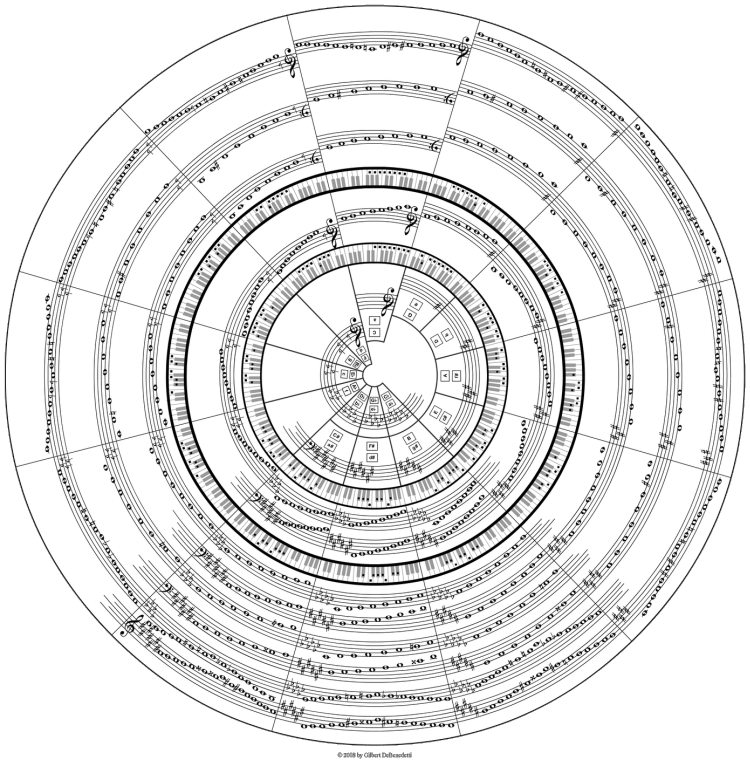
INSA | INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

Département Génie Mathématique

Projet de transcription de musique

Projet semestriel – semestre 8



Rand ASSWAD
Ergi DIBRA
Yuge SUN

A l'attention de :
Mme. Natalie FORTIER

11 juin 2018

Contents

"La vie sans musique est tout simplement une erreur, une fatigue, un exil." — **Friedrich Nietzsche**

1 Introduction

La musique peut être considérée comme la première langue parlée. Bien que la musique soit un art, sa perception est un effort artistique aussi bien que scientifique.

En effet, l'humain cherchait à caractériser la musique afin de pouvoir la conserver et de la partager. Cette science a commencé par l'écriture des chants hourrites sur des tablettes d'argile extraite d'Ougarit (actuellement Latakié, Syrie) qui remontent approximativement à 1400 av. J.-C. [Wikipédia, 2018]

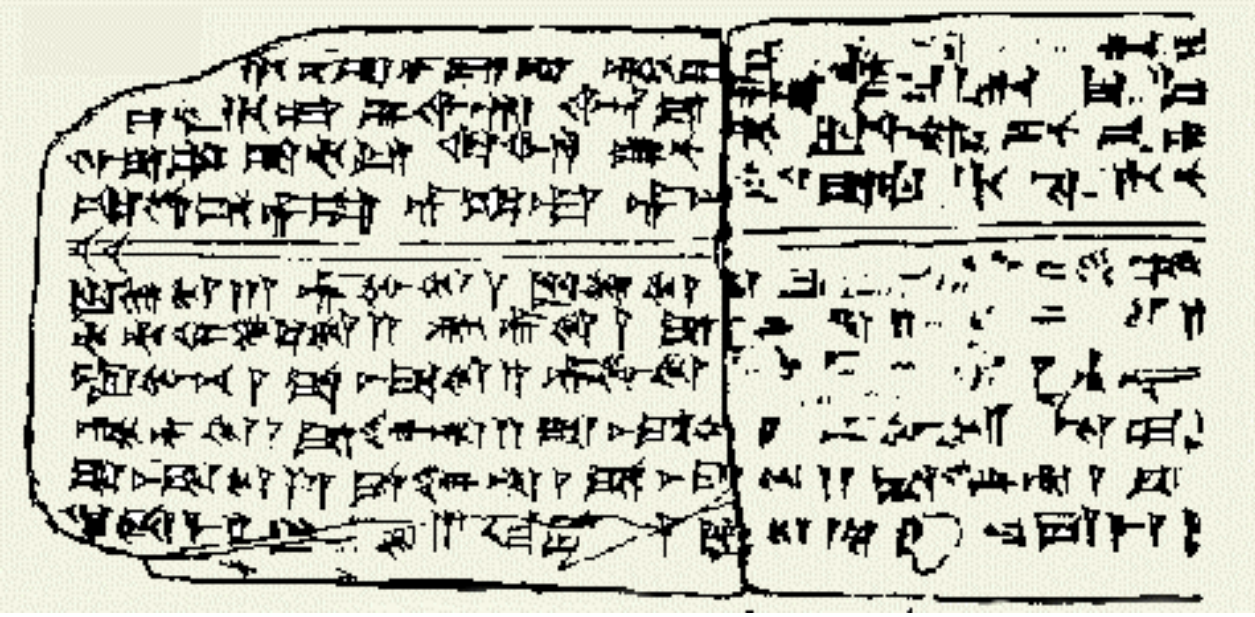


Figure 1: Dessin de la tablette de l'hymne à Nikkal (chant hourrite)

