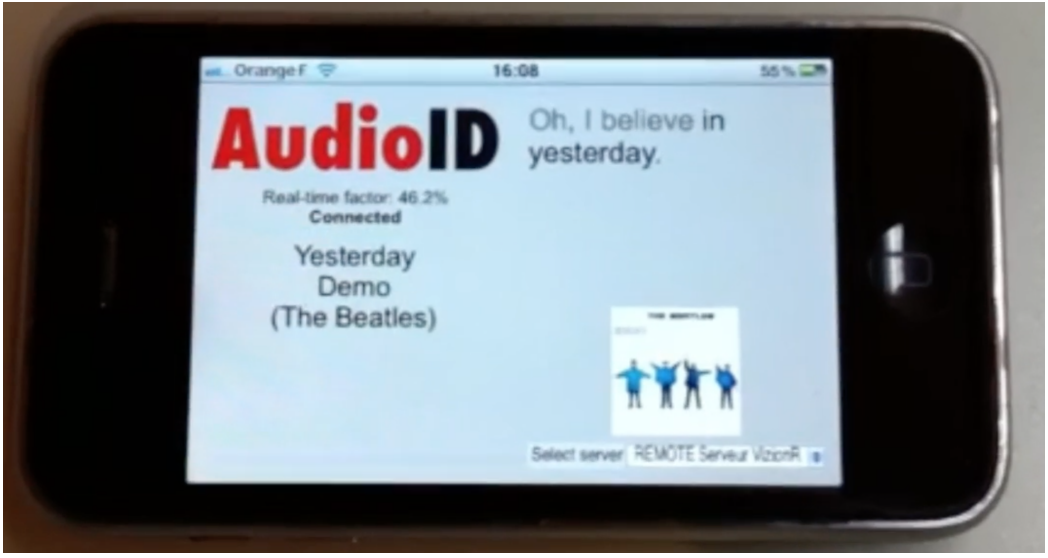
ENSEA 2ème AuPAM (2017-2018)
Traitement du signal audio musical, descripteurs et estimation

Geoffroy.Peeters@ircam.fr
UMR SMTS 9912 (IRCAM CNRS UPMC)

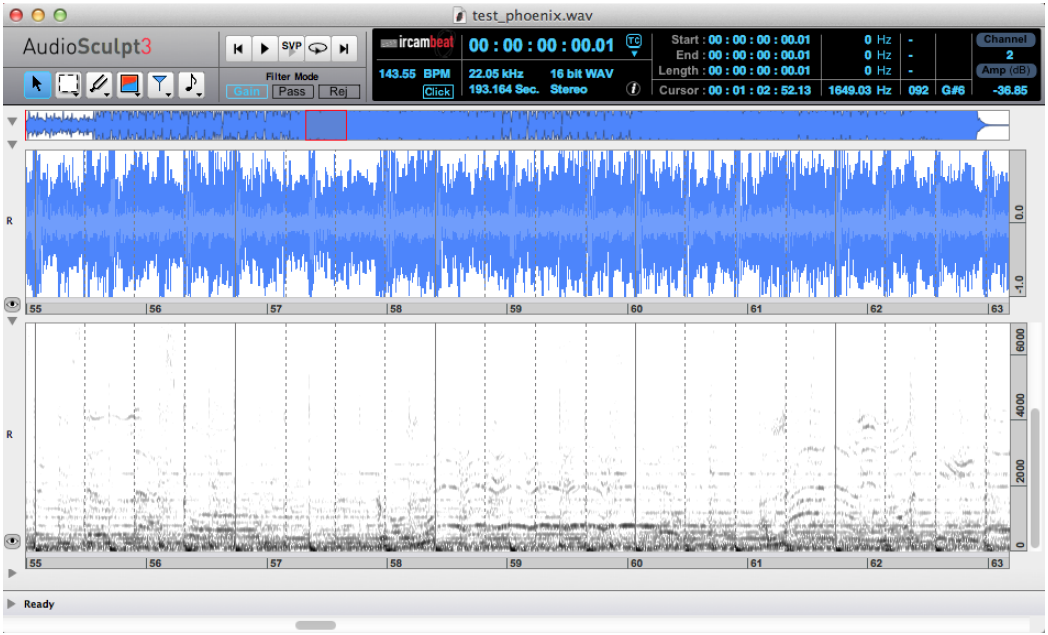
1. Introduction
1.1 Applications des techniques d'indexation audio pour la musique
2. Théorie : Traitement du signal
2.1 Transformée de Fourier (temps et fréquences continus)
2.2 Transformée de Fourier (temps et fréquences discrets)
2.3 Transformée de Fourier (à Court Terme) : TFCT
2.4 Transformée à Q-Constant (CQT)
3. Descripteurs Audio
3.1 Mel Frequency Cepstral Coefficients (MFCCs)
3.2 Chroma - Pitch Class Profile (PCP)
4. Applications
4.1 Identification audio
4.2 Estimation du tempo
4.3 Estimation de la structure musicale
5. Séparation de sources
5.1 Séparation de source par décomposition en partie répétée (rang faible) / partie sparse
5.2 Factorisation (décomposition) en matrices non-négatives
6. Transformation du signal
6.1 Application : dilatation/ contraction du temps par vocodeur de phase
6.2 Transformation du signal par la méthode P-SOLA

source : Gaël Richard

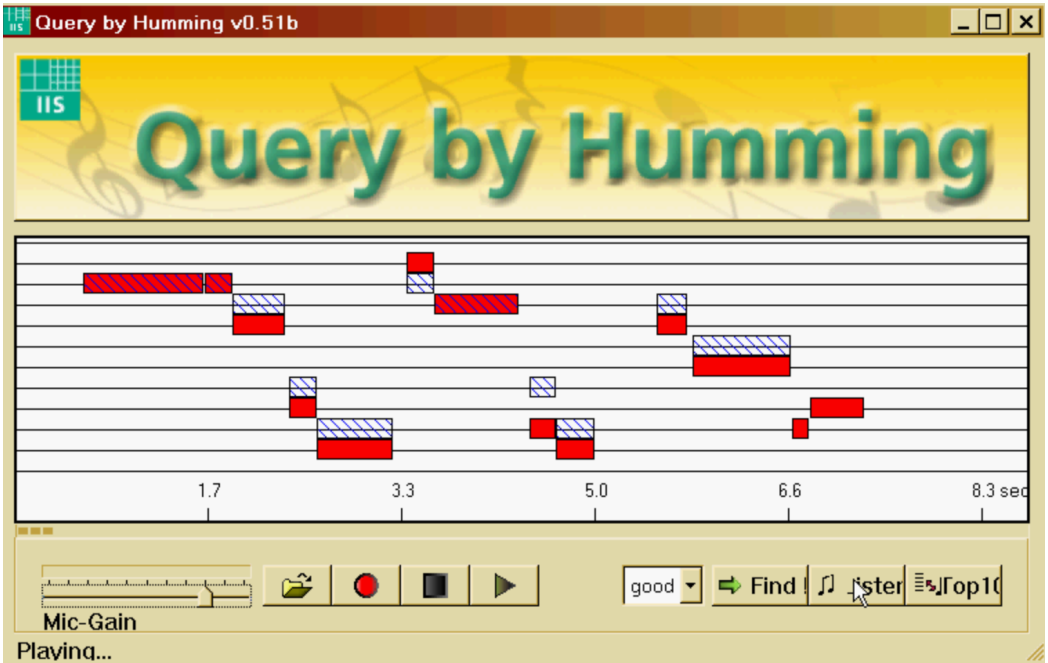
- Identification audio
  - recherche de doublons, gestion de copyright, attacher des méta données à une instance d'un morceau



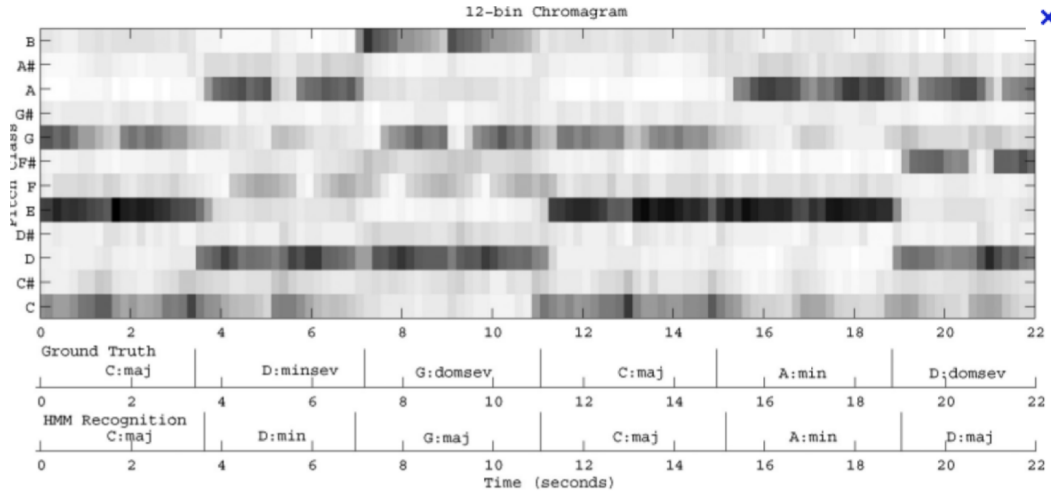
- Estimation du tempo, de la position des temps/ premier-temps
  - DJing, mainpulation du contenu (add swing ...)



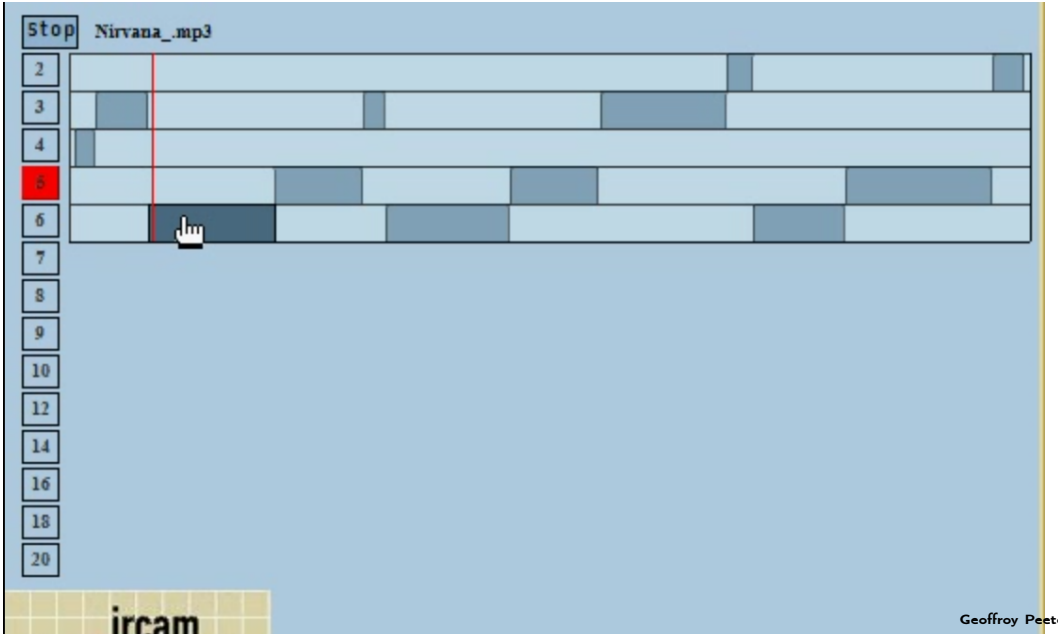
- Nouveaux modes de recherche :
  - par chantonnement/ sifflement



- Estimation des accords
  - obtenir des guitar-tab automatiquement



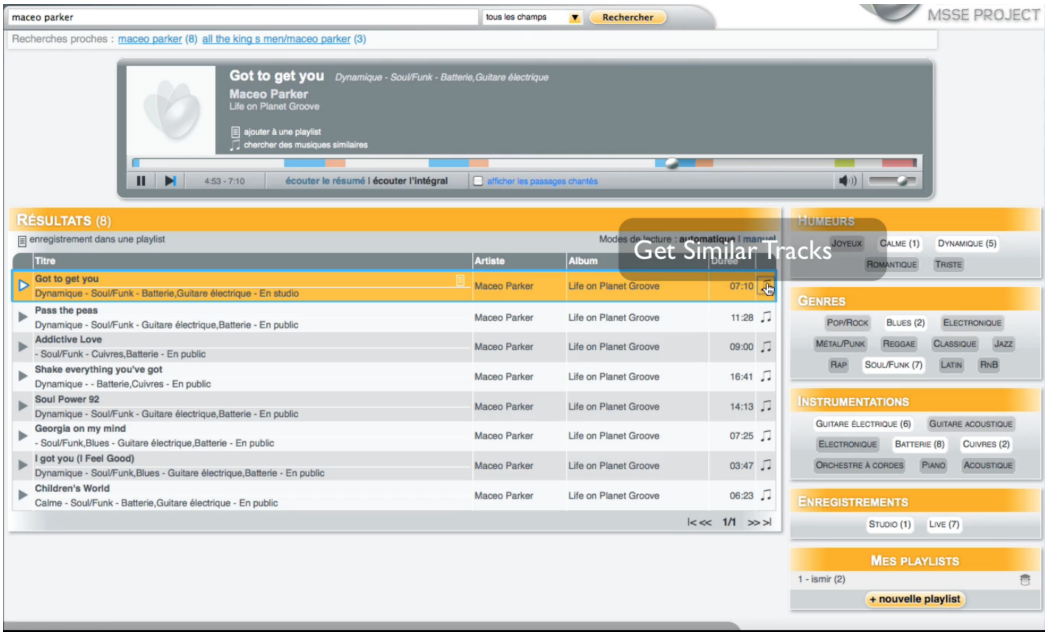
- Navigation à l'intérieur d'un morceau de musique par couplet/refrain
  - Génération automatique de résumé audio
- Dé-linéarisation d'un flux audio :
  - segmentation de flux radio, télé et étiquetage des parties



- Détection des cover, reprises ou ... des plagias

Titre	Artiste	Album	D.	Pop.
Let It Be	The Beatles Recovered Band	30 Beatles Top Hits	03:50	
Let It Be	The Hit Co., The Tribute Co.	A Tribute to the Beatles: The Lat...	03:42	
Let It Be	Labrinth	Let It Be	03:05	
Let It Be Me	Ray LaMontagne	Gossip in The Grain	04:41	
Let It Be – The Beatles Tribute	Let It Be	Let It Be – The Beatles Tribute	03:49	
Let It Be	Lois	Let It Be – The Voice 2	03:15	
Let It Be	The Yesteryears	A Tribute to #1 Beatles Hits – T...	03:48	
Let It Be	Aretha Franklin	This Girl's In Love With You	03:33	
Let It Be Sung	Jack Johnson, Matt Costa, Zach Gill,...	If I Had Eyes	04:09	
Let It Be	Vox Angeli	Gloria	03:26	
Let It Be	Paul McCartney	Good Evening New York City	03:54	
Hey Jude	Let It Be	Hey Jude	03:55	
Let It Be	Joan Baez	Greatest Hits And Others	03:51	

- Recherche d'un contenu audio dans une base de données
  - autrement que par "artistes", "titres" (Google musical)



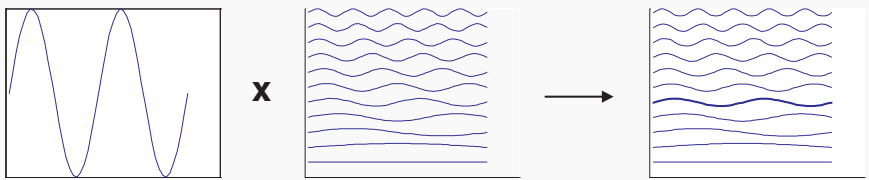
2- Théorie : Traitement du signal

2.1- Transformée de Fourier (temps et fréquences continus)

Transformée de Fourier (temps et fréquences continus)

$$X(\omega) = \int_{t=-\inf}^{+\inf} x(t)e^{-j\omega t}dt \quad X(f) = \int_{t=-\inf}^{+\inf} \exp(-j2\pi ft)dt$$

- Variables :
  - t est le **temps**
  - $\omega = 2\pi f$ les **fréquences continues** exprimées en radian,
  - $\exp(j2\pi ft) = \cos(2\pi ft) + j \cdot \sin(2\pi ft)$.
- Pourquoi la Transformée de Fourier ?
  - Difficile d'extraire des observations directement à partir de la forme d'onde $x(t)$
  - Reproduire la décomposition en fréquences de l'oreille humaine



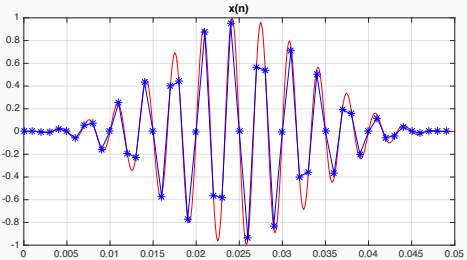
Propriété de la Transformée de Fourier (temps et fréquences continus)

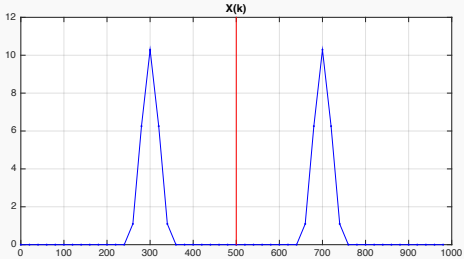
Propriétés	$x(t)$	$X(f)$
Similitude	$x(at)$	$\frac{1}{\|a\|}X(\frac{f}{\|a\|})$
Linéarité	$ax(t) + by(t)$	$aX(f) + bY(f)$
Translation	$x(t - t_0)$	$X(f)\exp(-j2\pi ft_0)$
Modulation	$x(t)\exp(j2\pi f_0 t)$	$X(f - f0)$
Convolution	$x(t) \circledast y(t)$	$X(f)Y(f)$
Produit	$x(t)y(t)$	$X(f) \circledast Y(f)$
Parité	réelle paire	réelle paire
	réelle impaire	imaginaire paire
	imaginaire paire	imaginaire paire
	imaginaire impaire	réelle impaire
	complexe paire	complexe paire
	complexe impaire	complexe impaire
	réelle	$X(f) = X^\star(-f)$
		$\Re(X(f))$ est paire
		$\Im(X(f))$ est impaire
	$x^\star(t)$	$X^\star(f)$

$$X(k) = \sum_{m=0}^{N-1} x(m) e^{-j2\pi \frac{k}{N} m} \quad \forall k \in [0, N]$$

