ENSEA 3ème SyM
Traitement du signal audio musical:
Transformation et séparation du son

Geoffroy.Peeters@ircam.fr
UMR SMTS IRCAM CNRS UPMC

1. Théorie : Traitement du signal fréquentiel

1.1 Transformée de Fourier (temps et fréquences continus)
1.2 Transformée de Fourier (temps et fréquences discrets)
1.3 Transformée de Fourier (à Court Terme) : TFCT
1.4 Transformée à Q-Constant (CQT)
1.5 Deux interprétations de la TFCT
1.6 Reconstruction du signal par addition/ recouvrement (TFTC inverse)
1.7 Application : filtrage constant au cours du temps
1.8 Application : débruitage par soustraction spectrale
1.9 Application : dilatation/ contraction du temps par vocodeur de phase

2. Séparation de sources

2.1 Séparation Harmonique Percussive (HPS)
2.2 Décomposition en matrice non-négatives (NMF)

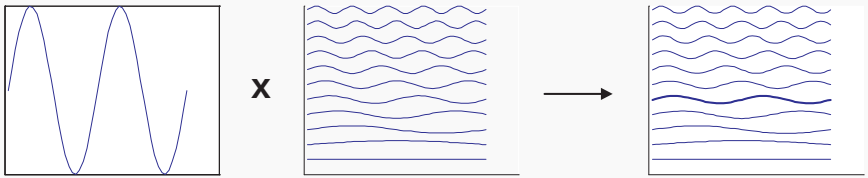
Théorie : Traitement du signal fréquentiel

1- Théorie : Traitement du signal fréquentiel

Transformée de Fourier (temps et fréquences continus)

$$X(\omega) = \int_{t=-\inf}^{+\inf} x(t) e^{-j\omega t} dt \quad X(f) = \int_{t=-\inf}^{+\inf} \exp(-j2\pi ft) dt \tag{1}$$

- 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



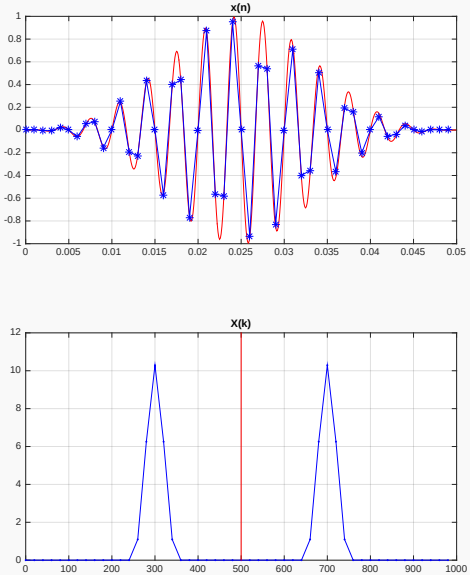
Propriété de la Transformée de Fourier (temps et fréquences continus)

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
	$x^\star(t)$	$X^\star(f)$

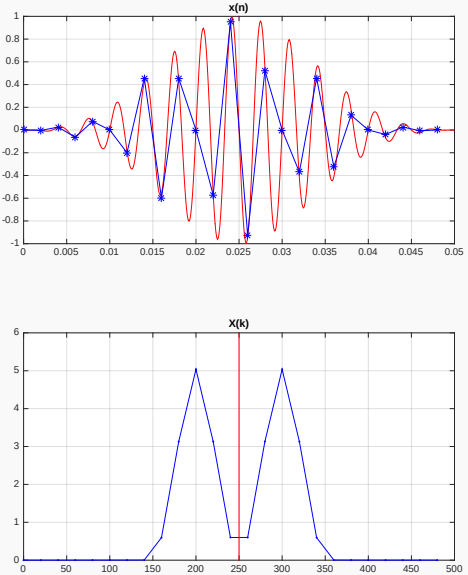
$$X(k) = \sum_{m=0}^{N-1} x(m) e^{-j2\pi \frac{k}{N} m} \qquad \forall k \in [0, N]$$

- Variables :
  - n 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** : $f_{Nyquist} = \frac{sr}{2} > f_{\text{max}}$

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



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

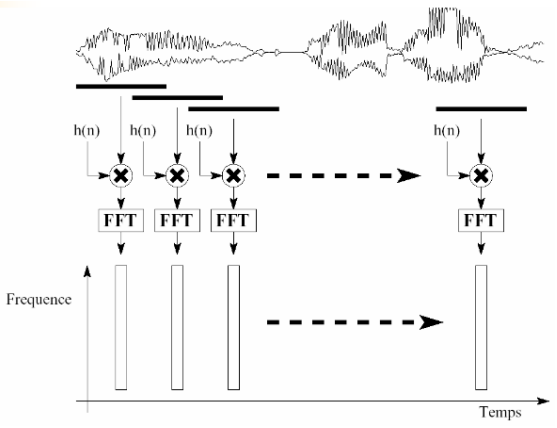


$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \qquad \forall k \in [0,N]$$

- 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

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

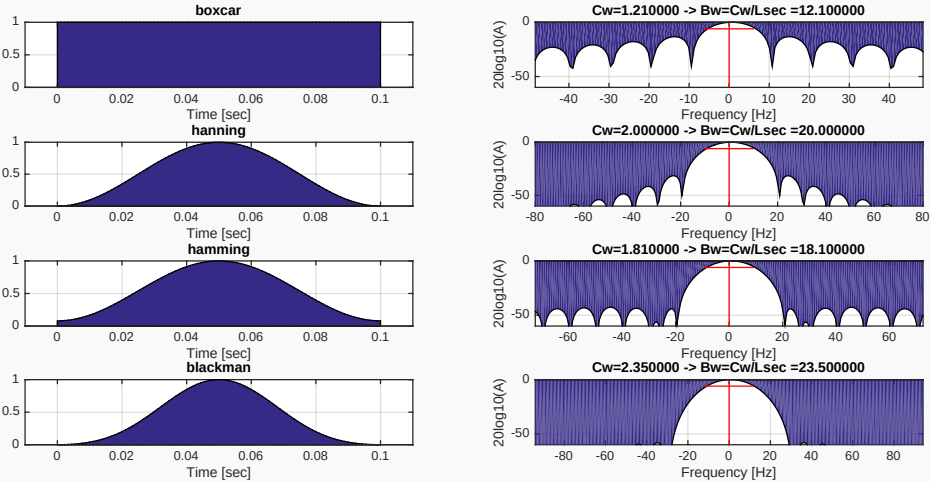
Fenêtre de pondération $w(t)$

- $x(t) \cdot w(t) \rightleftharpoons X(f) \circledast W(f)$
  - $w(t)$ 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}) : $Bw = \frac{Cw}{L}$
    - 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
  - $w(n) = a_0 - a_1\cos(\frac{2\pi n}{N-1}) + a_2\cos(\frac{2\pi n}{N-1})$
  - $Bw = 2.35$