La notation musicale a évolué au cours de l'histoire, le système occidental développé au moyen âge appelé "*solfège*" est adopté aujourd'hui par tous les musiciens du monde avec quelques variations. La transcription d'une œuvre musicale est appelée une *partition*.

Ce système de notation est fidèle à la perception humaine du son, vis-à-vis la linéarité de fréquences et la reconnaissance du rythme. En effet, on tappe les pieds suivant le rythme de la musique sans aucune connaissance musicale, et la majorité de personnes est capable de chanter une mélodie sans lire sa partition.

En revanche, un signal sonore contient les mêmes informations sous une forme différente; l'espace de fréquences n'est pas linéaire et le rythme n'est pas reconnaissable facilement.

Dans ce projet, nous avons étudié et implémenté des méthodes d'interprétation de signaux sonores dans l'objectif de pouvoir produire une partition de musique à partir d'un son enregistré.

Des nombreuses recherches ont été effectuées sur ce sujet et les résultats obtenus sont certainement intéressants. Malheureusement, ces méthodes ne traitent que des cas particuliers.

Nous nous sommes intéressés par ce domaine d'applications car il porte un fort potentiel dans le futur. Ce projet nous a permis de faire notre premier pas dans ce domaine et de mettre en place nos compétences en mathématiques, en informatique et en musique.

2 Signaux sonores

2.1 Son harmonique

Le son d'un résonateur acoustique comme une chorde ou une colonne d'air est une onde stationnaire. On dit que tel son évoque un **pitch défini**. Dans le cas des instruments de percussion, le son présente une *inharmonicité*. On dit que tel son évoque un **pitch indéfini**. Dans ce projet on ne s'intéressera qu'au sons harmoniques de pitch défini.

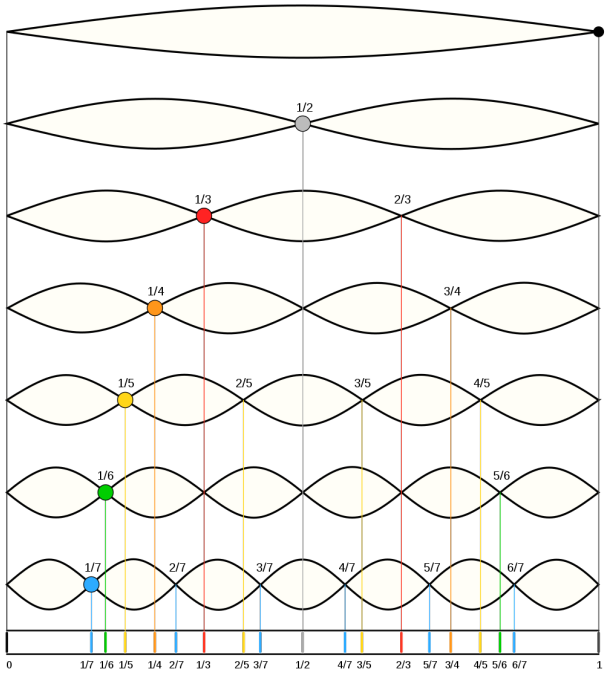


Figure 2: Les harmoniques d'une chorde vibrante

Un signal sonore de pitch défini, est une série harmonique de sons purs, représenté par des ondes sinusoïdales dont les fréquences sont des multiples **entiers** d'une fréquence dîte la **fondamentale** (où le **pitch**) notée f_0.

$$x(t) = \sum_{k \in \mathbb{N}} A_k \cdot \cos(2\pi k f_0 t)$$

où A_k est l'amplitude de la k$^{\text{ème}}$ harmonique.

On cherche donc à indentifier f_0 dans un signal harmonique donnée.

2.2 Discrétisation et échantillonnage

La numérisation d'un signal consiste à prélever des valeurs du signal à intervalles définis. Les valeurs obtenues sont appelées des *échantillons*.

La *période d'échantillonnage* T_s est l'intervalle de temps entre deux échantillons, on définit $f_s = \frac{1}{T_s}$ le nombre d'échantillons prélevés par secondes, f_s est dît *fréquence d'échantillonnage* ou en anglais **sample rate**.

On note $x[n] = x(t_n)$ où $t_n = n \cdot T_s = \frac{n}{f_s}$. Dans le reste du projet, on notera toujours $[\cdot]$ les valeurs discrètes.