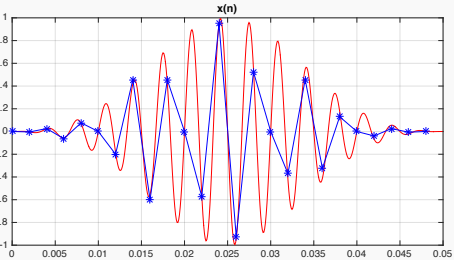
- Variables :
  - m le numéro d'**échantillon**
  - k les **fréquences discrètes**
- Fréquence d'échantillonnage (sampling rate) sr
  - sr définit à quelle fréquence le signal temporel va être échantillonné
  - Exemple :
    - Compact Disc $sr = 44100$ Hz
    - La distance temporelle entre deux échantillons (le pas d'échantillonnage) est de $\Delta t = \frac{1}{44100} = 0.000023\ s$.
- sr doit être $>$ à deux fois la $f_{\max}$ présente dans le signal
  - Sinon : repliement spectral
    - exemple : captation d'une roue d'une voiture accélérant dans les films
  - **Fréquence de Nyquist** : $f_{Nyquist} = \frac{sr}{2} > f_{\max}$

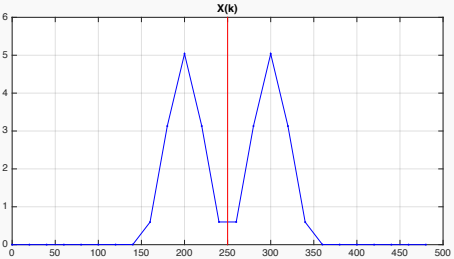
$f_{max} = 300$, $sr = 1000$





$f_{max} = 300$, $sr = 500$



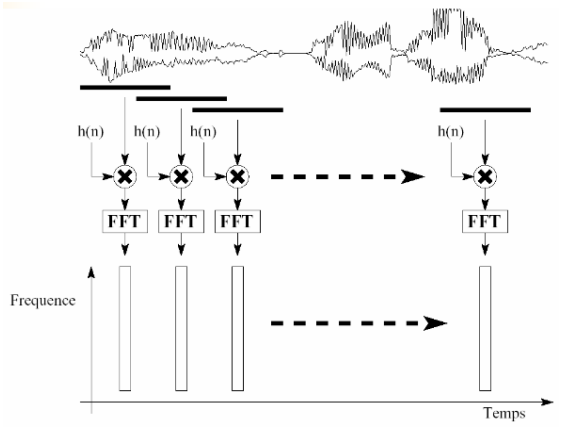


$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \qquad \forall k \in [0,N]$$

- Application de la TFD à une portion du signal centrée autour de l'échantillon n

Pourquoi la TFCT ?

- Signal audio = non-stationnaire
  - ses propriétés varient au cours du temps
- **Stationnaires "localement"** (en temps)
  - sur une durée de $\pm$ 40ms
- TFCT = suite d'analyses de Fourier sur des durées de $\pm$ 40ms
  - = analyse à Court Terme ("trames/frames" en vidéo)



source : Jean Laroche

$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \qquad \forall k \in [0,N]$$

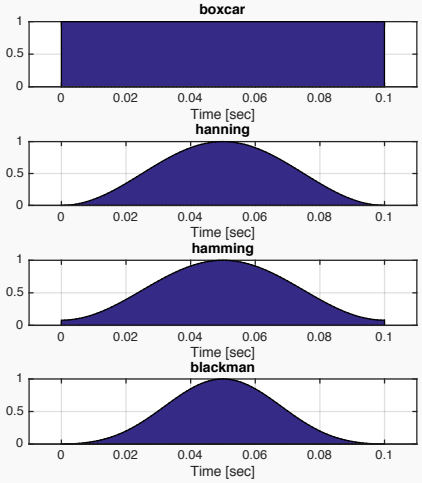
Fenêtre de pondération $w(t)$

- $x(t) \cdot w(t) \rightleftharpoons X(f) \circledast W(f)$
  - $w(t)$ est appelé **"fenêtre de pondération"**
  - $w(t)$ différents **types** de fenêtre
  - $w(t)$ définie sur un horizon fini (**longueur temporelle**) $[0, L]$.
  - Choix du type et de la longueur détermine les caractéristiques spectrales
    - Largeur de bande fréquentielle (à -6dB_{20}) : $Bw = \frac{C_w}{L}$
    - Hauteur des lobes secondaires

Choix du type de la fonction :

- rectangulaire
  - $w(n) = 1$
  - $Bw = 1.21$
- hanning
  - $w(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1}))$
  - $Bw = 2$
- hamming
  - $w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$
  - $Bw = 1.81$
- blackman
  - $w(n) = a_0 - a_1\cos(\frac{2\pi n}{N-1}) + a_2\cos(\frac{2\pi n}{N-1})$
  - $Bw = 2.35$

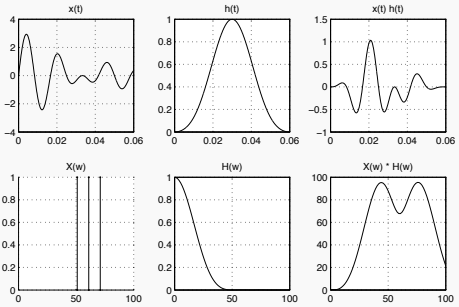
Influence du type de la fonction



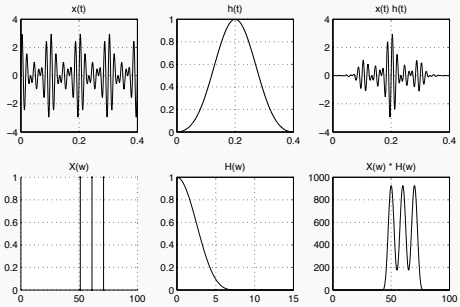
Choix de la longueur temporelle L :

- Au plus la fenêtre est courte,
  - au plus on observe précisément les temps.
- Au plus la fenêtre est longue,
  - au plus on observe précisément les fréquences.

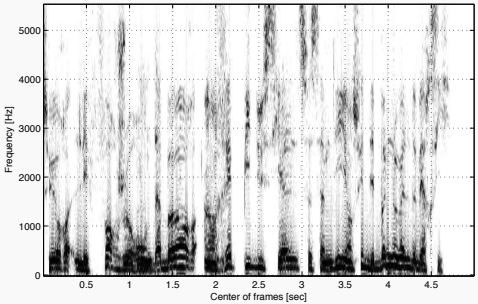
Influence de la longueur temporelle L ($L = 0.06$s.)



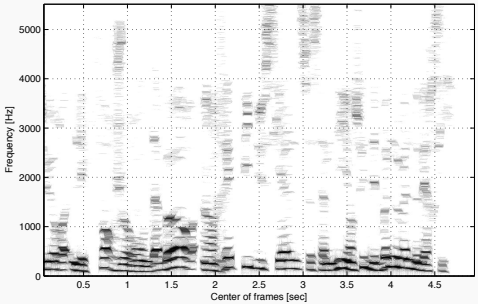
Influence de la longueur temporelle L ($L = 0.4$s.)



Influence de la longueur temporelle L ($L = 0.01$s.)

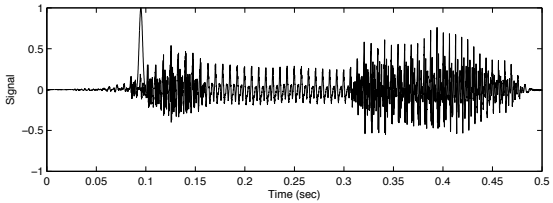


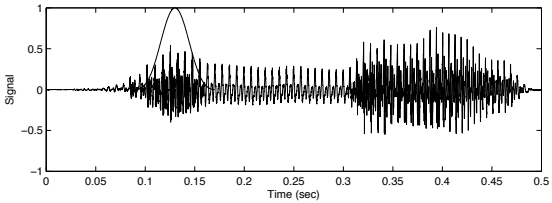
Influence de la longueur temporelle L ($L = 0.1$s.)

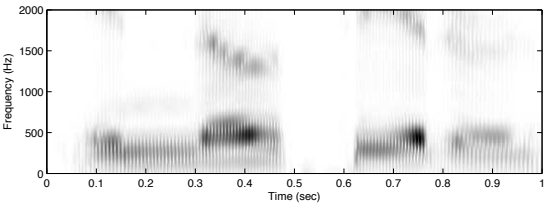


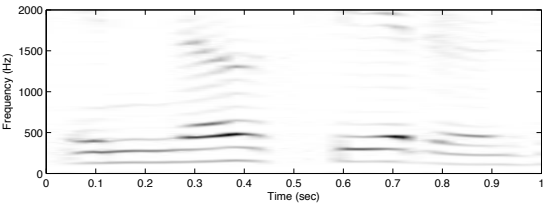
Paradoxe temps/ fréquence

- Pas possible d'avoir simultanément une bonne localisation en temps et en fréquence !









- Comme résoudre ce problème ?
  - Utiliser d'autres tranformées que celle de Fourier

2- Théorie : Traitement du signal

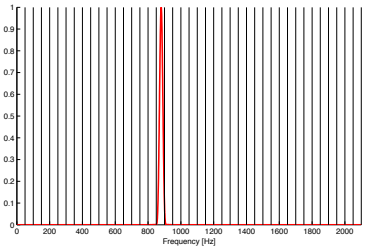
2.4- Transformée à Q-Constant (CQT)

Transformée à Q-Constant (CQT)

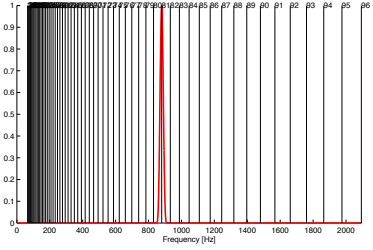
- La DFT
  - Définition : **La précision fréquentielle** : $\Delta f = \frac{sr}{N}$
    - c'est le pas d'échantillonnage du spectre
    - elle dépend de la taille de la DFT : N
    - on peut l'augmenter en augmentant N
  - Définition : **La résolution fréquentielle** : $Bw = \frac{Cw}{L}$
    - c'est le pouvoir de séparation entre deux fréquences présentes simultanément dans le spectre, le pouvoir de résoudre spectralement
  - Attention :
    - même si on augmente N (zero-padding) en gardant L constant on n'améliore pas la résolution !
- Dans la DFT, la précision et la résolution fréquentielle sont constantes à travers les fréquences

Transformée à Q-Constant (CQT)

- En audio musical
  - les fréquences sont logarithimiquement espacées
    - pour passer des fréquences aux hauteurs de notes : $m_k = 12 \cdot log_2 \frac{f_k}{440} + 69$
    - pour passer des hauteurs de notes aux fréquences : $f = 440 \cdot 2^{\frac{m-69}{12}}$
  - les hauteurs de notes sont plus rapprochées en basses fréquences, plus espacées en hautes fréquences
- La **résolution fréquentielle** de la DFT
  - n'est pas suffisante pour résoudre les hauteurs de notes adjacentes en basses fréquences,
  - est trop importante en hautes fréquences



Espacement linéaire de la DFT



Espacement logarithmique des hauteurs de notes

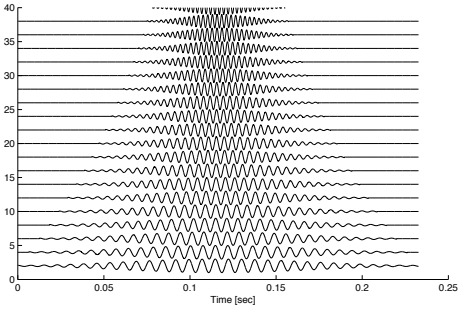
Transformée à Q-Constant

[J. Brown and M. Puckette. An efficient algorithm for the calculation of a constant q transform. JASA, 1992.]

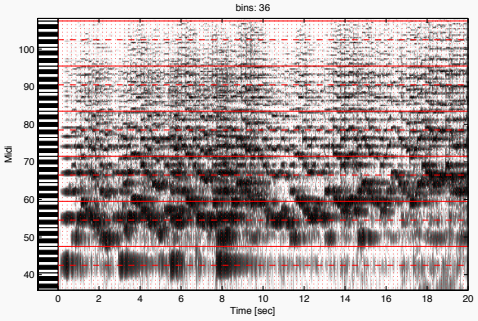
- Solution ?
  - Changer la **résolution fréquentielle** en fonction des fréquences considérées
- Comment ?
  - En changeant la longueur temporelle de la fenêtre pour chaque fréquence considérée
  - Le facteur $Q = \frac{f_k}{f_{k+1} - f_k}$ doit rester constant en fréquence

$$Q = \frac{f_k}{Bw} = \frac{f_k}{Cw/L} = \frac{f_k \cdot L}{Cw}$$

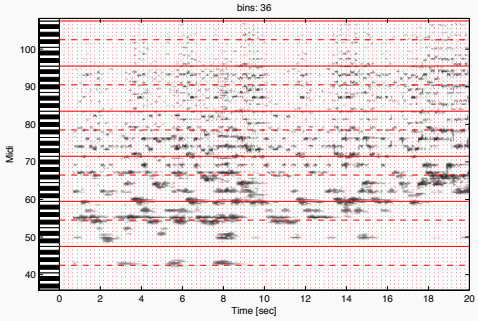
  - on choisit un L pour chaque fréquence f_k
    - $L_k = \frac{Q \cdot Cw}{f_k}$



Exemples (en utilisant la DFT)

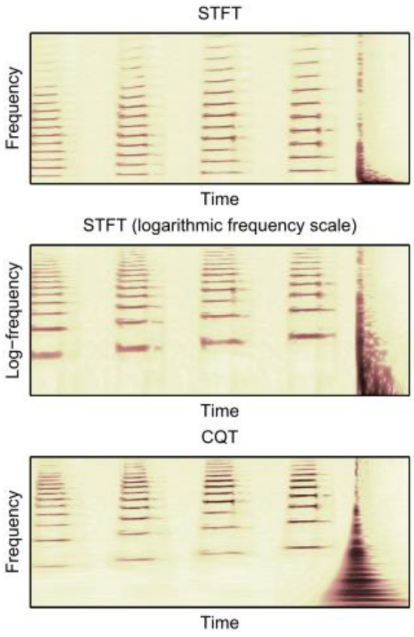


Exemples (en utilisant la CQT)



Transformée à Q-Constant (CQT)

- Sur une transformée à Q constant :
  - Une différence de pitch correspond à une translation sur l'axe des fréquences



source : Richard, 2012

4- Applications

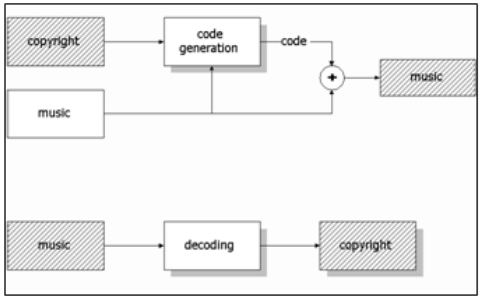
4.1- Identification audio

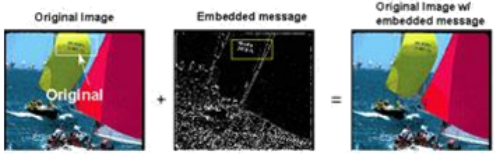
Identification audio

- Objectif :
  - Reconnaître un morceau diffusées sur radio, télé, Internet, bar, discothèque, ...
  - Identifier l'enregistrement (ISRC), pas l'oeuvre (ISWC)

Méthode du Watermarking

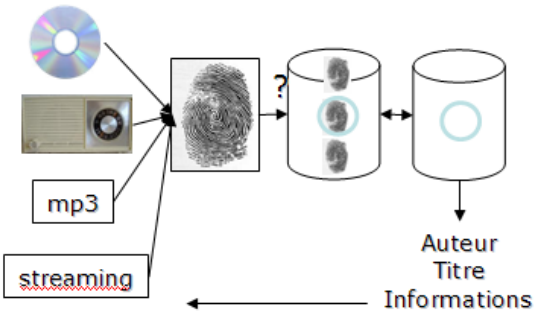
- Codage :
  - introduction d'un code identifiant robuste mais inaudible dans le signal sonore
- Décodage :
  - pour un nouveau signal : extraction du code (si il est présent) et recherche de ce code dans une base de données





Méthode du Fingerprint

- Shazam, Midomi, Philips, ...
- Codage :
  - prise d'empreinte du signal, stockage dans une base de données
- Décodage :
  - pour un nouveau signal, prise d'empreinte, comparaison avec les empreintes de la base de données
- Challenge :
  - déterminer un ensemble réduit de descripteurs audio extraits du signal sonore permettant d'identifier de manière unique un extrait musical

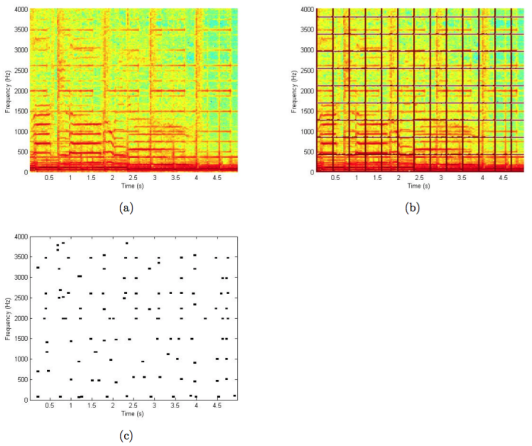


Algorithme de Fingerprint de Shazam

[A. L.-C. Wang. An industrial strength audio search algorithm. In Proc. of ISMIR, 2003.]