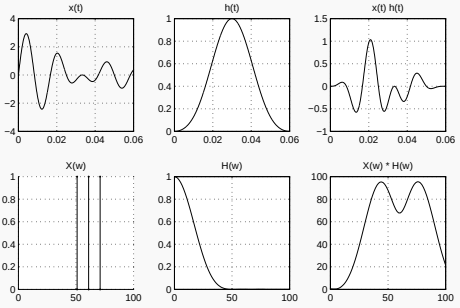
Influence du **type** de la fonction



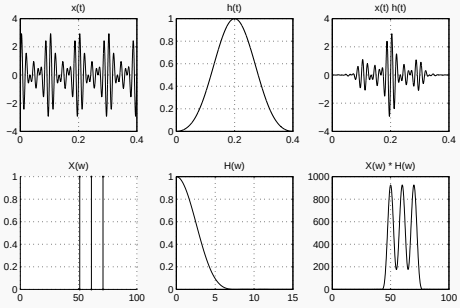
Choix de la **longueur temporelle** L :

- 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.

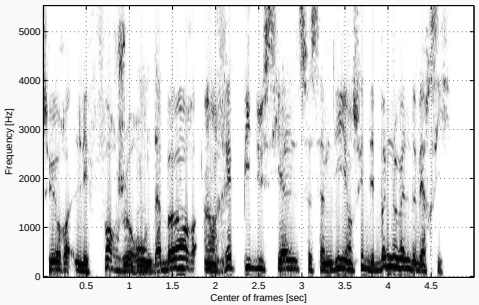
Influence de la **longueur temporelle** L ($L = 0.06$s.)



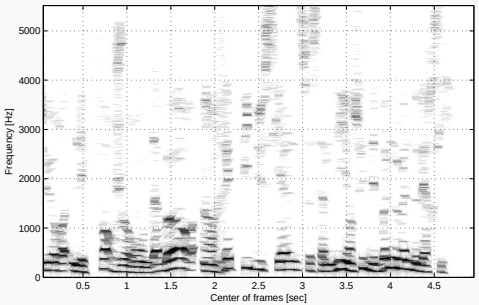
Influence de la **longueur temporelle** L ($L = 0.4$s.)



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

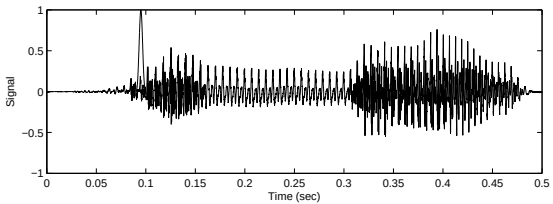


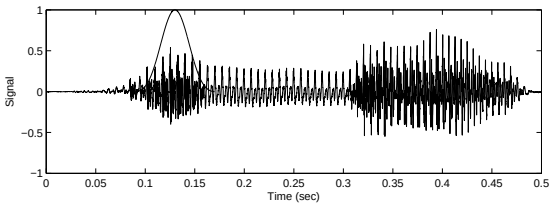
Influence de la **longueur temporelle** L ($L = 0.1$s.)

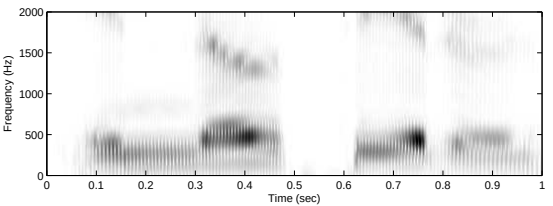


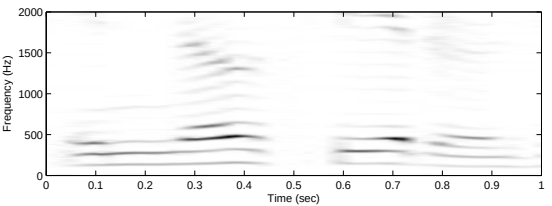
Paradoxe temps/ fréquence

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









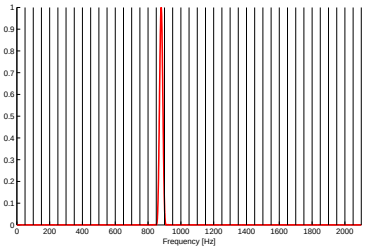
- Comme résoudre ce problème ?
  - Utiliser d'autres tranformées que celle de Fourier

Transformée à Q-Constant (CQT)

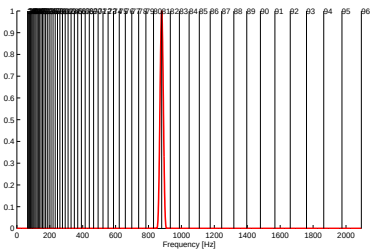
- La DFT
  - Définition : **La précision fréquentielle** : $\Delta f = \frac{sr}{N}$
    - c'est le pas d'échantillonnage du spectre
    - elle dépend de la taille de la DFT : N
    - on peut l'augmenter en augmentant N
  - Définition : **La résolution fréquentielle** : $Bw = \frac{Cw}{L}$
    - c'est 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

