche double. D'abord détecter les échantillons contenu dans l'extrait inconnu et ensuite déterminer les occurrences de chaque échantillon dans cette extrait. Pour détecter les échantillons contenus dans l'extrait inconnu, un score est attribué à chaque histogramme en fonctions des caractéristiques suivantes :

— valeur du max de l'histogramme $x_{hist,max}$,
— nombre de matchs entre extrait inconnu et échantillon n_{match},
— valeur moyenne de l'histogramme $\overline{x}_{hist}$.

Les histogrammes sont ensuite classés selon ce score. Pour déterminer les occurrences des échantillons dans un extrait à partir des histogrammes, il s'agit de détecter des pics francs correspondant à une accumulation de matchs entre l'extrait inconnu et l'échantillon à un instant donné. On considère un pic lorsque la valeur de l'histogramme est supérieure à une valeur seuil : $x_{seuil} = k \times \overline{x}_{hist}$ (avec k = 18 dans notre cas).

3 Test et résultats

3.1 Base de test

Pour tester l'algorithme il a fallu confectionner une base de test cohérente vis à vis de notre cas d'étude, à savoir détection de structure pour des musiques polyphoniques répétitives, polyphoniques signifiant ici qu'il s'agit de musiques composées de plusieurs éléments musicaux (ex : batterie, guitare, piano, basse) et répétitives signifiant que chaque voix de la polyphonie reproduit dans le temps un motif particulier, sans variation. N'ayant pas trouvé de base de test appropriée à nos besoin nous avons du la fabriquer de toute pièce. 6 extraits musicaux on ainsi été composés, chacun étant constitué d'un certain nombre d'échantillons, de durée variable (de moins d'une seconde à 30 secondes), répétés à différents instants de l'extrait. Voici les caractéristiques de chaque extrait.

- Extrait 1 : 4 voix (batterie, nappe synth, arpège e-piano, basse synth), durée : 1 : 30,
- Extrait 2 : 3 voix (batterie, arpège synth, basse électrique), durée 2 : 22,
- Extrait 3 : 4 voix (batterie, accords, piano motif, synth mélodie), durée : 1 : 56,
- Extrait 4 : 7 voix (4 percus, 2 nappe synth, basse synth), durée : 1 : 39,
- Extrait 5 : 5 voix (batterie, shaker, e-piano, synth, basse synth), durée : 1 : 41,
- Extrait 6 : 4 voix (batterie, e-piano, nappe synth, basse synth), durée : 2 : 01.

La base de données de référence est donc constituée de 27 échantillons sonores et les requêtes sont effectuées sur les extraits entiers. La structure de chaque extrait est connue et enregistrée dans un fichier texte (cf : dossier analyse/structure_ref).

3.2 Résultats

Les premiers résultats sont présentés dans cette section. Pour estimer les performances de la méthode, il s'agit de comparer la structure d'un extrait retrouvée par l'algorithme avec la réelle structure de cet extrait. Cette comparaison est facilement réalisable graphiquement. La figure 2 présente les

résultats pour l'extrait 1. Les histogrammes sont identifiés par l'échantillon auquel ils se réfèrent (ex : 1.3 correspond à l'échantillon 3 du mix 1).

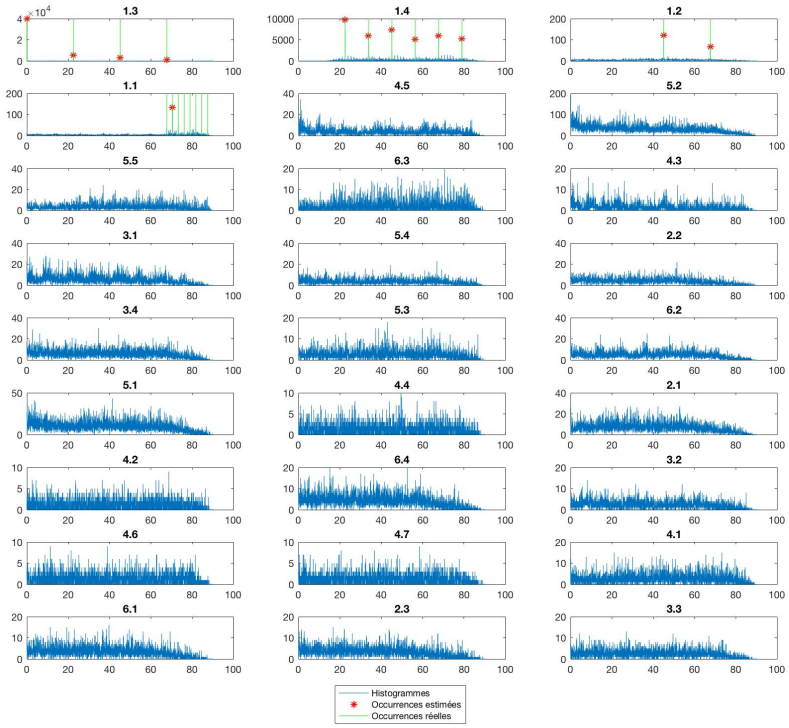

FIGURE 1 – Histogrammes obtenus par l'analyse de l'extrait 1, (classés par ordre de pertinence, selon le score décrit en section 2.5). L'axe des abscisses représente le temps (en s), l'axe des ordonnées représente le nombre de matchs entre l'extrait inconnu et l'échantillon.