L'échantillonnage d'un signal consiste à choisir une fréquence d'échantillonnage sans perdre de valeurs importantes du signal. En traitement de signaux sonores, f_s est souvent égale à $44.1kHz, 22.05kHz, 16kHz$, ou $8kHz$. La numérisation d'un signal dépend aussi d'autre facteurs comme le *bit depth* (i.e. le nombre de bits pour stocker chaque échantillon), mais nous ne nous intéressons pas par les détails; les bases de l'échantillonnage de signaux sont expliquées et démontrées par le théroème d'échantillonnage de **Nyquist-Shannon**.

2.3 La transformée de Fourier (FT)

La transformée de Fourier se définit par:

$$\hat{x}(f) = \int_{-\infty}^{\infty} x(t) \cdot e^{-2\pi j f t} \mathrm{d}t$$

Cette transformation permet d'identifier la fréquence d'une fonction périodique. En effet, La transformée de Fourier représente l'intensité d'une fréquence dans un signal, donc ses pics correspondent aux fréquences du signal.

Comme la transformée de Fourier est linéaire, la transformée d'un signal harmonique produit plusieurs pics.

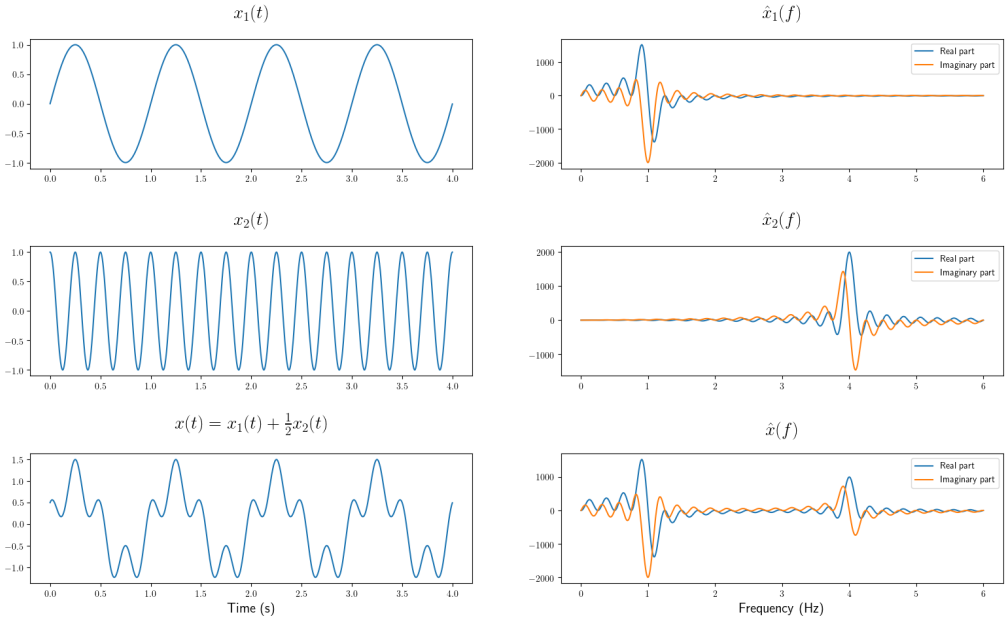


Figure 3: Linéarité de la transformée de Fourier

2.4 La transformée de Fourier discrète (DFT)

Soit N le nombre d'échantillons pris sur l'intervalle $[0, t_{\max}[$, soit f_s la fréquence d'échantillonage. Pour $n = 0, 1, \ldots, N-1$ on a:

$$\begin{aligned}
\hat{x}(f) &= \int_{0}^{t_{\max}} x(t) \cdot e^{-2\pi j f t} \mathrm{d}t \\
&= \lim_{f_s \to \infty} \sum_{n=0}^{N-1} x(t_n) \cdot e^{-2\pi j f t_n} \\
&= \lim_{f_s \to \infty} \underbrace{\sum_{n=0}^{N-1} x[n] \cdot e^{-2\pi j f \frac{n}{f_s}}}_{\hat{x}[f]} \\
&= \lim_{f_s \to \infty} \hat{x}[f]
\end{aligned}$$

La DFT de $x[n]$ se définit donc par:

$$\hat{x}[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-2\pi j k \frac{n}{f_s}}$$

Remarque: La DFT se calcule souvent matriciellement pour économiser les calculs. De plus, dans le cas où $N = 2^p, p \in \mathbb{N}$ on calcule la transformée de Fourier rapide (FFT) qui utilise le symétrie pour minimiser le nombre de calculs.

2.5 Fenêtrage

Nous avons souvent besoin de traiter le signal sur une durée limitée, on définit donc une fonction à support compact w et on étudie le produit de convolution du signal avec la fenêtre.