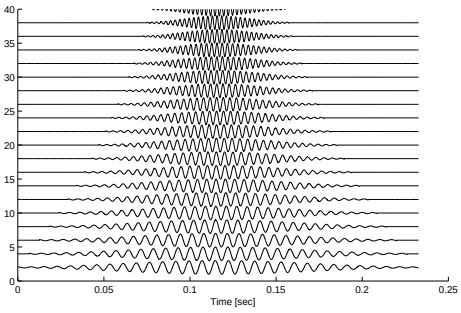
Transformée à Q-Constant

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

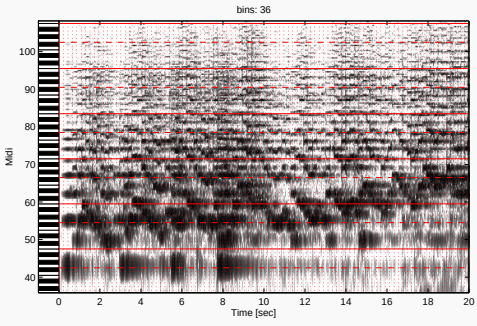
- Solution ?
  - Changer la **résolution fréquentielle** en fonction des fréquences considérées
- Comment ?
  - En changeant la longueur temporelle de la fenêtre pour chaque fréquence considérée
  - Le facteur $Q = \frac{f_k}{f_{k+1} - f_k}$ doit rester constant en fréquence

$$Q = \frac{f_k}{Bw} = \frac{f_k}{Cw/L} = \frac{f_k \cdot L}{Cw} \qquad (2)$$

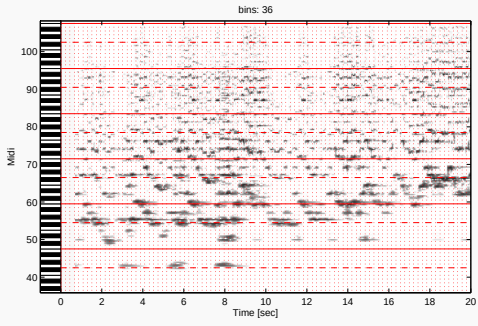
  - on choisit un L pour chaque fréquence f_k
    - $L_k = \frac{Q \cdot Cw}{f_k}$



Exemples (en utilisant la DFT)

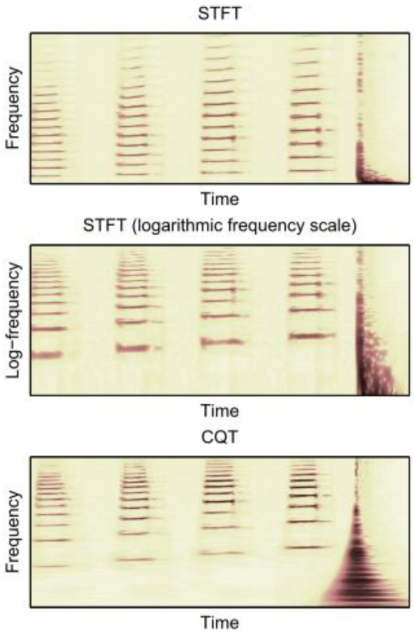


Exemples (en utilisant la CQT)



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

Deux interprétations de la TFCT

- Interprétation **passe-bas** :
  - on regarde l'évolution du signal à une fréquence f_0 donnée

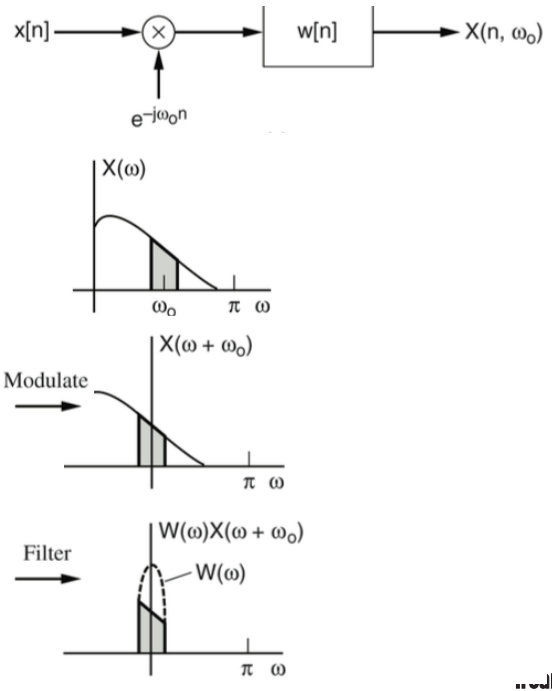
$$X(f,n) = \sum_{m=-\inf}^{+\inf} x(m)w(n-m)e^{-j2\pi fm}$$

$$X(f_0,n) = \sum_{m=-\inf}^{+\inf} \left[x(m)e^{-j2\pi f_0 m}\right] w(n-m)$$

$$= \sum_{m=-\inf}^{+\inf} x_0(m)w(n-m)$$

$$= x_0(n) \circledast w(n) \qquad (3)$$

- avec $x_0(m) = x(m)e^{-j2\pi fm}$ le signal modulé
- il s'agit d'une convolution de $x_0(m)$ par le filtre passe-bas $w_0(m)$



source : Patrick J. Wolfe, 2009

Deux interprétations de la TFCT

- Interprétation **passe-bande** :
  - on regarde l'évolution du signal à une fréquence f_0 donnée

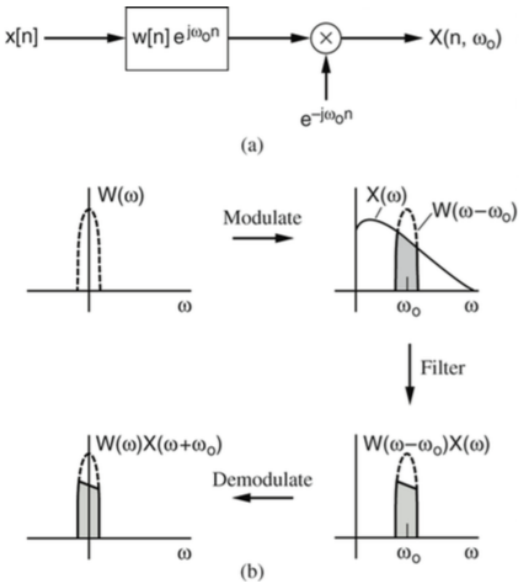
$$X(f,n) = \sum_{m=-\inf}^{+\inf} x(m)w(n-m)e^{-j2\pi fm}$$