[S. Fenet. Empreintes Audio et Stratégies d'Indexation Associées pour l'Identification Audio à Grande Echelle. PhD thesis, Télécom Paris-Tech, 2013.]

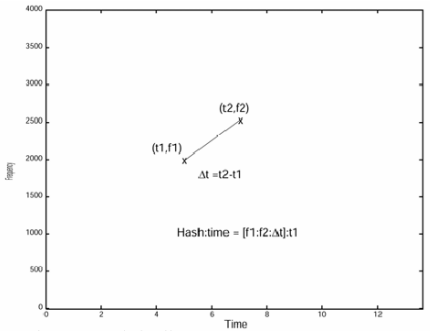
- Extraction de points saillants dans le plan temps/fréquence
  - Calcul du spectrogramme
    - fenêtre de Hamming, L=64 ms, S=32 ms
  - Dans chaque pavé du spectrogramme (Δt=0.4 s, Δf) :
    - détection du maximum → valeur = 1
  - = "constellation points"



source : Sebastien Fenet

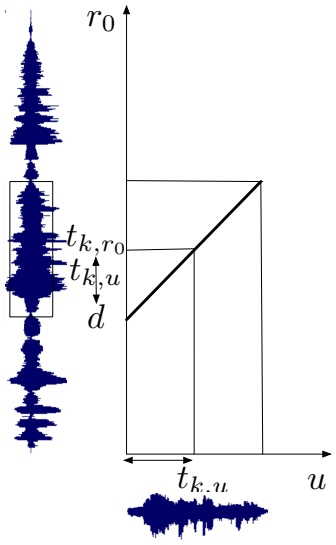
A) Partie stockage de signature

- Représentation des "constellation points" :
  - chaque point est pris comme un "anchor point" ayant une "target zone"
    - $[f_1, f_2, t_2 - t_1]$
    - $+$ le temps de l'anchor t_1
- Méthode de "pruning" des points
  - on ne garde que les pairs de points pour lesquels
    - $f_2 - f_1 < \Delta f_{\max} = 350Hz$
    - $t_2 - t_1 < \Delta T_{\max} = 3s$
- Stockage des triplets
  - $[f_1, f_2, t_2 - t_1]$ stocké sur 32 bits



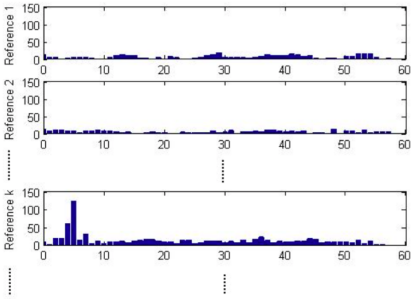
B) Partie matching de signature

- si le signal inconnu u est un extrait de r_0 démarrant au temps d
  - alors toutes les clefs apparaissant dans u doivent être trouvées dans r_0
  - une clef k de u au temps $t_{k,u}$ doit être trouvé dans r_0 au temps $t_{k,r_0} = d + t_{k,u}$
  - si on étudie l'ensemble des valeurs $\{t_{k,r_0} - t_{k,u}\}$ pour toutes les clefs k de u, on doit avoir un maximum d'accumulation en d

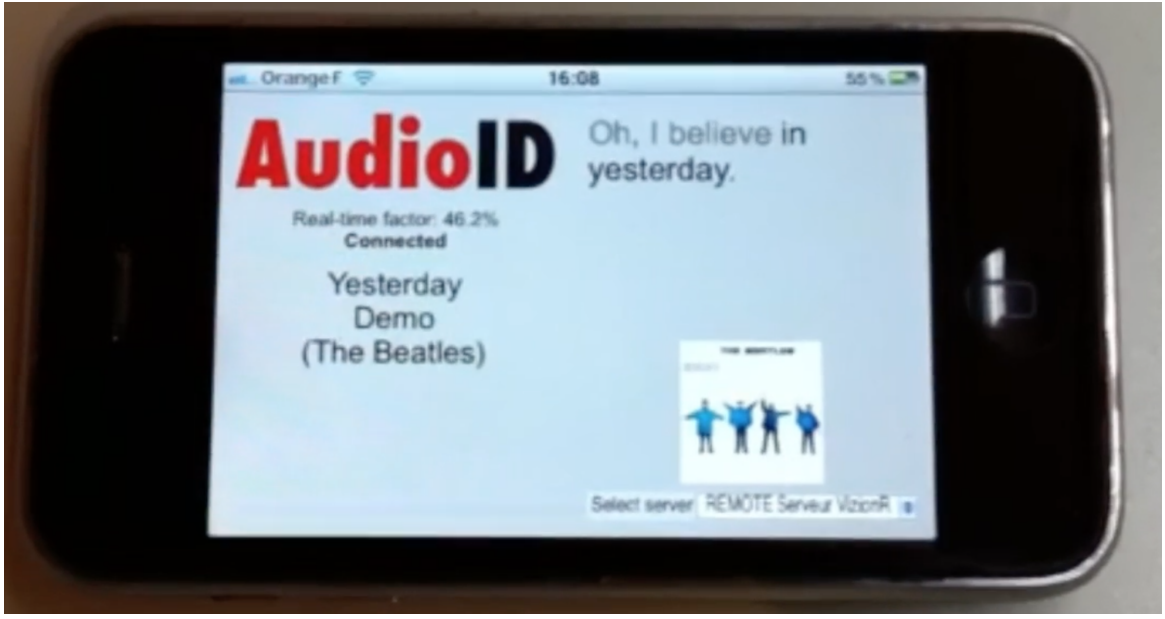


B) Partie matching de signature

- Méthode :
  - pour toutes les clefs k de u , pour chaque référence r_i, on stocke toutes les valeurs $\{t_{k,r_i} - t_{k,u}\}$ dans un histogramme
  - l'histogramme avec le plus grand maximum donne la référence du signal inconnu
  - la position du maximum dans cet histogram donne le point de démarrage d dans le signal de référence



source : Sebastien Fenet

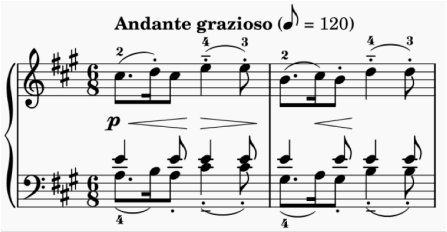

Orange F
16:08
55 %
AudioID
Oh, I believe in yesterday.
Real-time factor: 46.2%
Connected
Yesterday
Demo
(The Beatles)
Select server
REMOTE Serveur VizionR

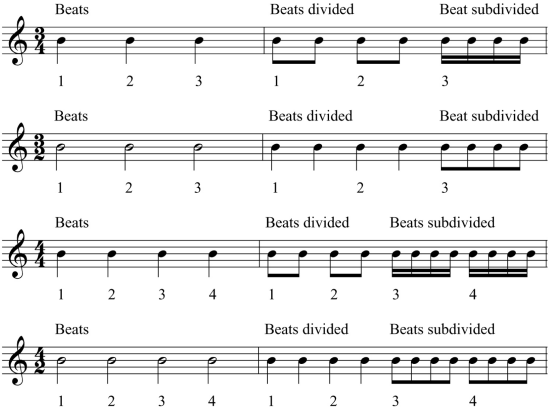
4- Applications

4.2- Estimation du tempo

Rhythme ?

- Tempo (beat)
  - indiquer sur une partition
  - "vitesse moyenne à laquelle les gens tapent du pied en écoutant la musique"
- Subdivision du rythme
  - mesure
    - entre deux barres, le groupement des noires
  - tactus
    - généralement la noire → le tempo
  - tatum
    - la vitesse la plus rapide
    - la subdivision de la noire en croches, triple-croches, double-croches



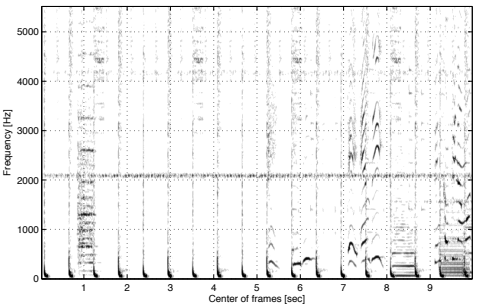


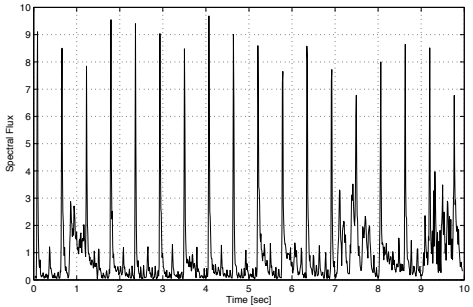
Estimation du tempo ?

- Détecter la périodicité des évènements dans le signal audio

Détection des évènements ?

- Début des évènements = onsets
- Méthode 1
  - détecter les maxima locaux de la fonction d'énergie du signal
    - $ener(m) = \sum_k X(k,m)^2$
- Méthode 2
  - détecter les maxima locaux du flux spectral :
    - $flux(m) = \sum_k \mathrm{HWR}[X(k,m) - X(k,m-1)]$
    - $\mathrm{HWR}(x) = x \quad si \quad x > 1$
    - $\mathrm{HWR}(x) = 0 \quad sinon$



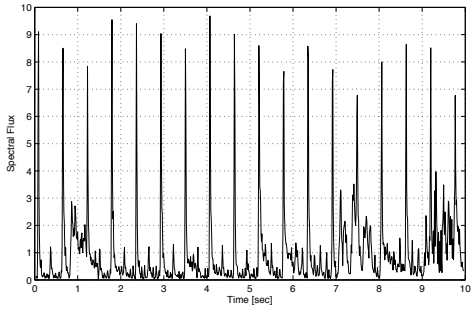


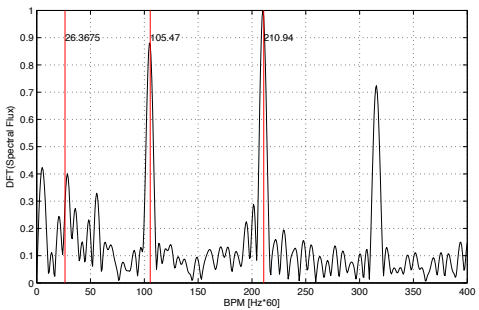
Périodicité des évènements ?

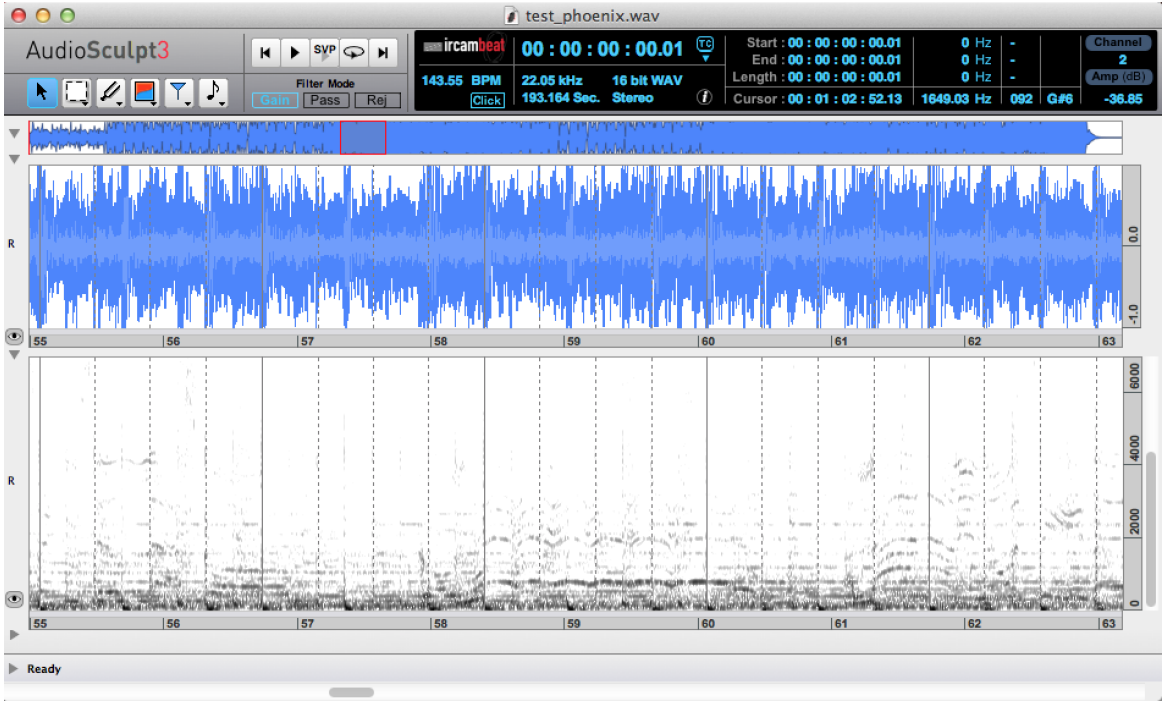
- Calcul de la transformée de Fourier (DFT) du flux spectral :
  - $FLUX(k) = \sum_{n=0}^{N-1} flux(n) \cdot \exp(j2\pi \frac{k}{N} n), \forall k$
- Calcul de la fonction d'auto-correlélation du flux spectral

Estimation du tempo ?

- Détecter le peak (fréquence f_k de la DFT) correspondant au tempo
  - Tempo= $60 f_k$ (BPM : Battement par Minute)
  - Peaks correspondant à la mesure, au tactus, au tatum






test_phoenix.wav
AudioSculpt3
SVP
Filter Mode
Gain
Pass
Rej
ircambeat
00 : 00 : 00 : 00.01
TC
143.55 BPM
Click
22.05 kHz
193.164 Sec.
16 bit WAV
Stereo
Start : 00 : 00 : 00 : 00.01
End : 00 : 00 : 00 : 00.01
Length : 00 : 00 : 00 : 00.01
Cursor : 00 : 01 : 02 : 52.13
0 Hz
0 Hz
0 Hz
1649.03 Hz
Channel
2
Amp (dB)
092
G#6
-36.85
R
0.0
-1.0
55
56
57
58
59
60
61
62
63
6000
4000
2000
0
Ready

4- Applications

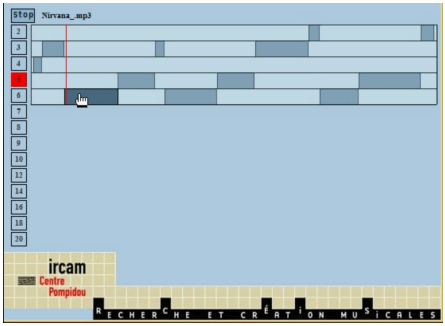
4.3- Estimation de la structure musicale

Structure musicale ?

- représentation du morceau comme une suite de parties (introduction, couplet, refrain, ...)

Estimation de la structure musicale ?