Voici quelques exemples de fenêtres:

- Fonction rectangulaire:

$$\text{rect}_{[0,T]}(t) = \begin{cases} 1 & \text{si } t \in [0,T] \\ 0 & \text{sinon} \end{cases}$$

- Fenêtre Hann:

$$w(t) = \sin^2\left(\frac{\pi t}{T}\right) \cdot \text{rect}_{[0,T]}(t)$$

- Fenêtre Welch (fenêtre parabolique):

$$w(t) = 1 - \left(\frac{2t - T}{T}\right) \cdot \text{rect}_{[0,T]}(t)$$

Nous avons choisi d'utiliser la fenêtre de Hann dans notre projet car elle attenue le phénomène **aliasing** qui rend les signaux indisntinguables lors de l'échantillonnage.

$$w[n] = \sin^2\left(\frac{\pi n}{N-1}\right) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$$

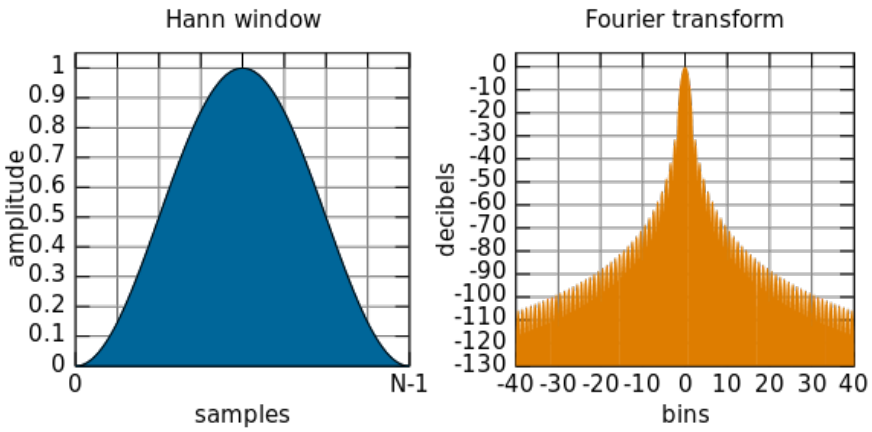


Figure 4: La fenêtre Hann est sa transformée de Fourier

2.6 La transformée de Fourier à court terme (STFT)

La transformée de Fourier nous permet d'obtenir les fréquences d'un signal harmonique. Or, la fréquence d'un signal peut changer en fonction du temps, on voudrait donc avoir la transformée de Fourier en fonction de la fréquence *et* du temps.

La transformée de Fourier à court terme $X(t, f)$ est la transformée de Fourier de x sur une fenêtre glissante w centrée en t (i.e. $w(\tau - t)$).

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau) \cdot w(\tau - t) \cdot e^{-2\pi j f \tau} \mathrm{d}\tau$$

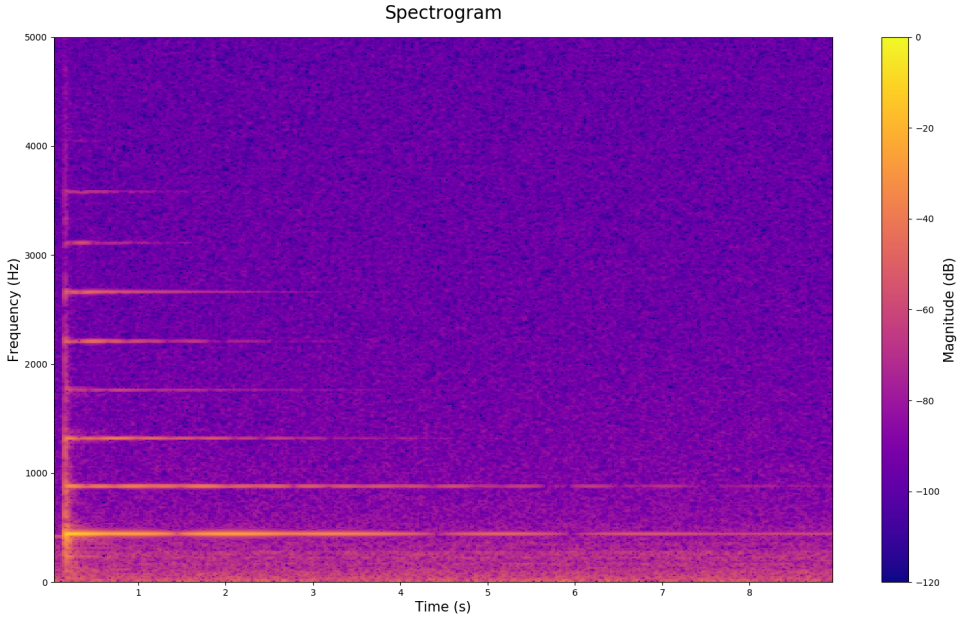
De même, la STFT discrète se définit:

$$X[n,k] = \sum_{n=0}^{N-1} x[m] \cdot w[m-n] \cdot e^{-2\pi j k \frac{m}{f_s}}$$

2.7 Spectrogramme

Le spectrogramme permet de visualiser les changement de fréquences en fonction du temps, il se définit par:

$$S(t,f) = |X(t,f)|$$



Remarque: Si on voudrait visualiser la puissance spectrale d'un signal, on prend le carré de la module de la STFT.