$$X(f_0,n) = \sum_{m=-\inf}^{+\inf} x(m)w_0(n-m)e^{-j2\pi f_0 n}$$

$$= e^{-j2\pi f_0 n} \sum_{m=-\inf}^{+\inf} x(m)w_0(n-m)$$

$$= e^{-j2\pi f_0 n} \cdot [x(m) \circledast w_0(n)] \quad (4)$$

- avec $w_0(m) = w(n-m)e^{j2\pi f_0(n-m)}$ la fenêtre démodulée
- il s'agit d'une convolution de $h(m)$ par le filtre passe-bande $h_0(m)$



source : Patrick J. Wolfe, 2009

On peut reconstruire le signal audio à partir des informations de la TFCT

- Méthode d'addition/recouvrement (OverLap-Add, OLA)

$$X(k,n) = \sum_m x(m) w(n-m) e^{-j2\pi \frac{k}{N} m}$$

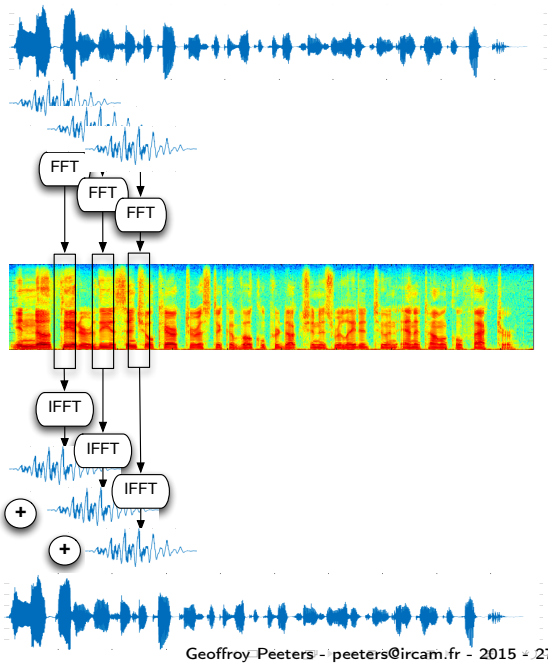
$$\frac{1}{N} \sum_{k=0}^{N-1} X(k,n) e^{+j2\pi \frac{k}{N} m} = x(m) w(n-m) \quad (5)$$

si n=m $\quad \frac{1}{Nw(0)} \frac{1}{N} \sum_{k=0}^{N-1} X(k,n) e^{+j2\pi \frac{k}{N} m} = x(n)$

On peut reconstruire le signal audio à partir des informations de la TFCT

- Si on note $n = RI$
  - r le numéro de la trame d'analyse
  - I le pas d'avancement

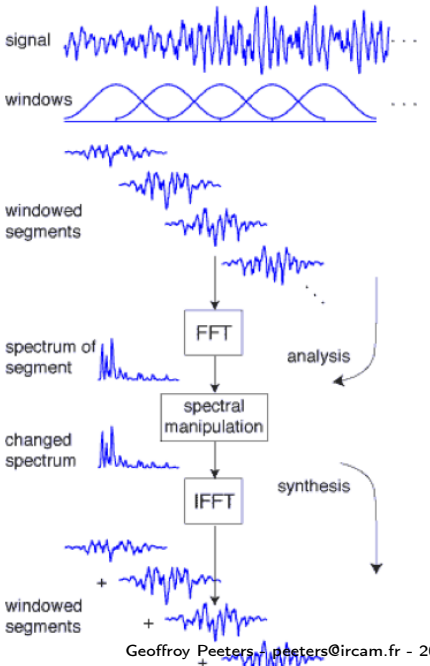
$$
\begin{aligned}
y(m, rI) &= \frac{1}{N} \sum_{k=0}^{N-1} X(k, rI) e^{+j2\pi \frac{k}{N} m} \\
&= x(m) w(rI - m) \\
y(m) &= \sum_r y(m, rI) \\
&= \sum_r x(m) w(rI - m) \\
&= x(m) \sum_r w(rI - m) \\
x(n) &= \frac{\sum_r y(m, rI)}{\sum_r w(rI - n)}
\end{aligned}
\tag{6}
$$



On peut reconstruire le signal audio à partir des informations de la TFCT

- Si on note $n = RI$
  - r le numéro de la trame d'analyse
  - I le pas d'avancement

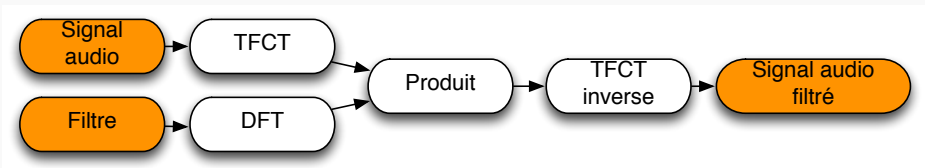
$$
\begin{aligned}
y(m, rI) &= \frac{1}{N} \sum_{k=0}^{N-1} X(k, rI) e^{+j2\pi \frac{k}{N} m} \\
&= x(m) w(rI - m) \\
y(m) &= \sum_r y(m, rI) \\
&= \sum_r x(m) w(rI - m) \\
&= x(m) \sum_r w(rI - m) \\
x(n) &= \frac{\sum_r y(m, rI)}{\sum_r w(rI - n)}
\end{aligned}
\tag{7}
$$



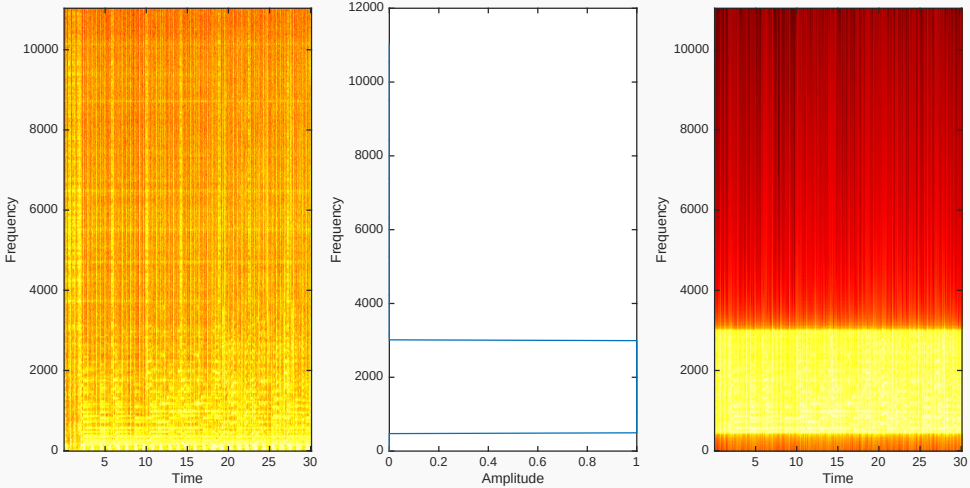
Application : filtrage constant au cours du temps