- recherche de parties temporelles homogènes et/ou répétées

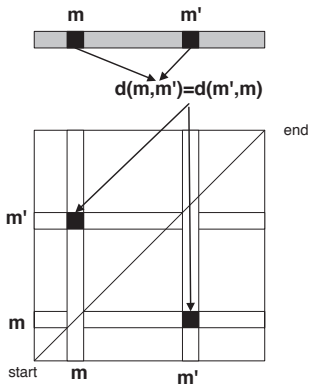


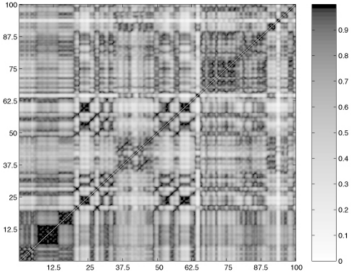
Matrice d'auto-similarité

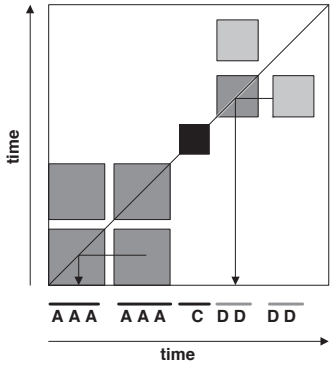
- représente la similarité entre les différents instants d'un même morceau

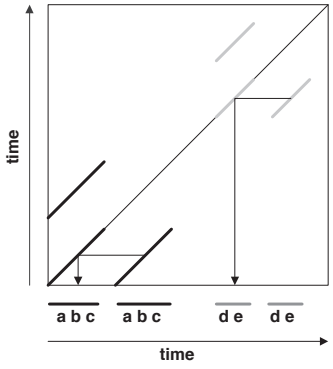
Méthode

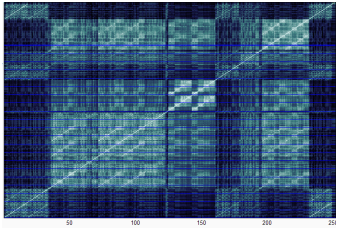
- On calcul la distance (euclidienne, cosinusoidale) entre les différents instants t d'un même morceau
- $d(t_1, t_2) = d(C(n, m_1), C(n, m_2) \quad \forall t_1, t2$
- On représente $d(t_1, t_2)$ sous forme d'une matrice (de similarité/ distance ou encore de co-occurrences)
- Utilisation :
  - localisation des couplets, refrain

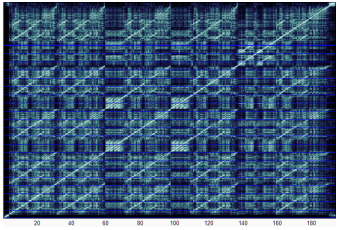


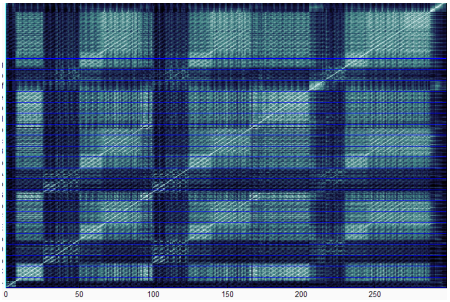


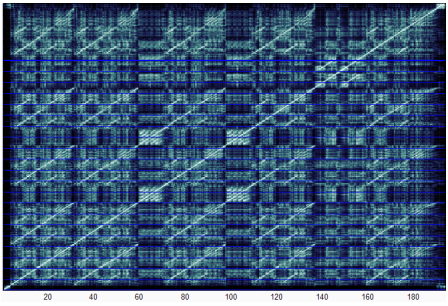

time
A A A
A A A
C
D D
D D
time

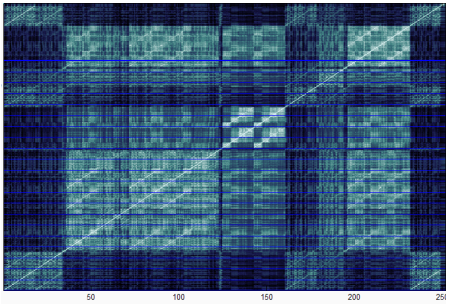

time
a b c
a b c
d e
d e
time

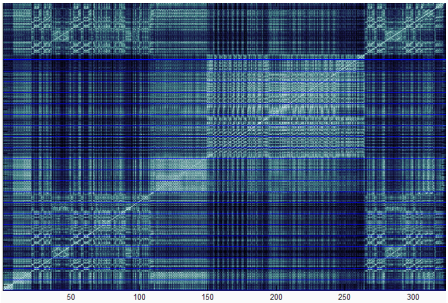

50
100
150
200
250

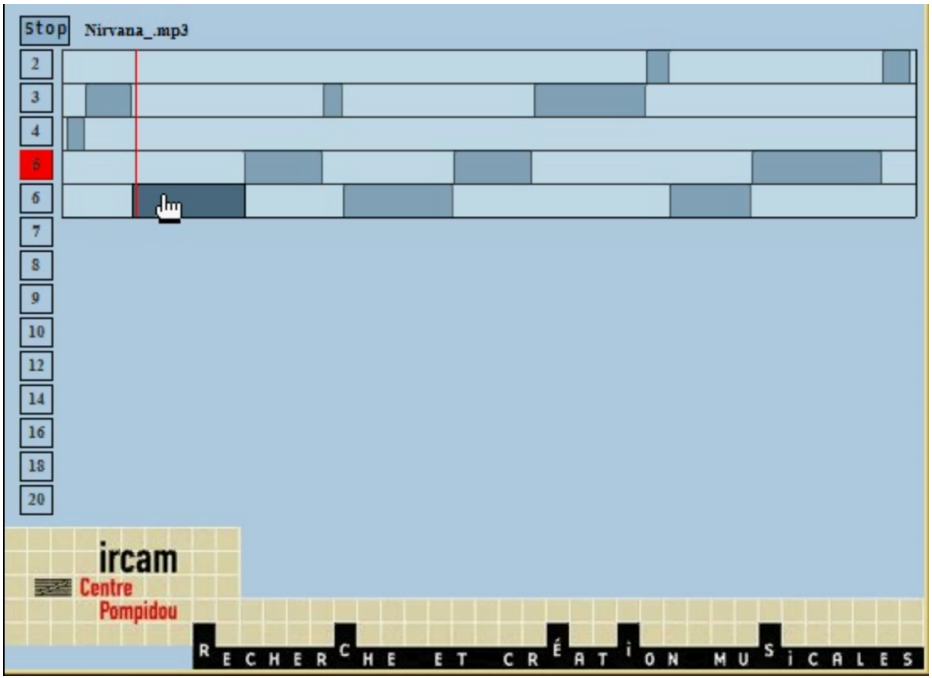

20
40
60
80
100
120
140
160
180


0
50
100
150
200
250


20
40
60
80
100
120
140
160
180


50
100
150
200
250


50
100
150
200
250
300


Stop
Nirvana_.mp3
2
3
4
5
6
7
8
9
10
12
14
16
18
20
ircam
Centre
Pompidou
RECHERCHE ET CRÉATION MUSICALES

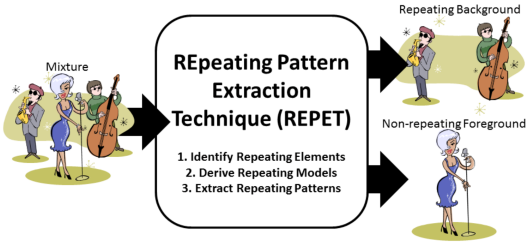
5- Séparation de sources

5.1- Séparation de source par décomposition en partie répétée (rang faible) / partie sparse

Introduction

[Zafar Rafii, Antoine Liutkus, and Bryan Pardo. "REPET for Background/Foreground Separation in Audio," Blind Source Separation, Springer, Berlin, Heidelberg, 2014]

- **REPET** : REpeating Pattern Extraction Technique (REPET)
- En musique, une pièce est souvent caractérisée par une structure sous-jacente répétitive par dessus laquelle des éléments non-répétés sont superposés
  - Il existe donc des patterns plus ou moins répétés en temps et en fréquence
  - Ces patterns répétés peuvent être identifiés en utilisant un masque temps-fréquence
  - Le masque T/F peut ensuite être appliqué pour extraire les patterns répétés



source : Rafii

REpeating Pattern Extraction Technique (REPET)

- Identifier les éléments répétés
- Dérivé un modèle de répétition
- Extraite la structure répétée

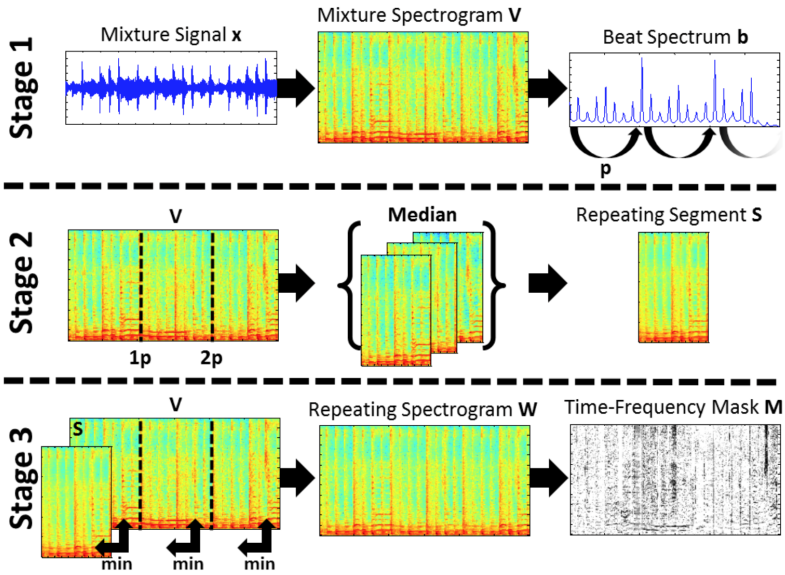
Méthode simple de séparation musique - voix

- $X_{mix} = X_{background} + X_{foreground}$
- $X_{mix} = X_{repeat} + X_{non-repeat}$
- $X_{mix} = X_{dense,low-rank} + X_{sparse}$
- Structure répétée = accompagnement musical
- Structure non-répétée = voix

Hypothèses

- L'arrière plan répété est dense et de rang-faible
  - Souvent vrai pour la voix dans un mélange musique + voix

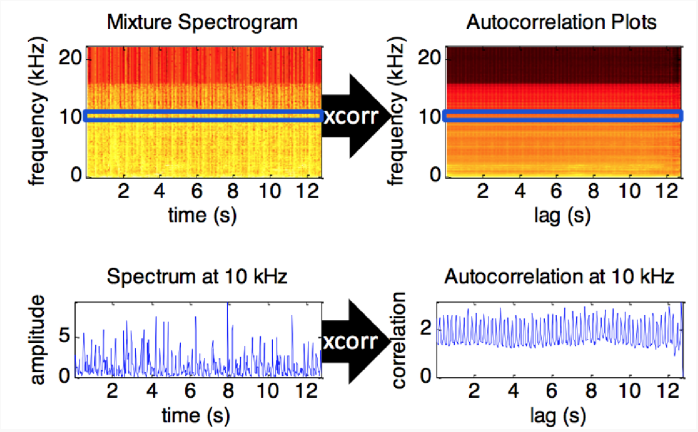
Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

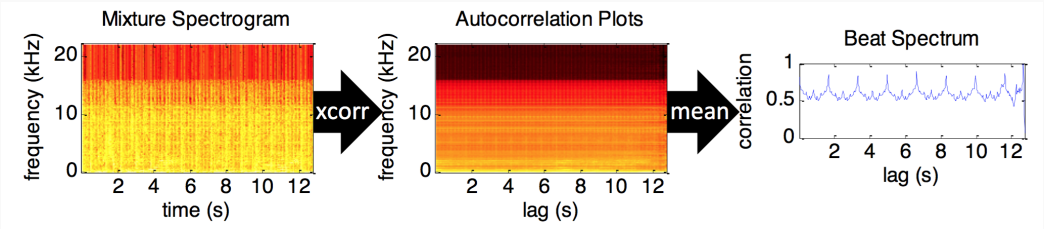
- **1. Détecter la période de répétition**
  - On calcul l'auto-corréléation de chaque ligne du spectrogramme X_{mix}



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

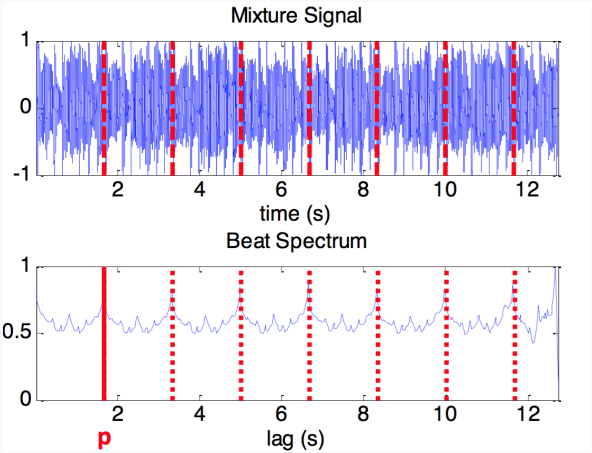
- **1. Détecter la période de répétition**
  - On prend la moyenne des fonctions d'auto-corrélation pour obtenir le "beat spectrum"



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

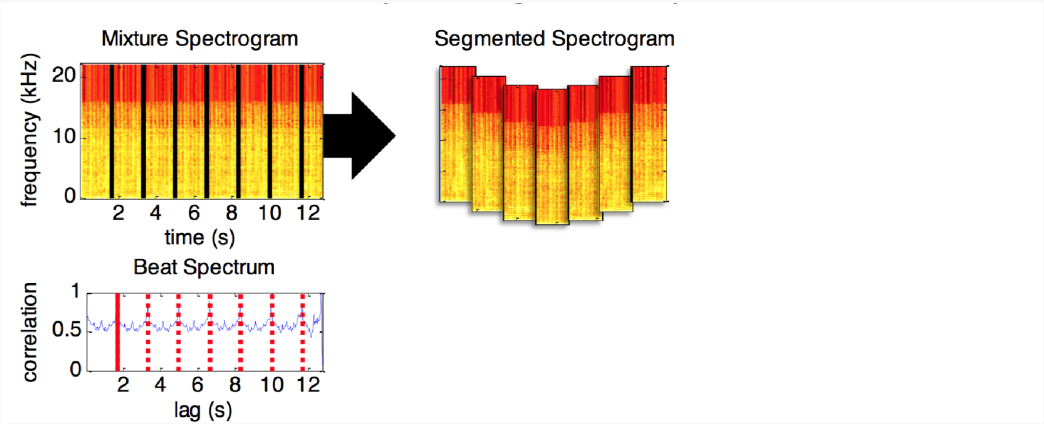
- **1. Détecter la période de répétition**
  - Le "beat spectrum" révèle la période de répétition p de la structure sous-jacente répétée
  - On suppose ici que l'arrière-plan (accompagnement) est plus dense et de rang plus faible que l'avant-plan



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

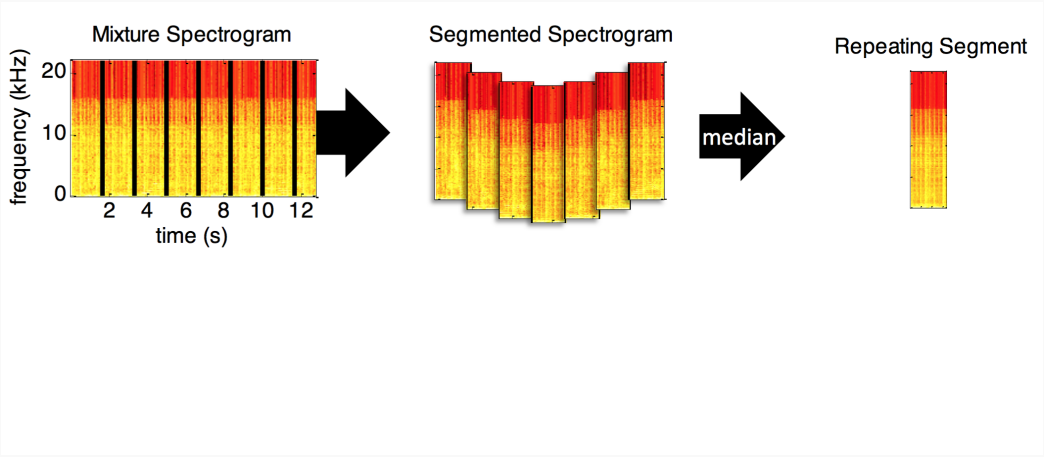
- **2. Segment répété**
  - La période de répétition est alors utilisée pour segmenter le spectrogramme du mélange à la vitesse de la période



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

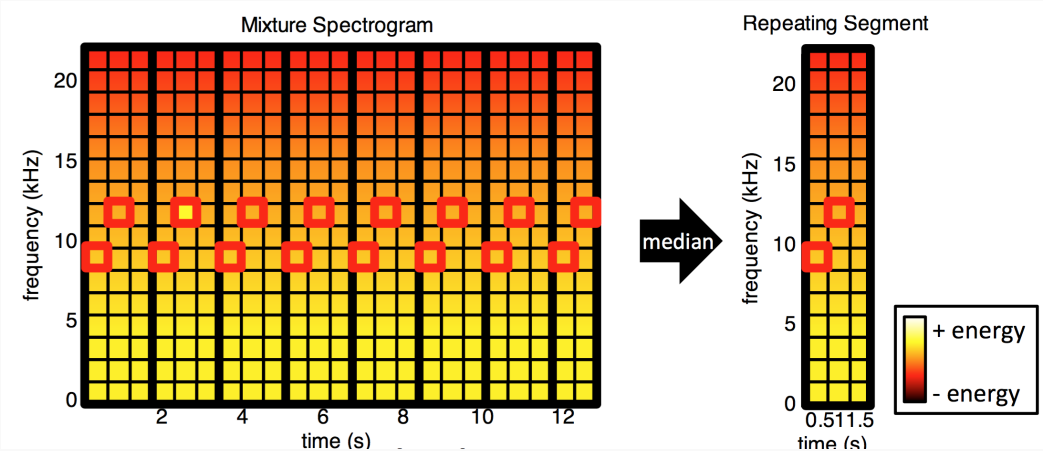
- **2. Segment répété**
  - Le segment répété est calculé comme la valeur médiane élément-à-élément des segments



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

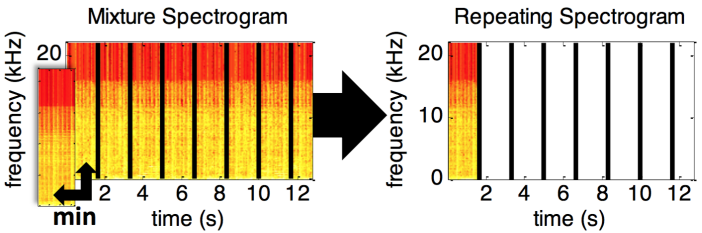
- **2. Segment répété**
  - La médiane aide à deriver un segment répété propre, en retirant les valeurs aberrantes non répétées
  - Nous supposons que l'avant-plan est plus "sparse" et varié que l'arrière plan



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

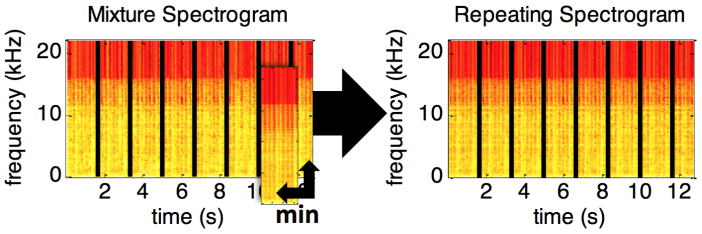
- **3. Structure répétée**
  - On prend la valeur minimum élément-à-élement entre le segment répété et les segments



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

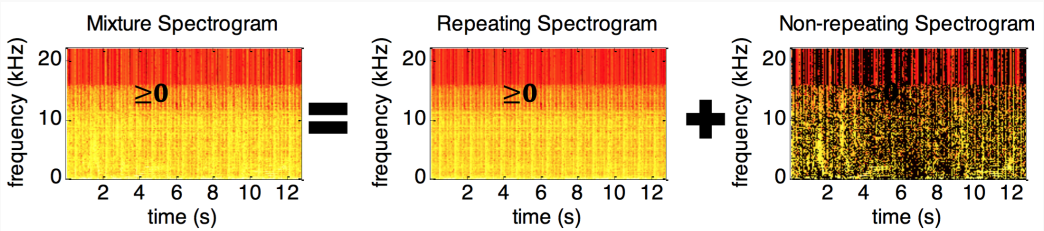
- **3. Structure répétée**
  - On obtient le modèle du spectrogram répété pour l'arrière-plan répété