3 Pitch

Ils existent plusieurs algorithmes de détection de fréqences fondamentales, il y'en a deux types : applications sur le domaine temporel et sur le domaine fréquenciel. Les applications sur le domaine fréquenciel calculent les fréquences à partir de la transformée de Fourier du signal, où les méthodes du domaine temporel les calcules à partir du signal sans passer par la transformée de Fourier.

Chaque type présente des avantages et des inconvénients. Nous avons décidé d'implémenter une de chaque type:

3.1 YIN

L'algorithme de YIN *(Kawahara et de Cheveigné, 2002)* est une méthode robuste pour la reconnaissance du pitch, il s'agit d'un modèle temporel. Son principe est la séléction de fréquences candidats parmi toutes les fréquences détéctés sur l'intervalle de fenêtrage.

La méthode propose que l'expression $x(t) - x(t+\tau)$ atteint son minimum quand τ est égale à la période du signal (i.e. $\frac{1}{f_0}$). En diffinissant la fonction de différence à l'instant t fixé:

$$d_t(\tau) = \int_t^{t+T_w} \left(x(t) - x(t+\tau)\right)^2 \mathrm{d}t$$

où T_w est la taille de la fenêtre w, on appelle τ le *retard* (anglais: *lag*).

Soit en temps discret:

$$d_n[m] = \sum_{i=n+1}^{n+N_w} \left(x[n] - x[n+m]\right)^2$$

Par la suite, on calcule la fonction de la moyenne cumulative définie par:

$$d'_t(\tau) = \begin{cases} 1 & \mathrm{si} \tau = 0 \\ d_t(\tau) / \frac{1}{\tau} \int_0^\tau d_t(u) \mathrm{d}u & \mathrm{sinon} \end{cases}$$

Soit en temps discret:

$$d'_n[m] = \begin{cases} 1 & \mathrm{si} m = 0 \\ d_n[m] / \frac{1}{m} \sum_{i=0}^{m} d_t[i] & \mathrm{sinon} \end{cases}$$

Les candidats sont les minimums locaux de d'_n.

3.2 YIN spectrale

L'algorithme de YIN spectrale *(Paul Brossier, 2006)* est une méthode qui utilise la même logique de l'algorithme de YIN et l'applique sur la STFT.

La fonction de différence est définie par:

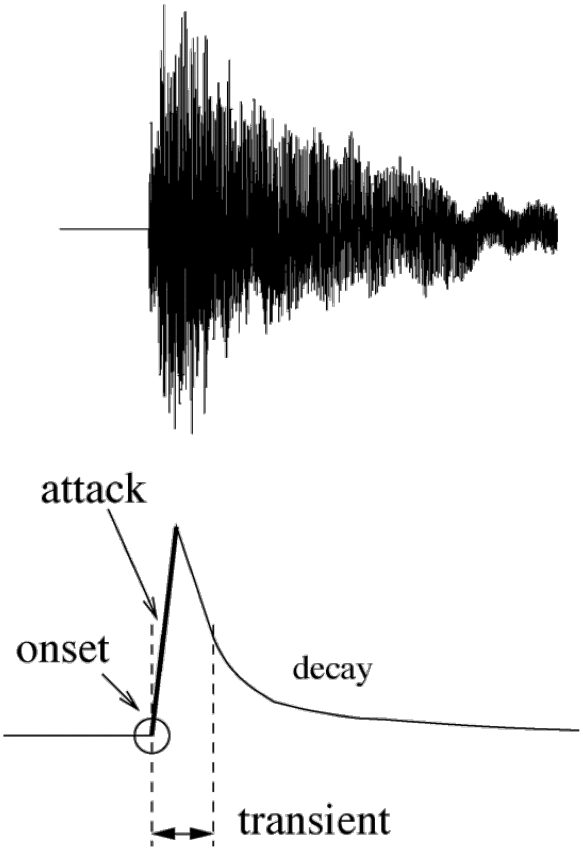
$$\hat{d}_n[m] = \frac{2}{N} \sum_{k=0}^{\frac{N}{2}+1} \left| \left(1 - e^{2\pi jkm/N}\right) X[n,k] \right|^2$$

La fonction de la moyenne cumulative se calcule de façon analogue à l'algorithme de YIN. L'algorithme cherche le minimum globale de cette dernière.

4 Segmentation temporelle

L'étape fondamentale dans la reconnaissance du son est la segmentation temporelle. Il s'agit de trouver les frontières des objets sonores, c'est-à-dire:

- Le début de la note – dît ***onset***.
- La fin de la note – dît ***offset***.



Attack, transient, decay, onset IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 13, NO. 5, SEPTEMBER 2005

Cette étape dépend fortement sur le type du son produit; les instrument à chordes pincées (guitare, piano, oud, etc.) ont un profile différent de celui des instruments à cordes frottées (la famille du violon) ou de celle des instruments à vent.