Filtrage dans le domaine fréquentiel= très économique en coût de calcul

- $x(t) \circledast h(t) \rightleftarrows X(\omega)H(\omega)$
- convolution en temps $\rightleftarrows$ produit en fréquence
- utilisation de l'algorithme FFT

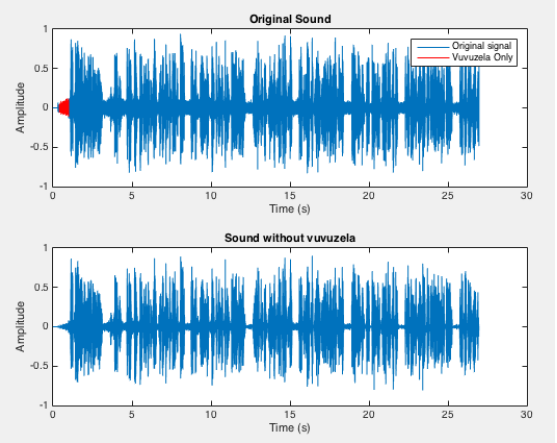


Application : filtrage constant au cours du temps



Application : débruitage par soustraction spectrale

- soit $x(t) = s(t) + n(t)$
  - $s(t)$ est un signal de parole
  - $n(t)$ est un bruit additif
  - on peut écrire le modèle : $X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega})$
- **Méthode**
  - On cherche un filtre fréquentiel $H(e^{j\omega})$ permettant de retirer le bruit additif
  - Amplitude de ce filtre
    - = valeur moyenne de $|N(e^{j\omega})|^2$ calculée sur un segment ne contenant pas de parole
  - Phase
    - = la phase de X : $\theta_x(\omega)$



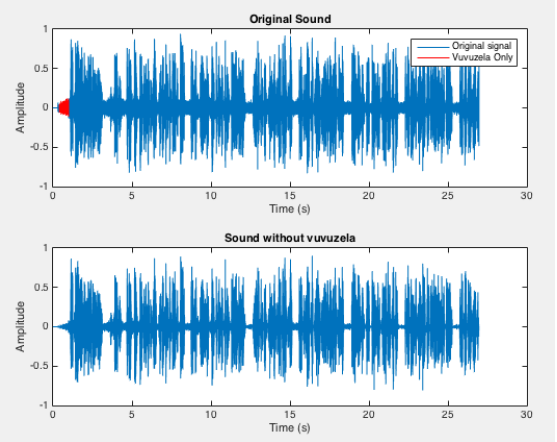
Application : débruitage par soustraction spectrale

- Soustraction :

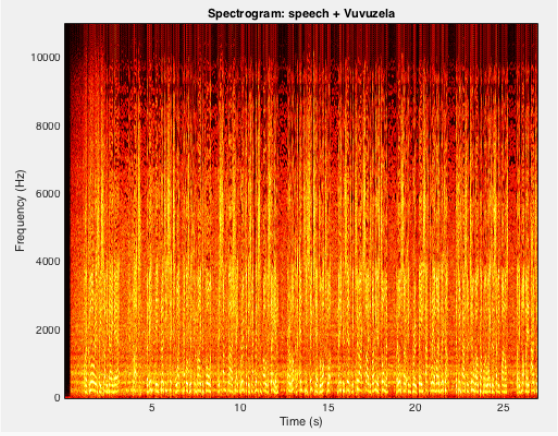
$$\begin{aligned}\hat{S}(e^{j\omega}) &= \left[|X(e^{j\omega})| - \mu(e^{j\omega})\right] e^{j\theta_x} \\ &= H(e^{j\omega})X(e^{j\omega})\end{aligned} \tag{8}$$

- avec $H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|}$
- avec $\mu(e^{j\omega}) = E\{|N(e^{j\omega})|\}$

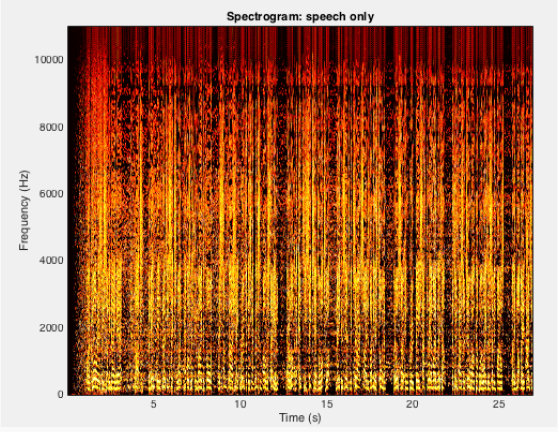
- **Amélioration** :
  - pour éviter des problèmes lorsque $|X(e^{j\omega})| < \mu(e^{j\omega})$ (quand le spectre d'amplitude est $<$ au spectre moyen du bruit)
  - rectification demi-onde (half-wave rectification) : $H_R(e^{j\omega}) = \frac{H(e^{j\omega}) + |H(e^{j\omega})|}{2}$



Spectrogramme speech+noise



Spectrogramme speech

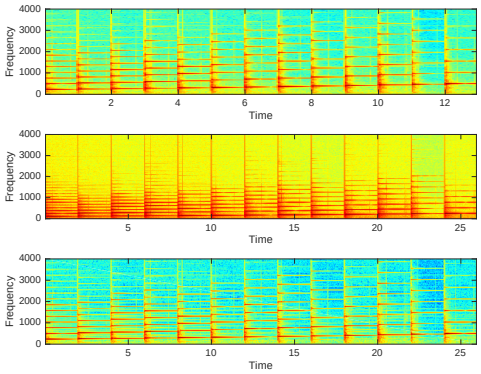


Technique de DJ pour changer le tempo

- ralentir la vitesse de lecture (du vynil, ce la bande magnétique)
- $x(at) \rightleftarrows \frac{1}{a} X\left(\frac{f}{|a|}\right)$
- si $a < 1$
  - on ralentit le temps
  - mais on contracte aussi les fréquences (on abaisse les hauteurs)
- si $a > 1$
  - on accélère le temps
  - mais on étend aussi les fréquences (on augmente les hauteurs)