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

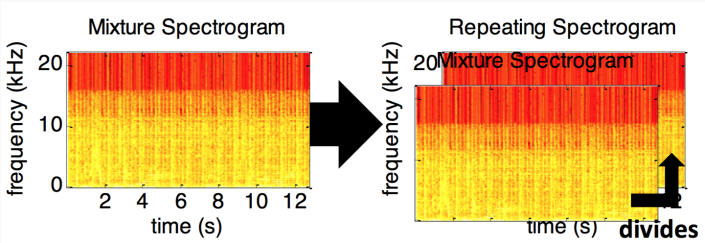
- **3. Structure répétée**
  - le spectrogram répété ne peut avoir des valeurs plus grandes que le spectrogramme mélangé



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

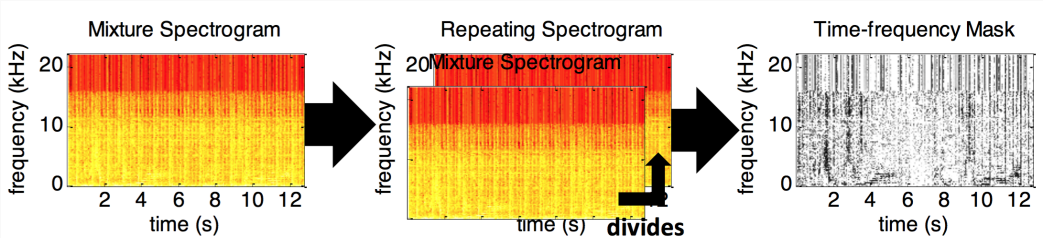
- **3. Structure répétée**
  - On divise élément-à-élement le spectrogramme répété par le spectrogramme mélangé



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

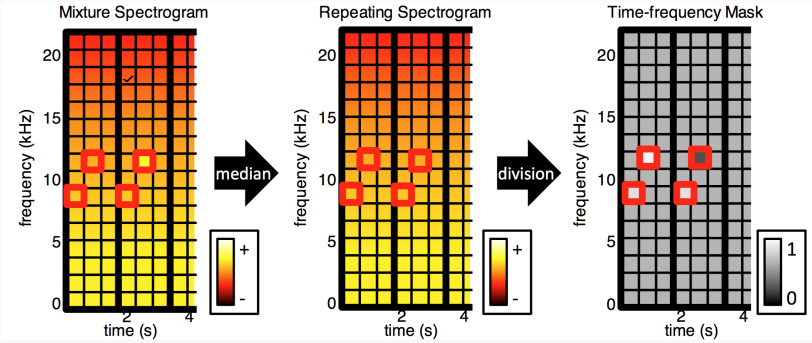
- **3. Structure répétée**
  - On obtient un "soft mask" (valeur dans [0,1])



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

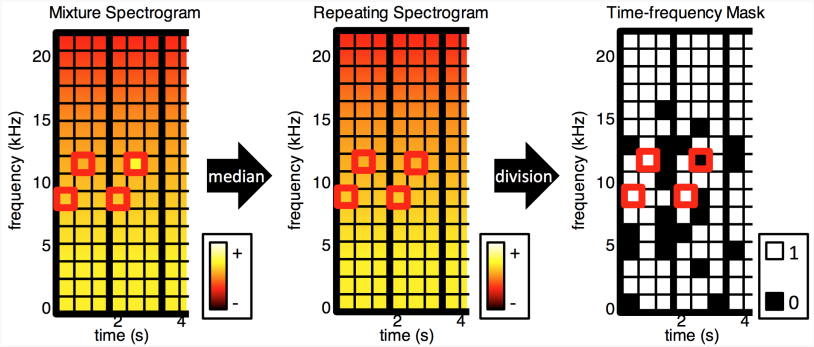
- **3. Structure répétée**
  - Dans le "soft mask",
    - au moins un bin T/F est répété au plus il tend vers 0
    - au plus un bin T/F est répété au plus il tend vers 1



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

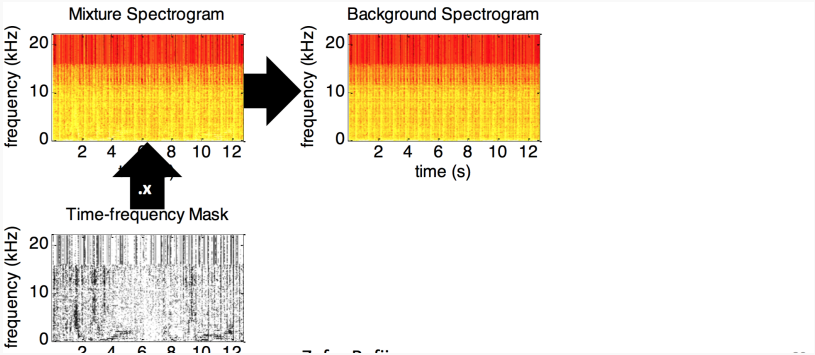
- **3. Structure répétée**
  - un "binary mask" peut également être dérivé en appliquant un seuil entre 0 et 1



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

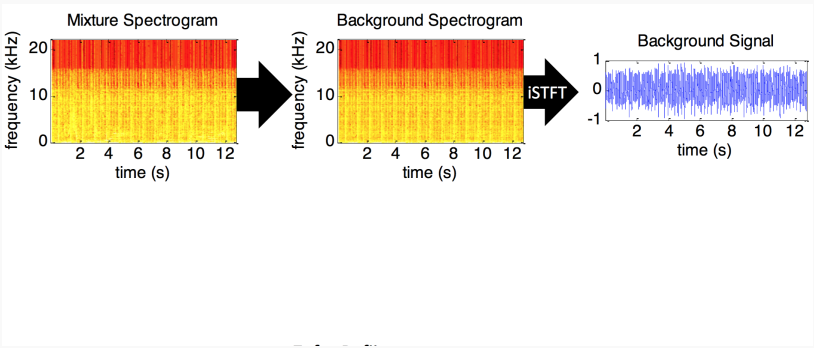
- **3. Structure répétée**
  - On multiplie le T/F mask avec la STFT du mélange pour extraire l'arrière-plan répété de la TFCT



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

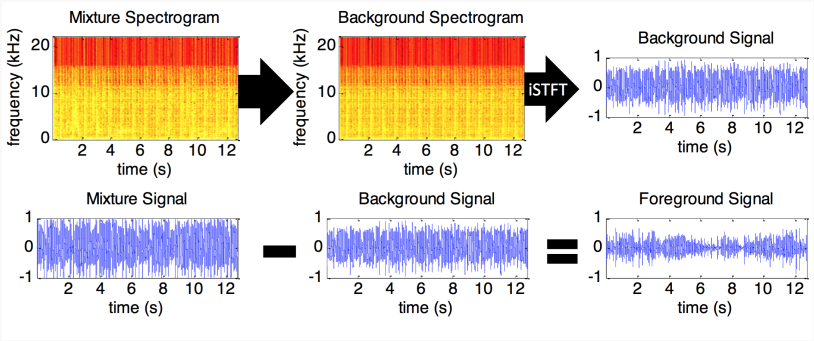
- **3. Structure répétée**
  - L'accompagnement répété est obtenu en inversant la TFCT dans le domaine temporel



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

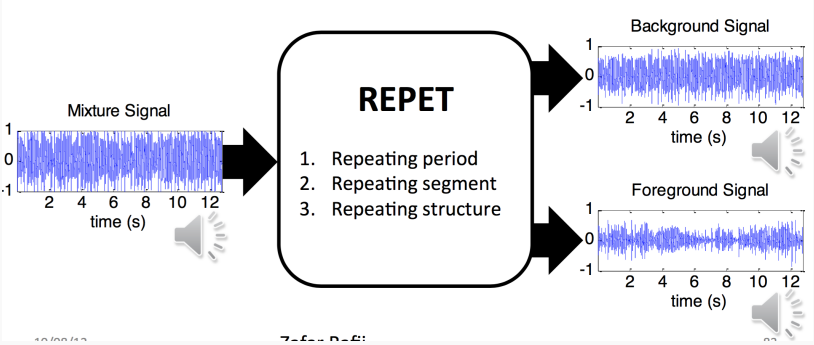
- **3. Structure répétée**
  - L'arrière-plan non répété est obtenu en soustrayant l'arrière-plan du mélange



source : Rafii

Comment identifier les répétitions ? 1) Estimer globalement le tempo et les battements du morceau : REPET

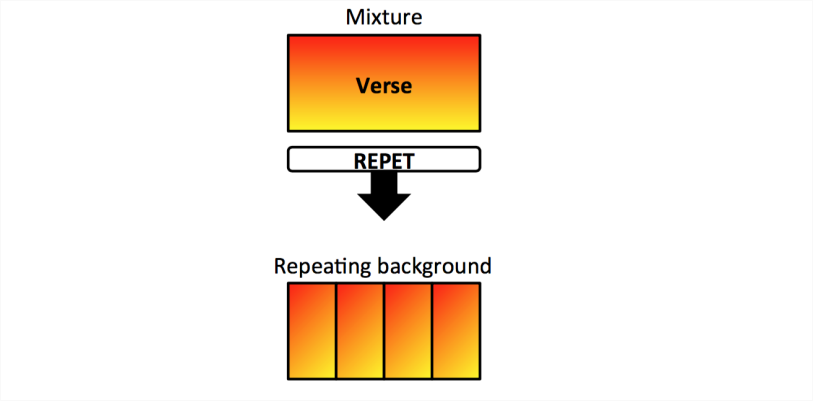
- **3. Structure répétée**
  - Arrière-plan répété = accompagnement musical
  - Avant-plan non-répété = voix



source : Rafii

Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET

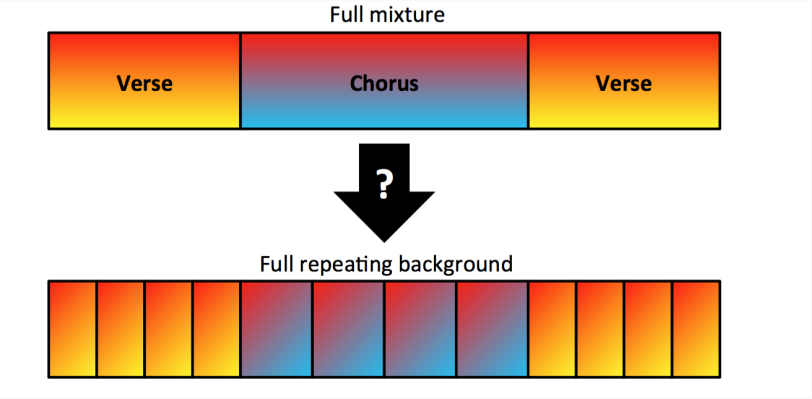
- REPET fonctionne bien sur des extraits qui ont un arrière-plan relativement stable (10 s du couplet)



source : Rafii

Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET

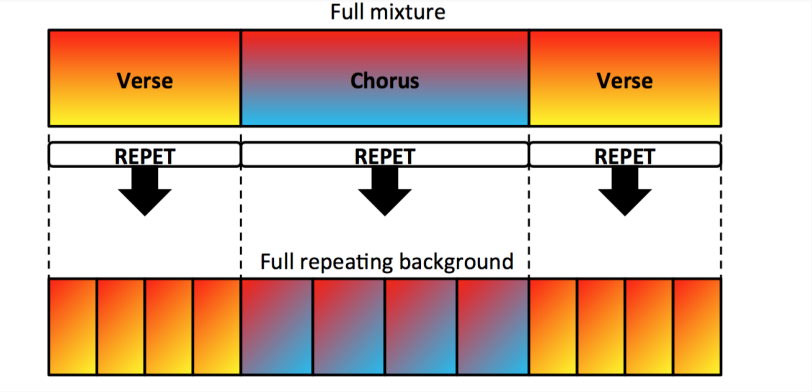
- Pour des morceaux entiers, l'arrière-plan répété est susceptible de changer au cours du temps (couplet/ chorus)



source : Rafii

Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET

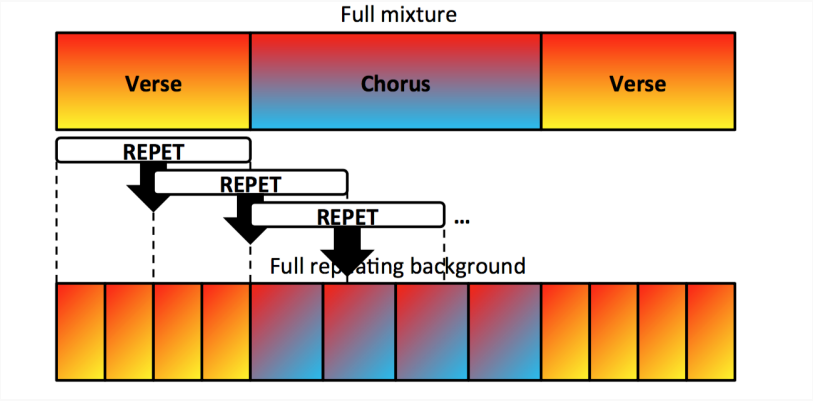
- On pourrait appliquer une segmentation du morceau et appliquer REPET sur chaque segment individuel



source : Rafii

Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET

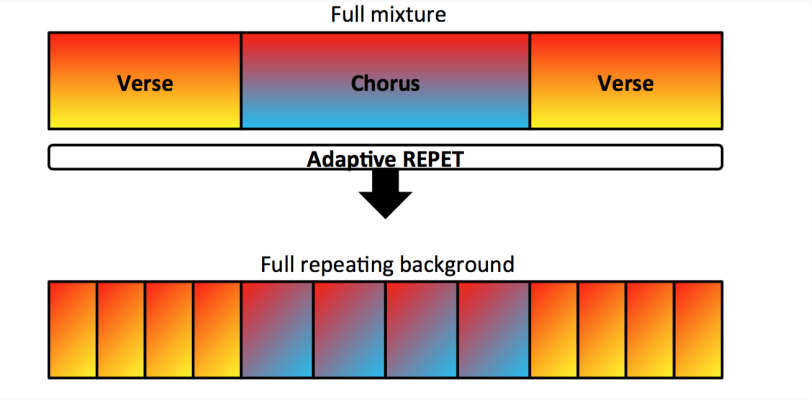
- On pourrait appliquer REPET aux segments locaux du morceau par utilisation d'une fenêtre glissante



source : Rafii

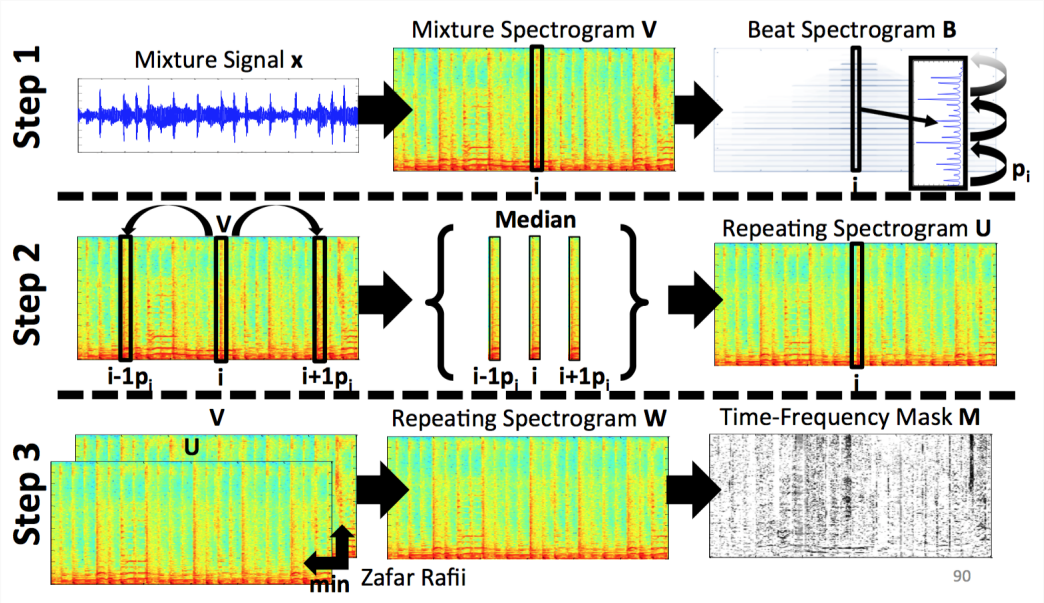
Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET

- On pourrait adapter REPET au cours du temps en modélisant l'arrière plan répété