Dans cette partie on expliquera les méthodes implémentés pour la reconnaissance du **onset**.

4.1 Méthode

La lecture scientifique indique une méthode rigoureuse qu'on a simplifier pour obtenir des résultats rapide.

Il s'agit de définir une fonction qui permet de quantifier la perturbation du signal à un moment donné, cette fonction est souvent appellée **Onset Detection Function** ou **Onset Strength Signal**, dans ce projet on fera référence à cette dernière par **Onset Detection Function** ou ***ODF***.

Théoriquement, les maximums locaux de l'ODF sont les onsets du signal, mais en pratique il s'agit d'un sous-ensemble de ces points. En effet, l'ODF est souvent très sensible et détectera la moindre des perturbations.

Ce problème pourra être résolu en définissant un seuil au dessous duquel aucun onset est considéré. Ils existent plusieurs méthodes pour définir tel seuil.

Soit un seuil fixe, ce qui minimise le coût des calculs au prix de la qualité des résultats. Soit de calculer un seuil variable, il s'agit de lisser la fonction ODF par des méthodes classiques comme la moyenne mobile.

La méthode consiste donc en trois étapes:

1. Calcul de l'**Onset Detection Function**.

2. **Thresholding**: calcul du seuil.
3. **Peak-picking**: la selection des onsets.

4.2 Onset Detection Function (ODF)

Ils existent plusieurs fonction de détéction d'onsets, on expliquera quelques unes qui se basent sur la STFT.

4.2.1 High Frequency Content (HFC)

Il s'agit de priviligier les fréquences élevées dans un signal:

$$HFC[n] = \sum_{k=1}^{N} k \cdot |X[n,k]|^2$$

4.2.2 Phase Deviation (Phi)

Il s'agit de calculer les différences de phases en dérivant l'argument complex de la STFT, on note

$$\varphi(t,f) = \arg(X(t,f))$$

$$\hat{\varphi}(t,f) = \text{princarg}\left(\frac{\partial^2 \varphi}{\partial t^2}(t,f)\right)$$

où

$$\text{princarg}(\theta) = \pi + ((\theta + \pi) mod(-2\pi))$$

donc la ODS de phase se calcule par la formule:

$$\Phi[n] = \sum_{k=0}^{N} |\hat{\varphi}[n,k]|$$

Dans notre implémentation, nous avons approximé la dérivée partielle seconde de la phase par un schéma de Taylor d'ordre 2.

4.2.3 Complex Distance

Cette méthode permet de qualifier les changements spectraux du signal ainsi que les changements en phase. Il s'agit de calculer une prédiction du spectre du signal, et puis le comparer par sa valeur. On reprend la fonction calculée en $\hat{varphi}(t,f)$ de la méthode précédante. On définit la prédiction :

$$\hat{X}[n,k] = |X[n,k]| \cdot e^{j\hat{\varphi}[n,k]}$$

Donc la distance complexe se calcule:

$$DC[n] = \sum_{k=0}^{N} \left|\hat{X}[n,k] - X[n,k]\right|^2$$

4.3 Thresholding

Nous avons décidé de lisser la fonction ODF par une moyenne mobile echelonnées par la fenêtre Hann, il s'agit du produit de convolution de l'ODF avec la fonction Hann.

5 Théorie de musique

5.1 Introduction du problème

L'espace de notes est un espace linéaire discrèt, mais l'espace de fréquences est continue non-linéaire. Le problème consiste à trouver une fonction qui associe les fréquences fondamentales obtenues avec des valeurs entières.

5.2 Gammes et intervalles

En acoustique, un **intervalle** désigne le rapport de fréquences de deux sons. Or, chaque intervalle est caractéristique d'une échelle musicale, elle-même varie selon le type de musique.

En musique, une **gamme** (*en:* **scale**) est une suite de notes conjointes où la fréquence de la dernière est le double de celle de la première. Une gamme se caractérise par sa première note et la suite d'intervalles qui séparent les notes conjointes.

L'**armure** — ou l'**armature** (*en:* **key signature**) — est un ensemble d'altérations réunies à la clé. Elle est composée soit exclusivement de dièses, soit exclusivement de bémols — en dehors du cas particulier constitué par le changement d'armure. Ces altérations correspondent à la tonalité principale des mesures suivant la clé.

À chaque tonalité majeure est associée une tonalité en mode mineur, présentant la même armure de clef et appelée relative mineure.