Objectif

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



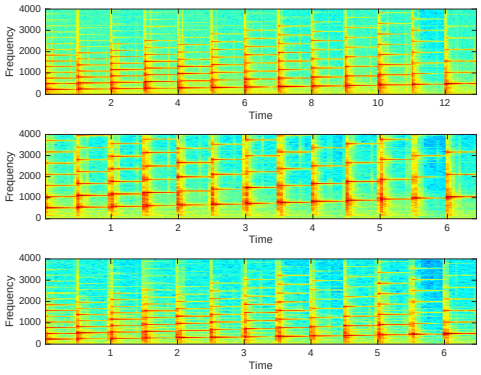
[haut] : signal original, [milieu] $a < 1$ par ré-échantillonnage, [bas] : $a < 1$ par vocodeur de phase

Technique de DJ pour changer le tempo

- ralentir la vitesse de lecture (du vynil, ce la bande magnétique)
- $x(at) \rightleftarrows \frac{1}{a} X\left(\frac{f}{|a|}\right)$
- si $a < 1$
  - on ralentit le temps
  - mais on contracte aussi les fréquences (on abaisse les hauteurs)
- si $a > 1$
  - on accélère le temps
  - mais on étend aussi les fréquences (on augmente les hauteurs)

Objectif

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

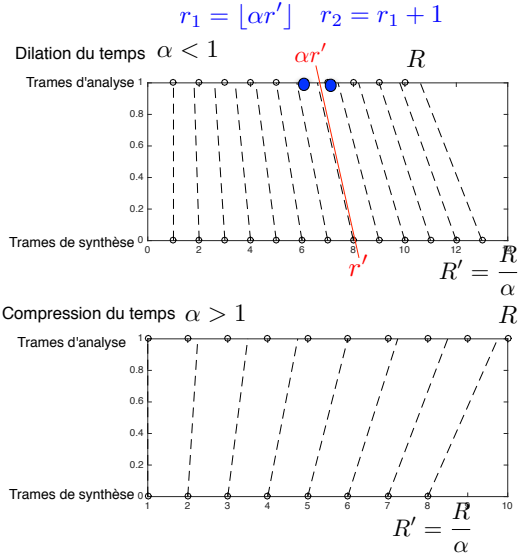


[haut] : signal original, [milieu] $a > 1$ par ré-échantillonnage, [bas] : $a > 1$ par vocodeur de phase

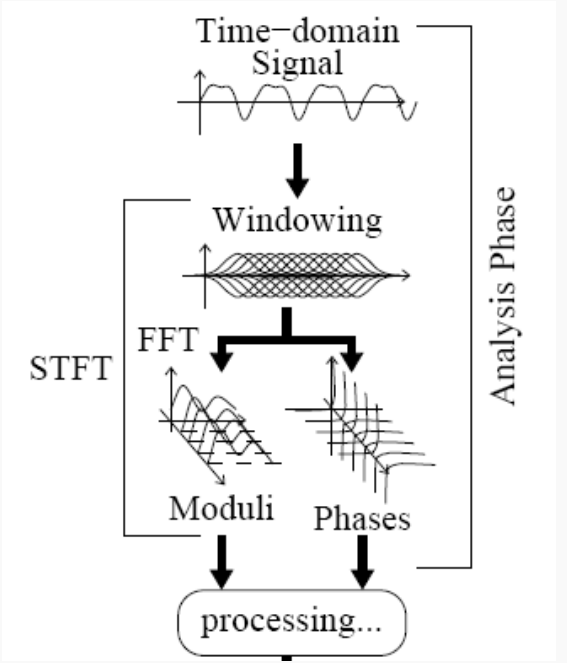
Le vocodeur de phase

- **Méthode** :
  - **pour raccourcir/ rallonger le signal, on va changer le nombre de trames utilisées pour la resynthèse par TFCT inverse**
- Soit R : le nombre de trames d'analyse de la TFCT
- Soit $R' = \alpha R$: le nombre de trames de synthèse (utilisées pour la resynthèse par TFCT inverse)
  - si $\alpha < 1$, on dilate le temps du signal (on le ralentit)
  - si $\alpha > 1$, on compresse le temps du signal (on l'accélère)
- 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
  - $r_1 = \lfloor \alpha r' \rfloor$ e
  - $r_2 = \lceil \alpha r' \rceil$

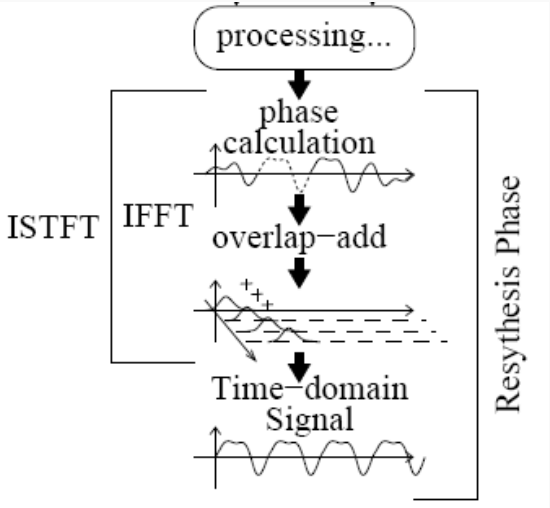
$r_1 = \lfloor \alpha r' \rfloor \quad r_2 = r_1 + 1$