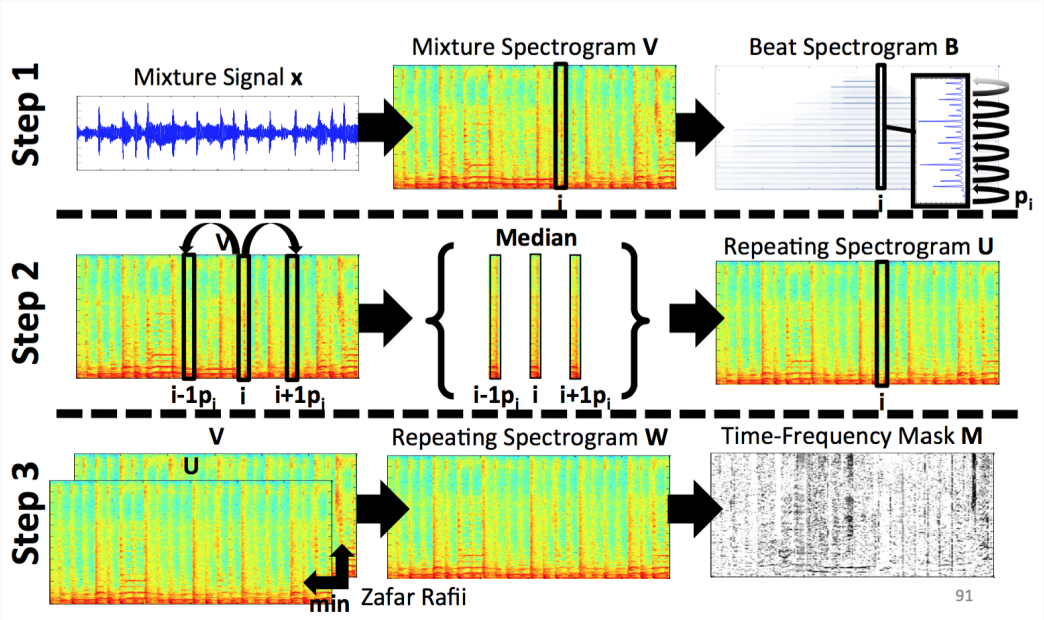
source : Rafii

Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET



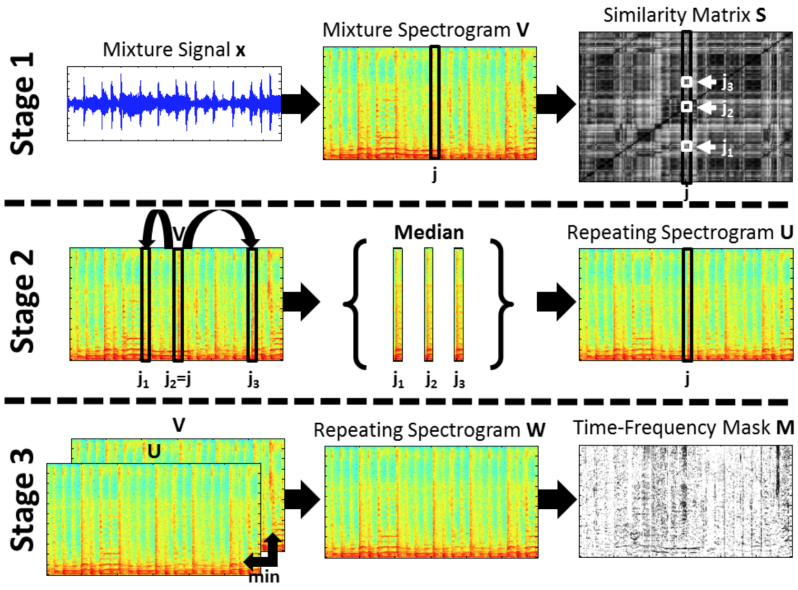
source : Rafii

Comment identifier les répétitions ? 2) Estimer localement le tempo et les battements du morceau : Adaptive REPET



source : Rafii

Comment identifier les répétitions ? 3) Esimer l'auto-similarité (matrice d'auto-similarité) du morceau : REPET-SIM



source : Rafii

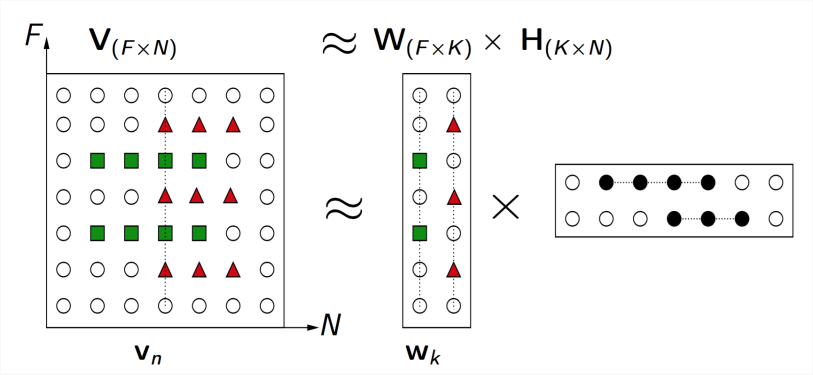
5- Séparation de sources

5.2- Factorisation (décomposition) en matrices non-négatives

Introduction

[D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 1999.]

- **NMF** : Non-Negative Matrix Factorization

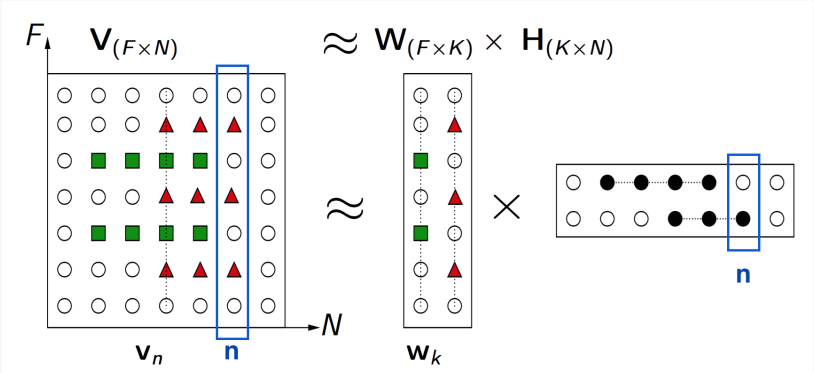


source : Cédric Févotte

- $V_{(F,N)} \simeq W_{(F,K)} H_{(K,N)}$
  - $V_{(F,N)}$: matrice de **données**, observée (spectrogramme d'énergie), définie positive : $V_{fn} \geqslant 0$
  - $W_{(F,K)}$: matrice de **bases**, dictionnaires, définie positive : $W_{fk} \geqslant 0$
  - $H_{(K,N)}$: matrice d'**activation**, définie positive : $H_{fn} \geqslant 0$
  - K : le nombre de bases du dictionnaire

Introduction

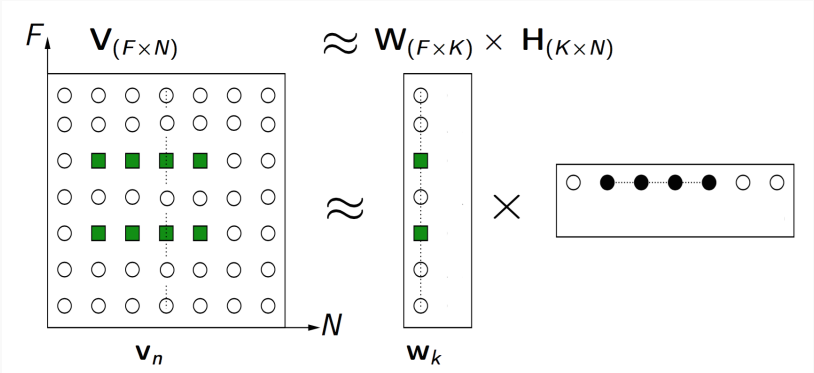
- Chaque trame **n** est reconstituée comme l'**activation** H d'un certain nombre de **bases** W
  - $V_{(1:F,\mathbf{n})} \simeq \sum_{k=1}^{K} W_{(1:F,k)} H_{(k,\mathbf{n})}$



source : Cédric Févotte

Introduction

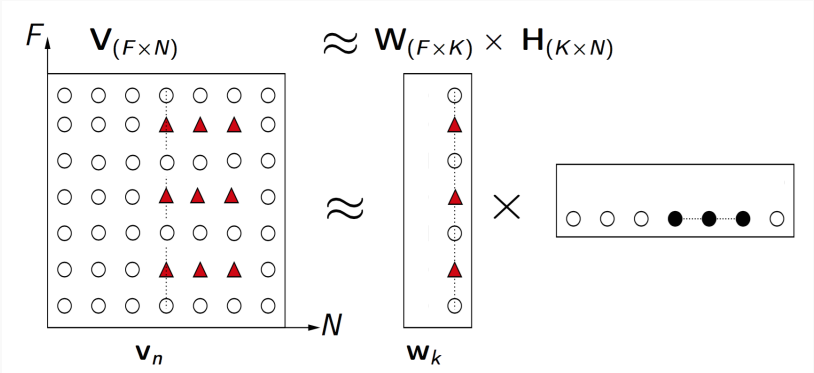
- Le signal d'une **source** k est reconstitué comme
  - $V^k_{(1:F,1:N)} = W_{(1:F,\mathbf{k=1})} H_{(\mathbf{k=1},1:N)}$



source : Cédric Févotte

Introduction

- Le signal d'une **source** k est reconstitué comme
  - $V^k_{(1:F,1:N)} = W_{(1:F,\mathbf{k=2})} H_{(\mathbf{k=2},1:N)}$



source : Cédric Févotte

Estimation des paramètres de la NMF

- $V_{(F,N)} \simeq W_{(F,K)} H_{(K,N)}$
- **Minimisation** de
  - $\min_{\underline{W},\underline{H} \geqslant 0} D(\underline{V}|\underline{WH})$
  - $\min_{\theta} C(\theta) \stackrel{def}{=} D(\underline{V}|\underline{WH})$ avec $\theta = \{W, H\}$
- D/d est une **divergence séparable**
  - $D(\underline{V}|\underline{\hat{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn})$
- Choix de D/d :
  - Distance Euclidenne :

$$d_{EUC}(x, y) = (x - y)^2$$

  - Divergence de Kullback-Leibler :

$$d_{KL}(x, y) = x \log \frac{x}{y} - x + y$$

  - Divergense d'Itakura-Saito :

$$d_{IS}(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

Dérivation du critère pour la distance Euclidenne

- Non Negative Matrix Factorization

$$\underset{(f,n)}{V} \simeq \underset{(f,k)}{W} \underset{(k,n)}{H}$$

- Erreur de reconstruction : $e = V - WH$
- Minimisation de la SSE (Sum of Squared Error) ou de la norme de Frobenius de $SSE = ||V - WH||_F^2$
- Norme de Frobenius : $||A||_F = \sqrt{\sum_i \sum_j a_{ij}^2}$

Dérivation du critère pour la distance Euclidenne

$$
\begin{aligned}
SSE &= ||V - WH||_F^2 \\
SSE &= (V - WH)^T (V - WH) \\
&= (V^T - H^T W^T)(V - WH) \\
&= V^T V - V^T WH - H^T W^T V + H^T W^T WH \\
&= V^T V - 2V^T WH + H^T W^T WH \\
\frac{\partial sse}{\partial H} &= -2W^T V + 2W^T WH \\
&= 2W^T (WH - V) \\
\frac{\partial sse}{\partial W} &= -2VH^T + 2WHH^T \\
&= -2(V - WH)H^T
\end{aligned}
$$

Propriétés utilisées (Matrix CookBook)

- $\frac{\partial a^T x}{\partial x} = a$
- $\frac{\partial a^T Xb}{\partial X} = ab^T$
- $\frac{\partial x^T Bx}{\partial x} = (B + B^T)x$
- $\frac{\partial b^T X^T Xc}{\partial X} = X(bc^T + cb^T)$

Algorithme de descente de gradient

- Descente de gradient ?
  - déplacement dans la direction opposée au gradient, de manière à faire décroître la fonction
- Le gradient : $\frac{\partial sse}{\partial H} = \underbrace{2W^T WH}_{\nabla_+} - \underbrace{2W^T V}_{\nabla_-}$
- Mise à jour de H

$$H \leftarrow H + \eta \cdot [-\text{gradient}]$$
$$H \leftarrow H + \eta \cdot [\underbrace{W^T V}_{\nabla_-} - \underbrace{W^T WH}_{\nabla_+}]$$

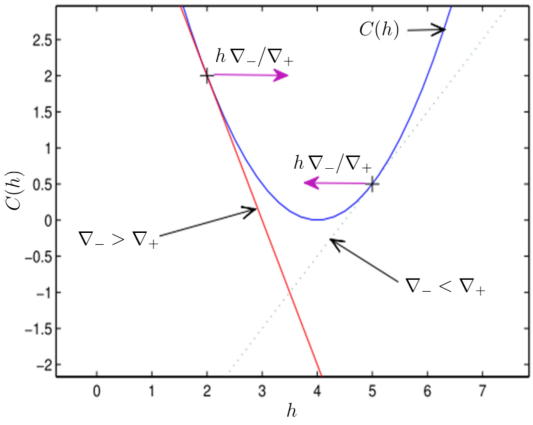
- si on choisit $\eta = \frac{H}{W^T WH}$

$$H \leftarrow H + \frac{H}{W^T WH}(W^T V - W^T WH)$$
$$H \leftarrow H + \frac{HW^T V}{W^T WH} - H$$
$$H \leftarrow H \cdot \frac{\overbrace{W^T V}^{\nabla_-}}{\underbrace{W^T WH}_{\nabla_+}}$$

Mise à jour multiplicative

- permet de garantir que les valeurs restent positives ! ! !
- Séparation du gradient en contribution **positive** et **négative**

$$\nabla_h C(h) = \nabla_+ - \nabla_-$$



Algorithme complet de NMF dans le cas Euclidéen : $\underset{(f,n)}{V} \simeq \underset{(f,k)}{W}\,\underset{(k,n)}{H}$

- Calcul de la TFCT : $V(f,n) = |X(f,n)|$
- Choix du nombre de bases K du dictionnaire W
- Initialisation de W et H : valeurs aléatoires positives
- Itérations
  - Mise à jour des bases W étant donné les activations H

$$W \leftarrow W \cdot \frac{VH^T}{WHH^T}$$

  - Mise à jour des activations H étant donné les bases W

$$H \leftarrow H \cdot \frac{W^T V}{W^T WH}$$

  - Prise en compte de l'invariance d'échelle
    - normalisations des colonnes de H
    - OU
    - normalisation des lignes de W
  - Arrêt lorsque la SSE cesse de décroitre

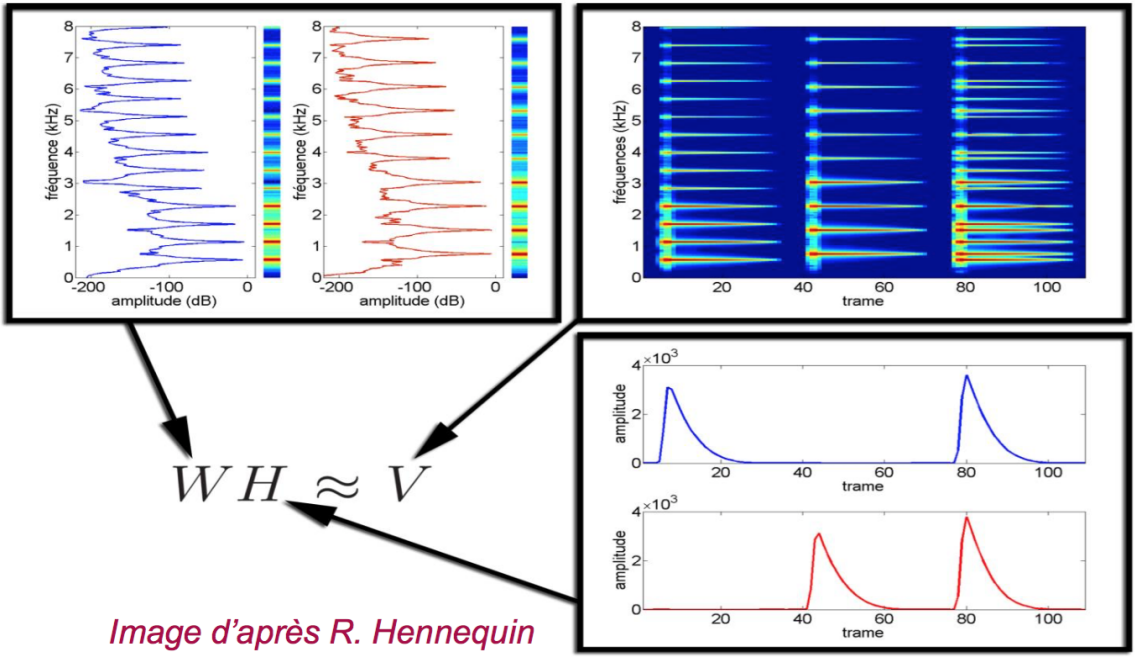
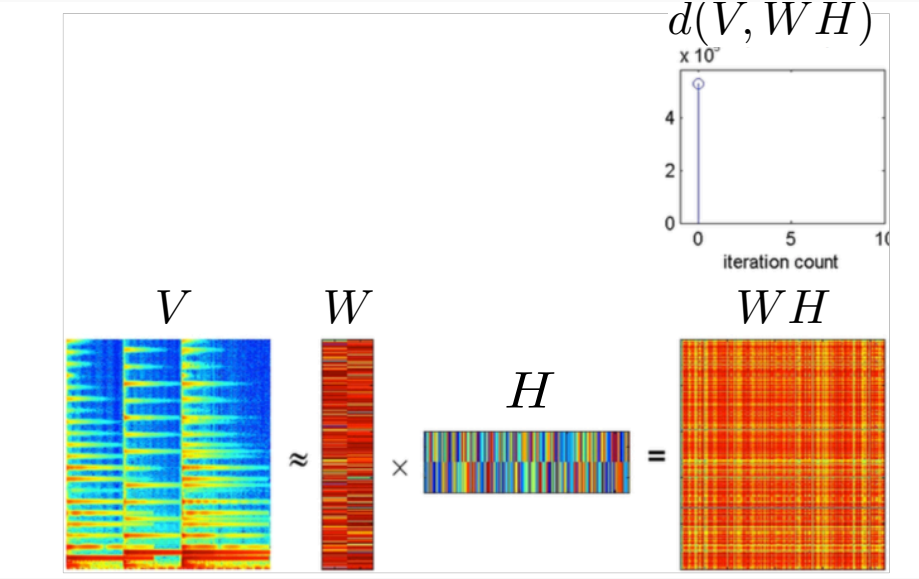