Pour le premier extrait il est possible de constater que les 4 premiers histogrammes correspondent bien aux 4 échantillons qui constituent cet extrait. Les autres histogrammes ont globalement une valeur moyenne moins élevée et aucun pic n'émerge du bruit moyen, ce qui signifie qu'il y a peu de paires communes entre l'extrait testé et les échantillons associés à ces histogrammes et que ces paires sont le fruit du hasard. En revanche les histogrammes 1.3, 1.4, 1.2 et 1.1 présentent des particularités. La figure 2 est un zoom sur ces

4 histogrammes.

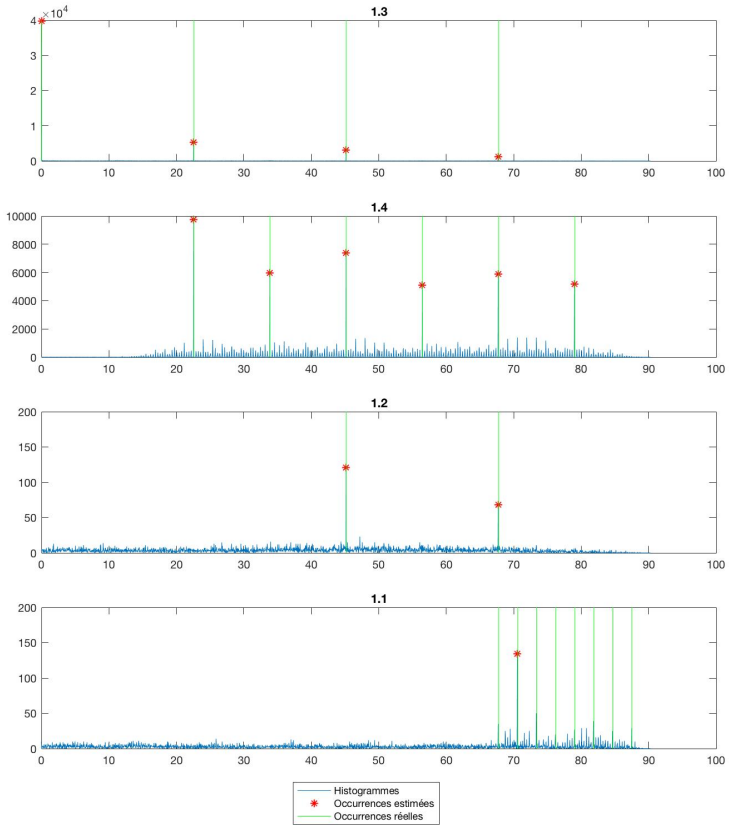

FIGURE 2 – Zoom sur les 4 histogrammes correspondant aux 4 échantillons qui constituent l'extrait 1.

Sur la figure 2, on observe que chaque occurrence réelle des échantillons dans l'extrait (représentées en vert) concorde avec un maximum local de l'histogramme. La méthode a permis de retrouver la structure quasiment intégrale de l'extrait 1. Toutefois on peut voir pour l'échantillon 1.1 que toutes les occurrences n'ont pas été retrouvées. En effet, cet échantillon étant très court par rapport aux autres échantillons du même extrait, peu de paires lui sont associées dans la base de données ce qui signifie également que le nombre de matchs entre l'extrait et cet échantillon est forcement bien plus faible

que pour des échantillons plus long. On remarque également que l'histogramme 1.2 a une valeur maximum nettement inférieure à celle des histogrammes 1.3 et 1.4. Cela peut s'expliquer par le fait que l'échantillon 1.2 est un échantillon de basse. Or nous avons pu constater que la méthode actuelle présente des performances mitigées en basses fréquences. En effet, la résolution fréquentielle utilisée pour l'analyse est une résolution linéaire or en musique l'évolution des fréquences est logarithmique (un écart d'une octave correspondant à un rapport de fréquence égal à 2). Ainsi, une solution pour améliorer les performances en basse fréquence serait par exemple d'analyser le signal avec une transformée à Q constant au lieu d'une simple FFT. Enfin, les histogrammes relatifs à des séquences rythmiques (type percussion) présentent de bons résultats. De plus à l'instar de l'histogramme 1.4, des sous-structures en émergent, ce qui pourrait dans un second temps donner des informations sur le tempo et sur le type de mesure.