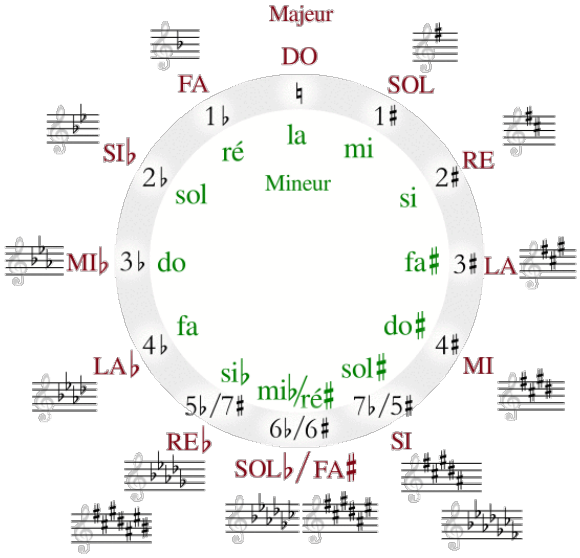


Figure 5: Le cycle des quintes

Pour simplifier, on va considérer la théorie de la musique occidentale basée sur l'accord tempéré (depuis le XVIII^{e} siècle). Dans ce cas, l'intervalle séparant la première et la dernière note d'une gamme est dite *octave*, une octave se divise en 12 écarts égales appelés *demi-tons*. La dernière note porte le même nom de la première dans la gamme.

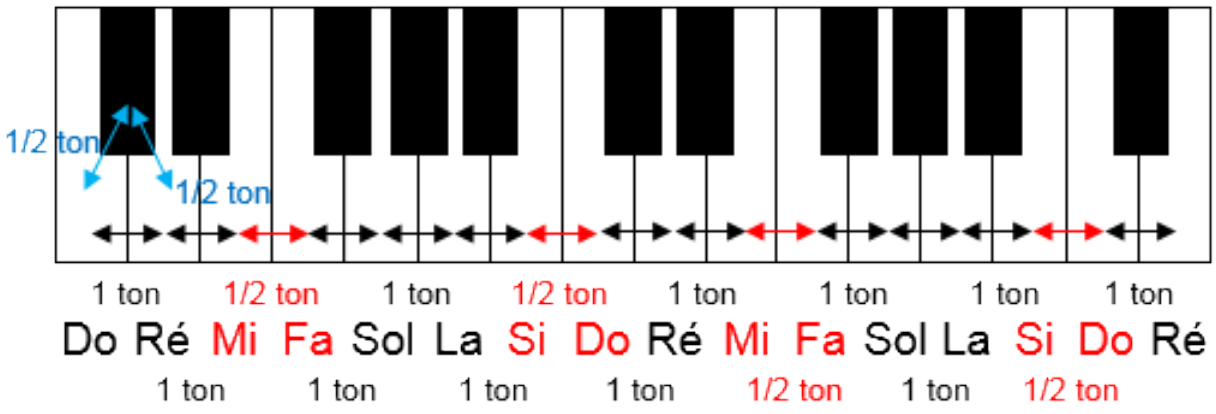


Figure 6: Les intervalles sur un piano

5.3 Nomenclature

Ils existent plusieurs systèmes de nomenclature de notes de musique. Le système utilisé en France adopte les noms en termes de *Do-Ré-Mi-Fa-Sol-La-Si*. De plus, il existe un système basé sur l'alphabet latin : *C-D-E-F-G-A-B*. Les deux systèmes sont très utilisés, dans ce projet on utilisera le dernière pour simplifier.

Vu que les noms des notes se répètent au bout d'un octave, il faut distinguer une note LA de fréquence $440Hz$ d'une autre de fréquence $220Hz$ ou $880Hz$.

Le système de notation scientifique **Scientific Pitch Notation** identifie une note par sont nom alphabetique avec un nombre identifiant l'octave dans laquelle elle se situe, où l'octave commence par une note C. Par exemple la fréquence $440Hz$ représente A_4 sans ambiguité, et les fréquences $220Hz$ et $880Hz$ représentent les notes A_3, A_5 respectivement.

Dans le protocole **MIDI**, le notes sont représentées par un nombre entier, il permet de coder plus de 10 octave en partant de la note C_{-1}.

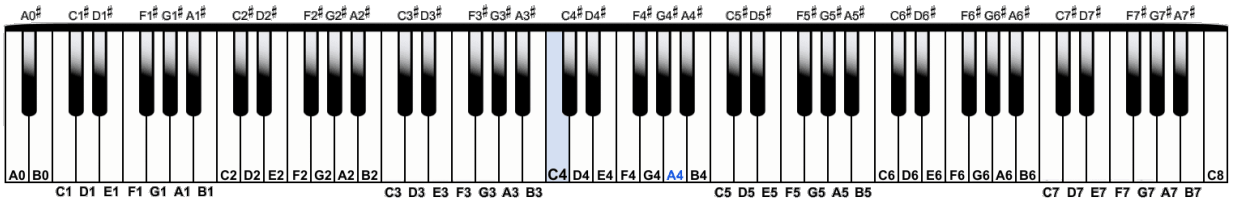


Figure 7: La notation scientifique sur un piano

5.4 Reconnaissance des notes

Un demi-ton est l'écart entre deux touches voisines sur un piano. On voudrais savoir le rapport r de fréquences associé à un demi-ton, sachant que l'octave double la fréquence on peut conclure facilement :

$$r^{12} = 2 \Rightarrow r = 2^{1/12}$$

On souhaite ramener l'espace de fréquences $(\mathbb{R}, \times)$ à l'espace $(\mathbb{N}, +)$ tel que $\boxed{\text{demi-ton} \equiv 1}$. On définit donc une bijection

$$\forall f \in]0, \infty[, f \mapsto 12 \log_2 f$$

En arrondissant le résultat à la valeur entière la plus proche, on obtient un espace linéaire discrèt correspondant aux notes.