Image d'après R. Hennequin

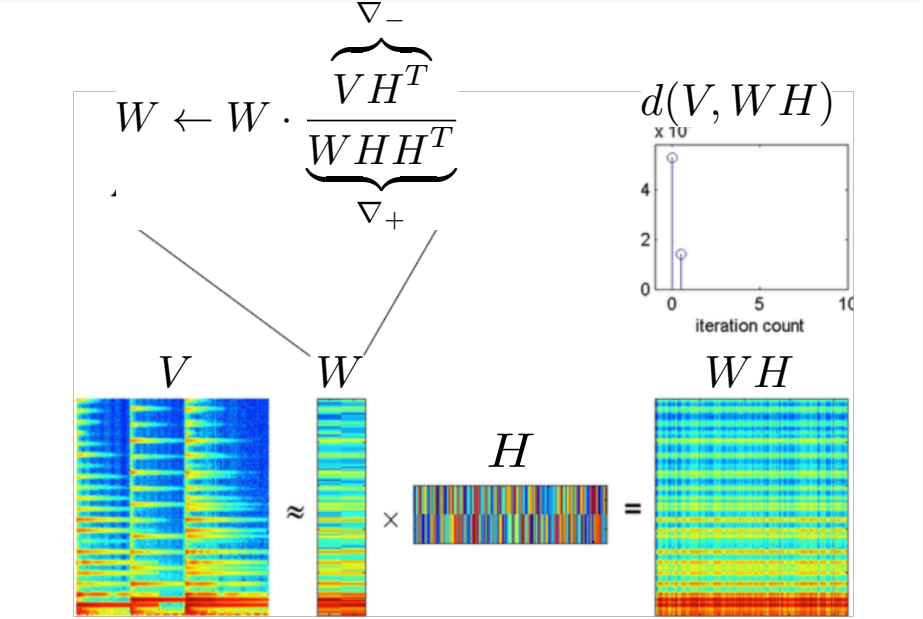
source : Romain Hennequin

Initialisation



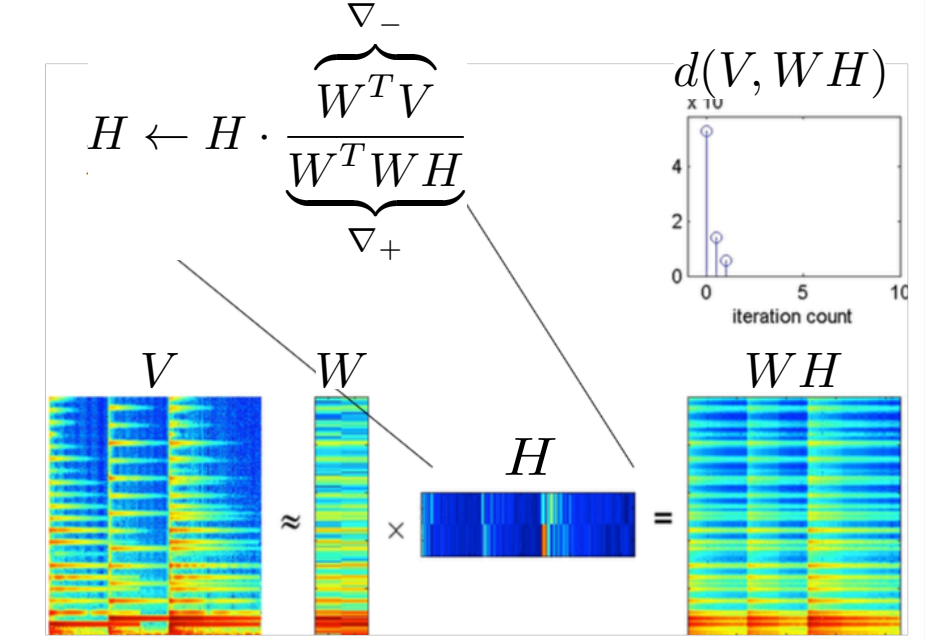
source : Tuomas Virtanen

Iteration 1 : Mise à jour de W



source : Tuomas Virtanen

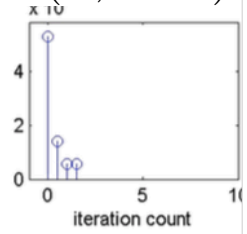
Iteration 1 : Mise à jour de H

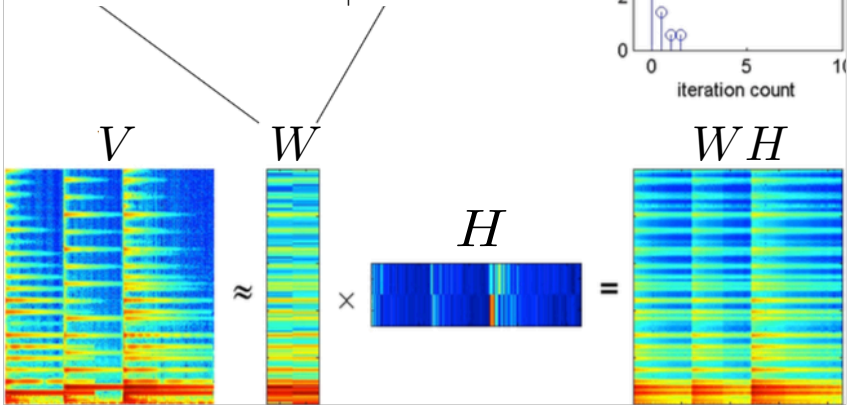


source : Tuomas Virtanen

Iteration 2 : Mise à jour de W

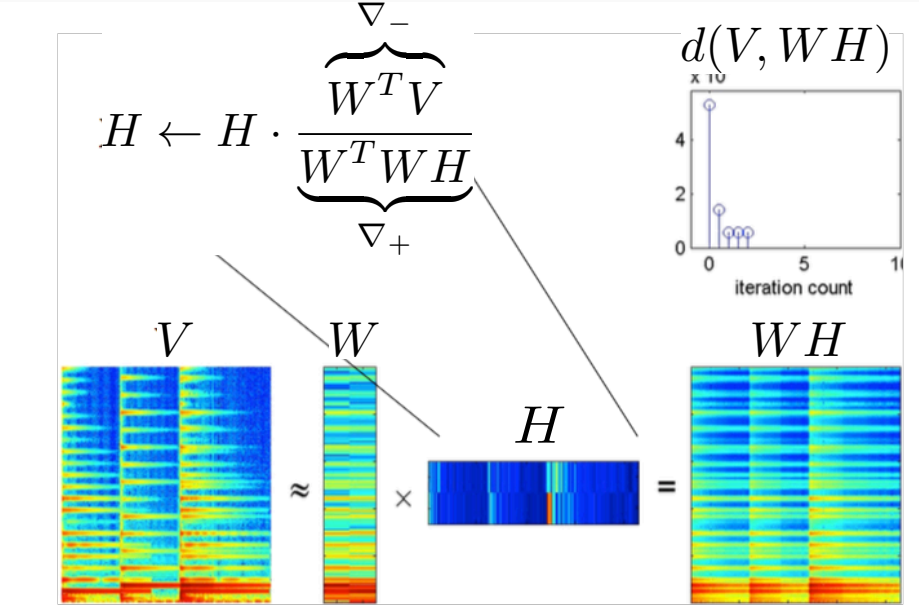
$$W \leftarrow W \cdot \frac{\overbrace{VH^T}^{\nabla_-}}{\underbrace{WHH^T}_{\nabla_+}}$$





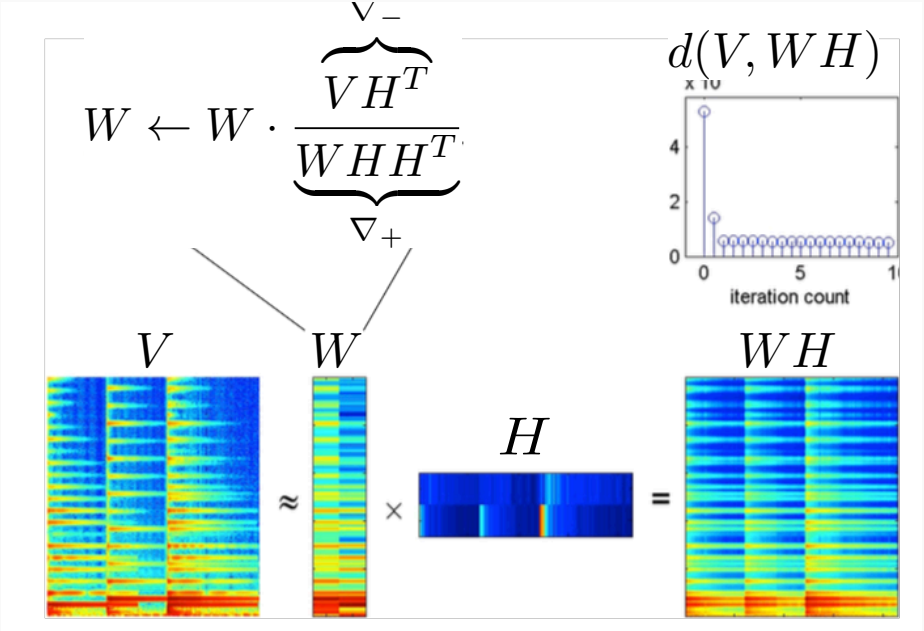
source : Tuomas Virtanen

Iteration 2 : Mise à jour de H



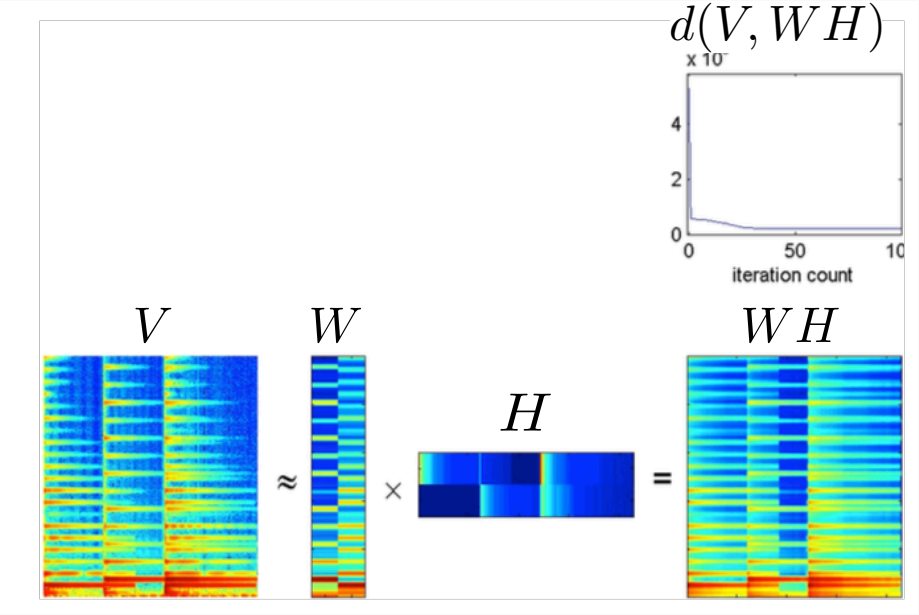
source : Tuomas Virtanen

Iteration 10 : Mise à jour de W



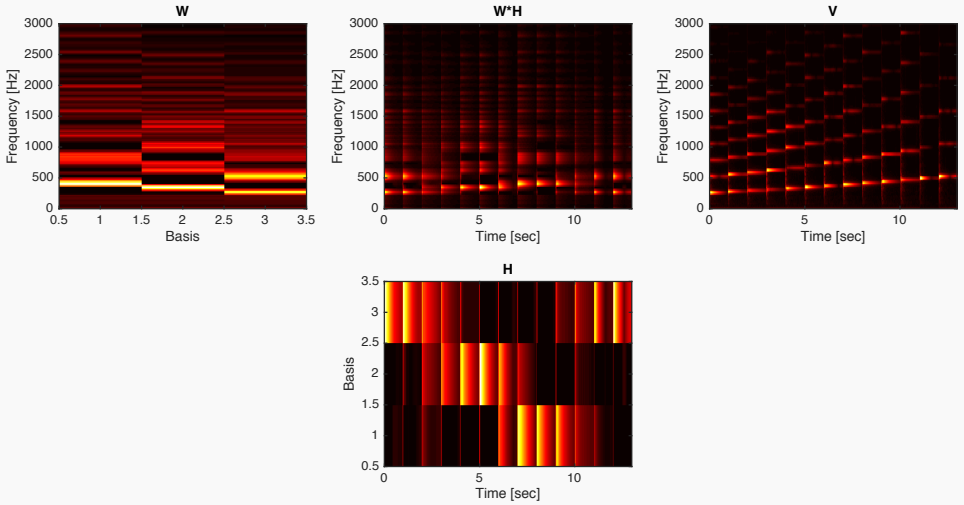
source : Tuomas Virtanen

Iteration 100

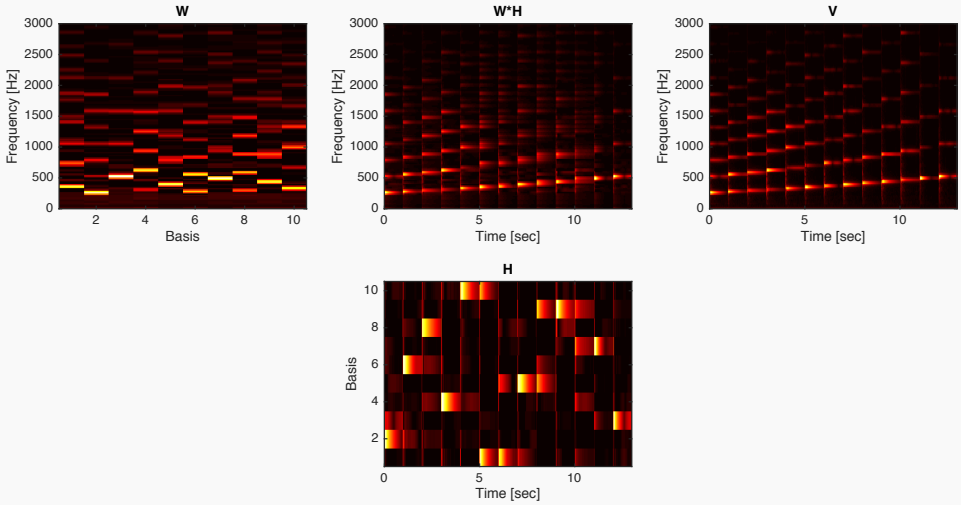


source : Tuomas Virtanen

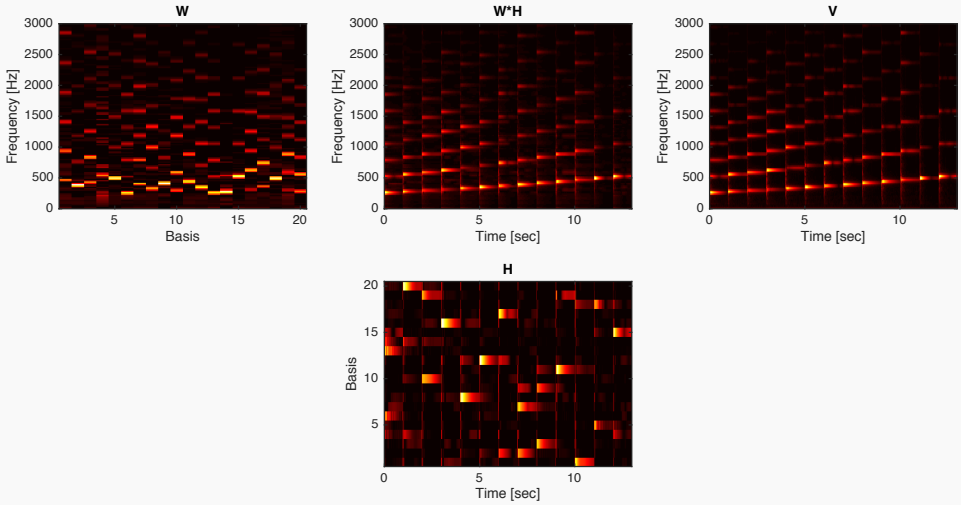
Choix du nombre de bases $K = 3$ (trop faible)



Choix du nombre de bases $K = 10$ (correcte)



Choix du nombre de bases $K = 20$ (trop grand)



6- Transformation du signal

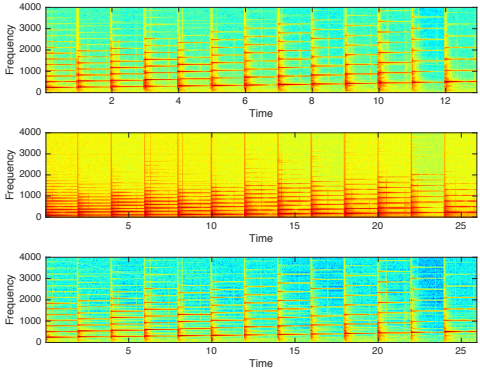
6.1- Application : dilatation/ contraction du temps par vocodeur de phase

Technique de DJ pour changer le tempo

- Ralentir/accélérer la vitesse de lecture (du vynil, de la bande magnétique)
- $x(\alpha t) \rightleftarrows \frac{1}{\alpha} X\left(\frac{f}{|\alpha|}\right)$
- Ralentir le temps : $\alpha < 1$
  - mais contracte aussi les fréquences (on abaisse les hauteurs)
- Accélérer le temps : $\alpha > 1$
  - mais étend aussi les fréquences (on augmente les hauteurs)

Objectif

- Changer le temps et les hauteurs de manière **indépendante**



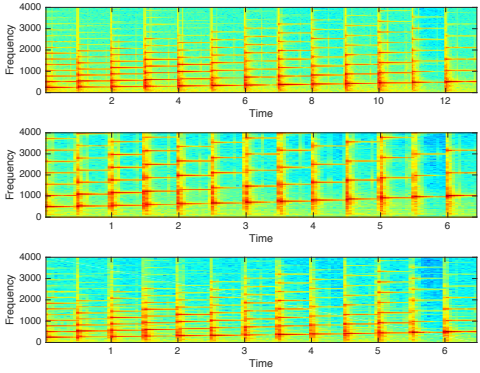
[haut] : signal original, [milieu] $\alpha < 1$ par ré-échantillonnage, [bas] : $\alpha < 1$ par vocodeur de phase