Le vocoder de phase : analyse

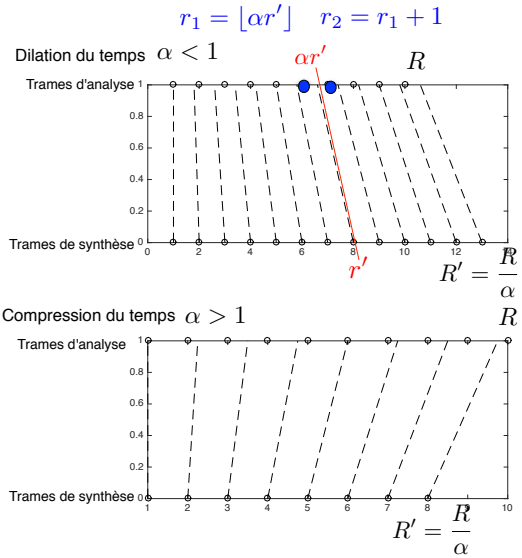


Le vocoder de phase : synthèse



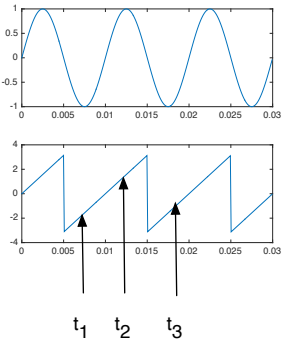
Le vocodeur de phase : spectre d'amplitude

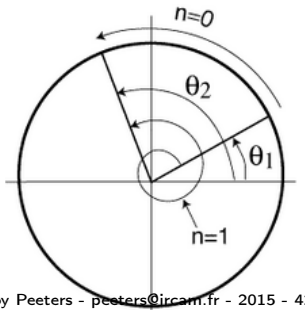
- Le spectre d'amplitude à la trame r', est obtenu par interpole linéaire des spectres d'amplitude en r_1 et $r_2 = r_1 + 1$:
  - $A(k, r') = (1 - \Delta)A(k, r_1) + \Delta A(k, r_2)$
  - avec $\Delta = \alpha r' - r_1$
- Le spectre de phase ?
  - **c'est peu plus compliqué ! ! !**



La phase et la fréquence instantanée

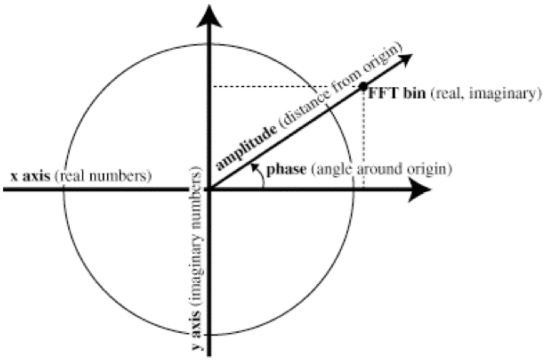
- Considérons un signal formé d'une sinusoide pure à la fréquence f_0 : $x(t) = \sin(\phi(t)) = \sin(2\pi f_0 t))$
  - entre les instants t_1 et t_2, sa phase a "tournée" de $\phi(t_1)$ à $\phi(t_2)$
  - puisqu'il s'agit d'une sinusoide pure, elle a tournée de $\phi(t_2) = \phi(t_1) + 2\pi f_0(t_2 - t_1)$
  - on peut donc estimer la fréquence 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 $[-\pi, \pi]$
  - donc en pratique le $\hat{\phi}(t_2)$ qu'on observe n'est pas tel
    - $\phi(t_2) = \phi(t_1) + 2\pi f_0(t_2 - t_1)$
  - mais est
    - $\hat{\phi}(t_2) + n2\pi = \phi(t_1) + 2\pi f_0(t_2 - t_1)$, avec n indéterminé
  - pour déterminer 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)

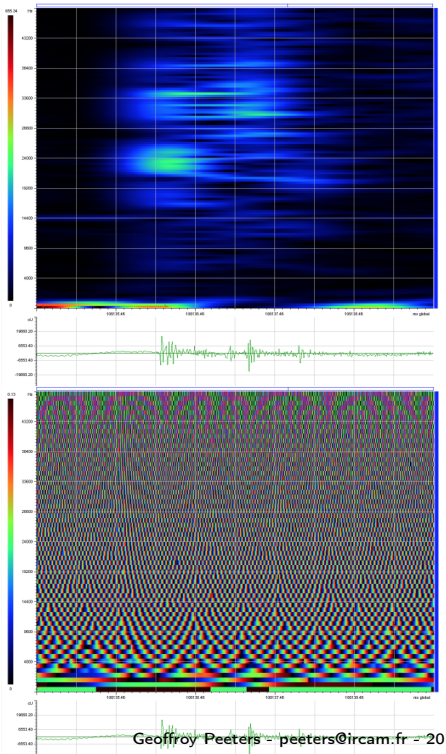
- Pour chaque trame n et fréquence k la TFTC est un nombre complexe
  - $X(k,n) = \sum_m x(m)w(n-m)e^{-j2\pi \frac{k}{N} m}$
- Il peut se décomposer en amplitude (module) et phase :
  - $X(k,n) = A(k,n)e^{j\phi(k,n)}$



FFT Cartesian to Polar Conversion

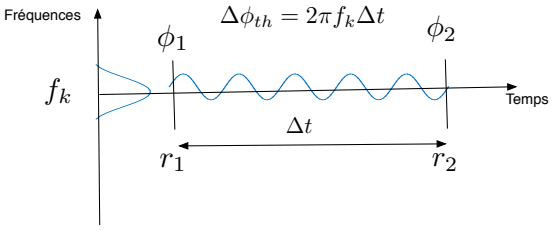
Phase dans le Transformée de Fourier à Court Term (TFCT)

- 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 $(n-1) \rightarrow n$.