Il sera convenient d'obtenir les mêmes notes du protocole **MIDI** vu qu'il est très bien établi et très utilisé. Pour cela, on effectue une petite translation, en partant de la note de référence $A4 \equiv 69_{\text{MIDI}} \equiv 440Hz$.

$$\begin{cases} \varphi : f \mapsto 12 \log_2 f + c_{\text{ref}} \\ \varphi(440) = 69 \end{cases} \Rightarrow c_{\text{ref}} = 69 - 12 \log_2 440$$

Par conséquent, la bijection φ est définit par :

$$\varphi :]0, \infty[\to \mathbb{R} : f \mapsto 12 \log_2 f + c_{\text{ref}} \quad \text{avec } c_{\text{ref}} = 69 - 12 \log_2 440$$

On note $\bar{\varphi}$ la fonction définit par $\bar{\varphi}(f) = \lfloor \varphi(f) \rceil \in \mathbb{Z}$ où $\lfloor \cdot \rceil$ est la fonction d'arrondissement à l'entier le plus proche.

On peut donc obtenir les nombres MIDI de notes à partir des fréquences fondamentales grâce à la fonction $\bar{\varphi}$.

Néanmoins, le nombre MIDI n'est pas suffisant pour identifier une note, car certaines notes ont la même fréquence en accord tempéré et donc le même nombre midi (i.e. la même touche sur un piano), par exemple $\text{MIDI}(C\#) = \text{MIDI}(D\flat)$. Pour distinguer ces notes il est nécessaire de trouver la gamme du morceau.

5.5 Reconnaissance de la gamme/l'armature

Dans cette étude, on ne s'intéressera aux notes dans une octave. On introduit donc la fonction ψ :

$$\psi :]0, \infty[\to [0, 12[: f \mapsto \psi(f) \mod 12$$

De même, on définit la fonction $\bar{\psi}$ telle que $\bar{\psi}(f) = \lfloor \psi(f) \rceil$. On voit que $\mathrm{Im}(\bar{\psi}) = \mathbb{Z}/12\mathbb{Z}$

Note	C		D		E	F		G		A		B
$\bar{\psi}(f)$	0	1	2	3	4	5	6	7	8	9	10	11

En musique classique, ils existent 4 types de gammes, on ne s'intéressera qu'à un : *la gamme majeure*. Comme on l'a déjà dit, une gamme est caractérisée par sa première note et la suite des intervalles. Dans la gamme majeure, les intérvalles en fonction du ton sont : 1–1–½–1–1–1–½.

La gamme *Do/C Majeur* contient donc les notes {0, 2, 4, 5, 7, 9, 11}.

De même, la gamme *Sol/G Majeur* contient les notes {7, 9, 11, 0, 2, 4, 6}. Ces gammes diffèrent par une note, la note 5≡F est remplacée par la note 6 qui correspond à F# ou G♭. Dans le contexte du Sol Majeur, on sait que 6≡F# car la gamme contient déjà 7≡G.

On voit bien que l'identification de la gamme est *nécessaire* pour la distinction entre certaines notes.

Une gamme peut être alors identifié par son ensemble de notes qu'on notera G tel que $G \subset \mathbb{Z}/12\mathbb{Z}, |G| = 7$. On définit le vecteur $g \in \{0,1\}^{12}$ associé à G tel que

$$g_i = \mathbb{1}_G(i) = \begin{cases} 1 & \text{si } i \in G \\ 0 & \text{sinon} \end{cases}$$

On définit donc E l'ensemble de gammes majeures.

Soit F l'ensemble de fréquences fondamentales obtenues, soit $S = \bar{\psi}(F) \subset \mathbb{Z}/12\mathbb{Z}$, soit $p : Z/12\mathbb{Z} \to \mathbb{N} : n \mapsto$ le nombre d'occurances de n dans le morceau. On note $p_{\max} = \max_{n \in S} p(n)$. On définit le vecteur $x \in [0,1]^{12}$ tel que $x_i = \frac{p(i)}{p_{\max}}$.
La gamme du morceau est alors la solution du problème d'optimisation

$$\min_{g \in E} \quad \|g - x\|$$

En musique classique, $|E| = 12$ donc le problème d'optimisation ne nécessite pas une résolution mathémtique avancée.

References

Wikipédia. Chants hourrites — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Chants_hourrites&oldid=145548895.