Technique de DJ pour changer le tempo

- Ralentir/accélérer la vitesse de lecture (du vynil, de la bande magnétique)
- $x(\alpha t) \rightleftarrows \frac{1}{\alpha} X\left(\frac{f}{|\alpha|}\right)$
- Ralentir le temps : $\alpha < 1$
  - mais contracte aussi les fréquences (on abaisse les hauteurs)
- Accélérer le temps : $\alpha > 1$
  - mais étend aussi les fréquences (on augmente les hauteurs)

Objectif

- Changer le temps et les hauteurs de manière **indépendante**

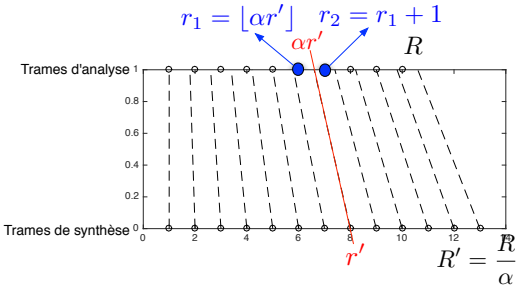


[haut] : signal original, [milieu] $\alpha > 1$ par ré-échantillonnage, [bas] : $\alpha > 1$ par vocodeur de phase

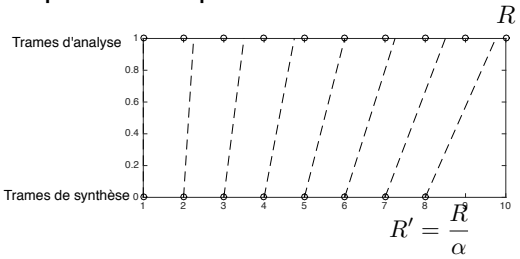
Le vocodeur de phase

- **Méthode pour rallonger/raccourcir le signal** :
  - changer le nombre de trames utilisées pour la resynthèse par TFCT^{-1}
- Soit R :
  - le nombre de trames d'**analyse** de la TFCT
- Soit $R' = \frac{R}{\alpha}$:
  - le nombre de trames de **synthèse** (utilisées pour la resynthèse par TFCT^{-1})
- $\alpha < 1 \rightarrow$ on dilate (ralentit) le temps
- $\alpha > 1 \rightarrow$ on compresse (accélère) le temps
- Le contenu d'une trame de synthèse $r' \in [1, R' = \frac{R}{\alpha}]$ est obtenu en recherchant les trames d'analyse r correspondantes les plus proches
  - $r_1 = \lfloor \alpha r' \rfloor$ et
  - $r_2 = r_1 + 1$

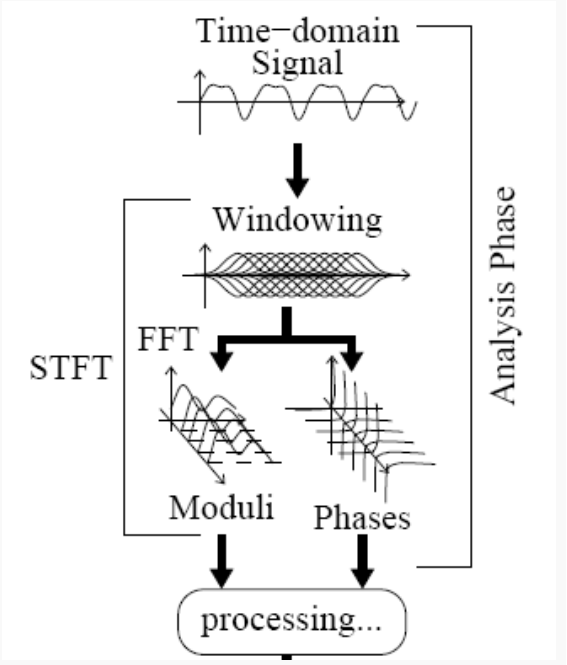
Dilation du temps $\alpha < 1$



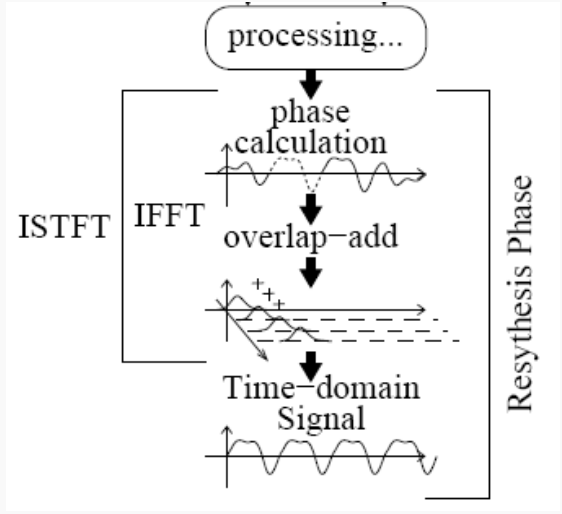
Compression du temps $\alpha > 1$



Le vocoder de phase : analyse



Le vocoder de phase : synthèse



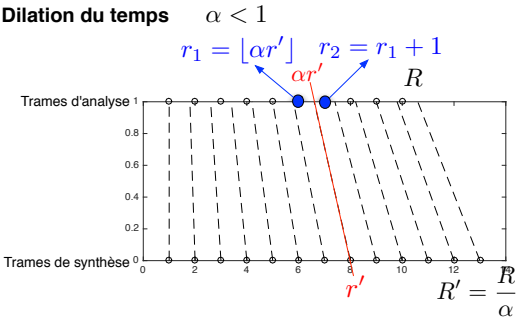
Le vocodeur de phase
1) Spectre d'amplitude

- Le spectre d'amplitude à la trame r', est obtenu par interpolation linéaire des spectres d'amplitude en r_1 et $r_2 = r_1 + 1$:
  - $A(k, r') = (1 - \Delta)A(k, r_1) + \Delta A(k, r_2)$
  - avec $\Delta = \alpha r' - r_1$

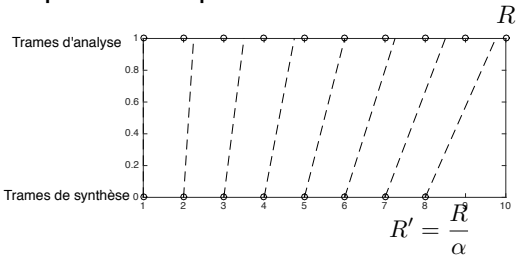
2) Spectre de phase

- **c'est plus compliqué ! ! !**

Dilation du temps $\alpha < 1$

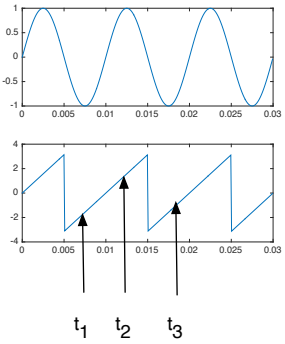


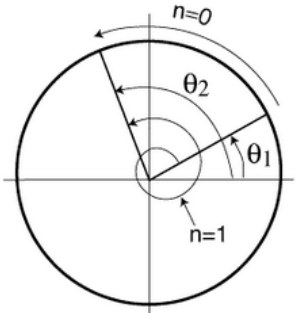
Compression du temps $\alpha > 1$



La phase et la fréquence instantanée

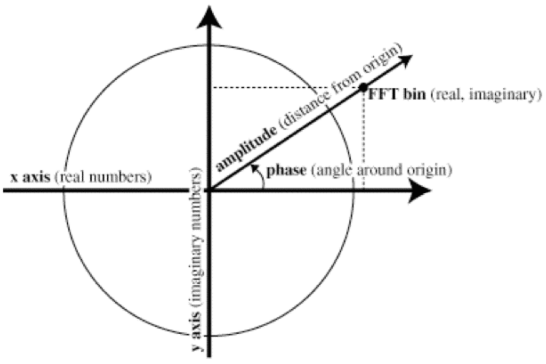
- Considérons un signal formé d'une sinusoide à la fréquence f_0 :
  - $x(t) = \sin(2\pi f_0 t)) = \sin(\phi(t))$
- Entre les instants t_1 et t_2, sa phase a "tournée" de $\phi(t_1)$ à $\phi(t_2)$
- Puisqu'il s'agit d'une sinusoide pure, elle a tournée de
  - $\phi(t_2) = \phi(t_1) + 2\pi f_0(t_2 - t_1)$
- On peut donc estimer f_0 à partir de la différence de phase
  - $f_0 = \frac{\phi(t_2)-\phi(t_1)}{2\pi(t_2-t_1)}$
- **Problème** :
  - la phase est uniquement définie dans l'intervalle $[-\pi, \pi]$
  - $\phi(t) \in [-\pi, \pi]$
  - donc en pratique le $\hat{\phi}(t_2)$ qu'on observe n'est pas $\phi(t_2)$ mais
    - $\hat{\phi}(t_2) = \phi(t_2) + n2\pi = \phi(t_1) + 2\pi f_0(t_2 - t_1) + n2\pi$
    - avec $n \in \mathbb{N}$ indéterminé
  - pour estimer f_0 il faut donc déterminer n
    - $f_0 = \frac{\phi(t_2)+n2\pi-\phi(t_1)}{2\pi(t_2-t_1)}$





Phase dans le Transformée de Fourier à Court Term (TFCT)

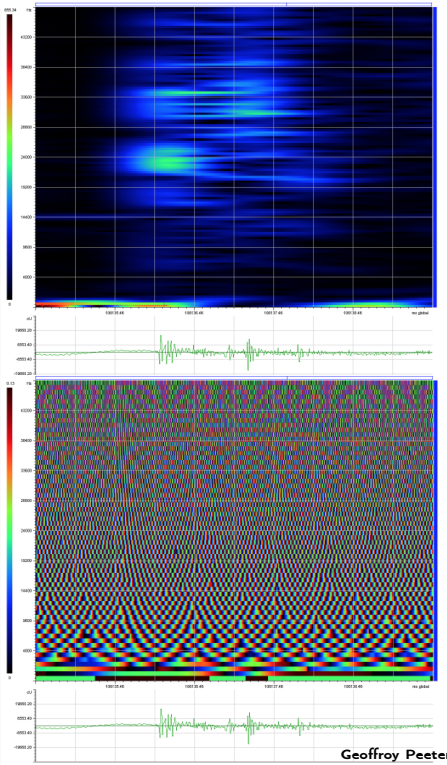
- Pour chaque trame n et fréquence k la TFTC est un nombre complexe
  - $X(k,n) = \sum_m x(m)w(n-m)e^{-j2\pi\frac{k}{N}m}$
- Il peut se décomposer en amplitude (module) et phase :
  - $X(k,n) = A(k,n)e^{j\phi(k,n)}$



FFT Cartesian to Polar Conversion

Phase dans le Transformée de Fourier à Court Term (TFCT)

- On a donc une valeur d'amplitude et de phase pour chaque (k, n)
- Spectrogramme
  - d'amplitude $A(k, n)$
  - de phase $\phi(k, n)$
- La phase indique la position de la cosinusoide,
- La variation temporelle de phase indique la fréquence instantanée
  - On peut donc calculer une fréquence instantanée pour chaque fréquence k et chaque couple de trames successives $(n-1) \rightarrow n$.

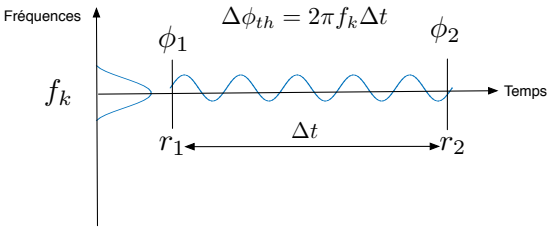


Le vocodeur de phase
2) Spectre de phase

- A la trame r' le spectre de phase dans le filtre k de la TFCT est obtenu en propageant la phase à partir de la fréquence contenu dans ce filtre
- 1) Solution **simplifiée** :
  - On suppose qu'à travers le filtre k de la TFCT on peut observer uniquement une sinusoide à la fréquence f_k
  - On propage l'évolution de la phase au cours du temps en utilisant la prédiction théorique de la phase : $\Delta\phi_{th} = 2\pi f_k \Delta t$
  - Donc

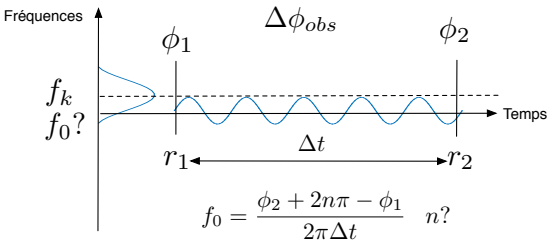
$$\begin{aligned}\phi(k, r') &= \phi(k, r'-1) + \Delta\phi_{th} \\ &= \phi(k, r'-1) + 2\pi f_k \Delta t\end{aligned}$$

  - avec comme phase **initiale** : $\phi(k, r'=1) = \phi(k, r=1)$



Le vocodeur de phase
2) Spectre de phase

- 2) Solution **correcte** :
  - En pratique, à travers le filtre k de la TFCT, on peut observer des sinusoides à des fréquences proches mais différentes de f_k
    - ceci est du à largeur du lobe principale, aux lobes secondaires
  - Il faut **estimer cette fréquence** f_0 que l'on observe à travers le filtre f_k pour ensuite appliquer la propagation de phase
    - $\phi(k, r') = \phi(k, r' - 1) + 2\pi f_0 \Delta t$



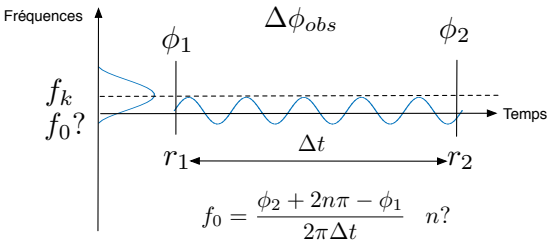
Le vocodeur de phase : spectre de phase

- **Estimer cette fréquence f_0 ?**
  - En utilisant la **fréquence instantanée** :
    - $f_0(n) = \frac{\phi_2 + n2\pi - \phi_1}{2\pi\Delta t}$
  - **Déterminer n ?**
    - on cherche n tel que $f_0 \simeq f_k$

n tel que $\min_n |f_0 - f_k|$

$$\min_n |\frac{\phi_2 + n2\pi - \phi_1}{2\pi\Delta t} - f_k|$$
$$\min_n |\phi_2 + n2\pi - \phi_1 - 2\pi\Delta t f_k|$$
$$\min_n |\phi_2 + n2\pi - \phi_1 - \Delta\phi_{th}|$$

- ce qui revient à
  - trouver la détermination principale (la valeur dans l'intervalle $[-\pi, \pi]$) de
    - $n = \lfloor(\phi_2 - \phi_1 - \Delta\phi_{th})/(2\pi)\rceil$
  - il s'agit de la différence de phase non-expliquée par le modèle théorique $\Delta\phi_{th}$

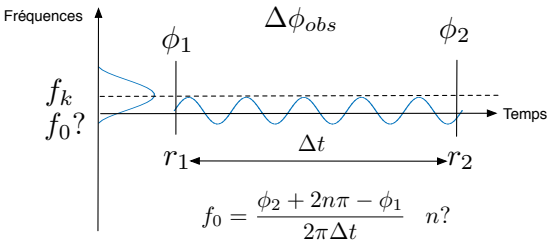


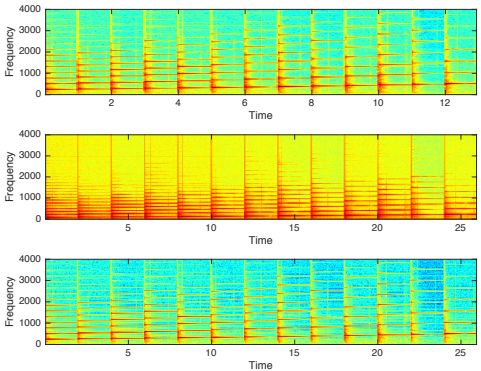
Le vocodeur de phase : spectre de phase

- 2) Solution **correcte** :
  - Finalement la phase est incrémentée de

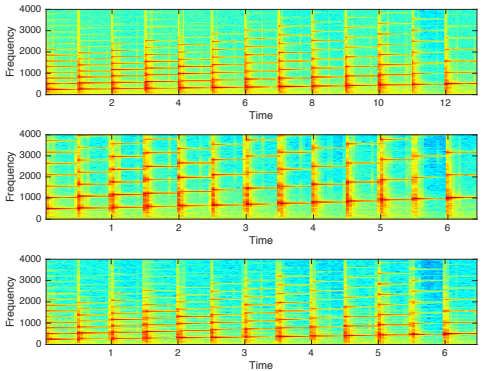
$$\begin{aligned}\phi(k, r') &= \phi(k, r'-1) + 2\pi f_0 \Delta t \\ &= \phi(k, r'-1) + \phi_2 + n2\pi - \phi_1\end{aligned}$$

  - avec comme phase **initiale** :
    $\phi(k, r'=1) = \phi(k, r=1)$





[haut] : signal original, [milieu] $a < 1$ **par ré-échantillonnage, [bas] :** $a < 1$ **par vocodeur de phase**



[haut] : signal original, [milieu] $a > 1$ **par ré-échantillonnage, [bas] :** $a > 1$ **par vocodeur de phase**

Changement de hauteur

- Ré-échantillonnage du signal pour correction de la longueur par phase-vocoder

Figure 8.24 Resampling of a time stretching algorithm.