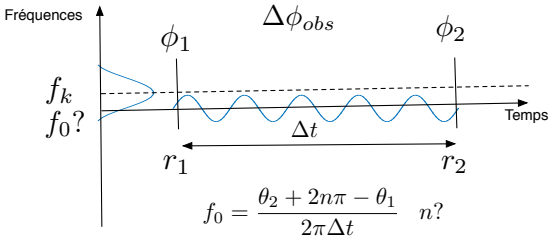
Le vocodeur de phase : spectre de phase

- 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 que le filtre k contient une sinusoide à la fréquence f_k
  - si on note
    - $\phi_1 = \phi(k, r_1)$
    - $\phi_2 = \phi(k, r_2)$
    - $\Delta t = t_2 - t_1$
  - puisque le pas de synthèse est égale au pas d'analyse : $r' - (r' - 1) = r_2 - r_1$
- on utilise la prédiction théorique de la phase : $\Delta\phi_{th}$
  - $\phi(k, r') = \phi(k, r' - 1) + \Delta\phi_{th}$
  - $\phi(k, r') = \phi(k, r' - 1) + 2\pi f_k \Delta t$
  - avec comme phase **initiale** : $\phi(k, r' = 1) = \phi(k, r = 1)$



Le vocodeur de phase : spectre de phase

- 2) Solution **correcte** :
  - En pratique il se peut qu'on observe à travers le filtre k une sinusoide à une fréquence proche mais différente 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 f_0 ?**
  - En utilisant la **fréquence instantanée** :
    - $f_0(n) = \frac{\phi_2 + 2\pi n - \phi_1}{2\pi\Delta t}$
  - **Comme déterminer n ?**
    - en cherchant n tel que $f_0 \simeq f_k$

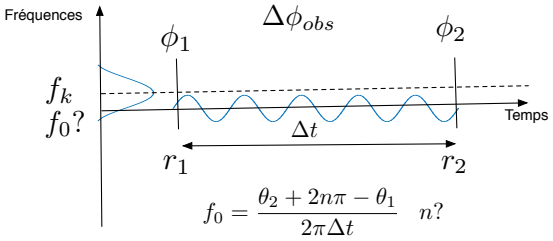
$$n \text{ tel que } \min_n |f_0 - f_k|$$

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

$$\min_n |\phi_2 + 2n\pi - \phi_1 - 2\pi\Delta t f_k|$$

$$\min_n |\phi_2 + 2n\pi - \phi_1 - \Delta\phi_{th}| \quad (9)$$

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

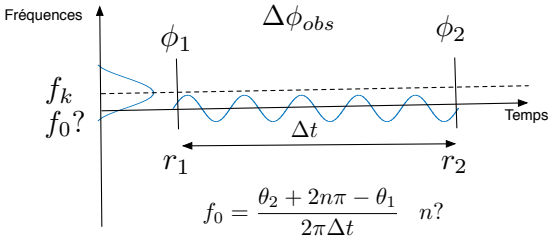


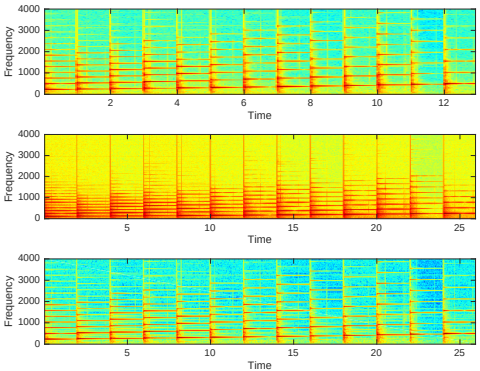
Le vocodeur de phase : spectre de phase

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

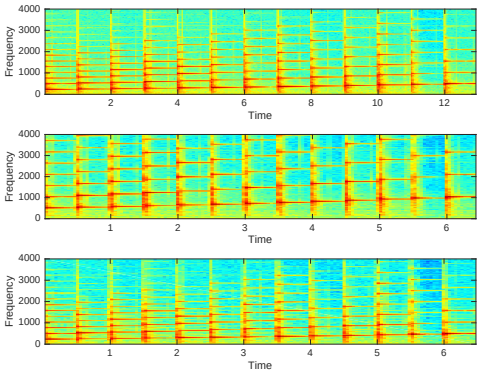
$$\begin{aligned}\phi(k, r') &= \phi(k, r'-1) + 2\pi f_0 \Delta t \\ &= \phi(k, r'-1) + \lfloor \phi_2 - \phi_1 - \Delta\phi_{th} \rceil_{[-\pi,\pi]}\end{aligned} \quad (10)$$

  - 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

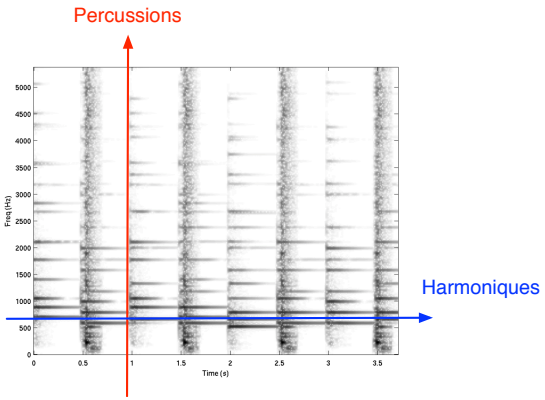
- Ré-échantillonnage du signal pour correction de la longueur par phase-vocoder

Figure 8.24 Resampling of a time stretching algorithm.

2- Séparation de sources

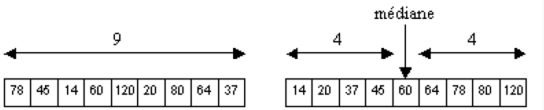
Séparation de la partie percussive et harmonique d'un morceau de musique

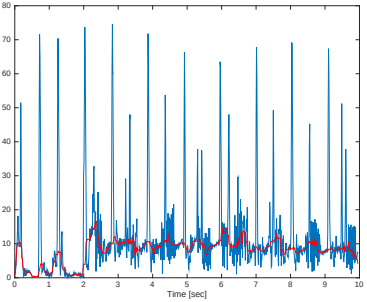
- On considère la TFCT comme l'addition des composantes harmoniques et percussives
  - $X(f,n) = H(f,n) + P(f,n)$
- Morphologie en temps/fréquence des instruments de musique
  - percussions : lignes verticales
  - harmoniques : lignes horizontales

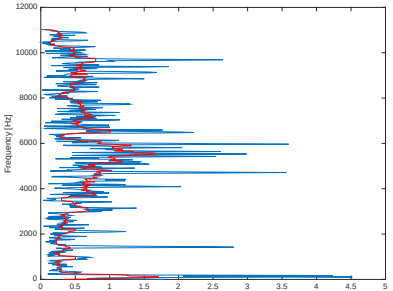


Séparation de la partie percussive et harmonique d'un morceau de musique

- Création d'un **spectrogramme harmonique** $H(f,n)$:
  - pour chaque f on applique un filtrage médian à travers les n de $X(f,n)$
- Création d'un **spectrogramme percussif** $P(f,n)$:
  - pour chaque n on applique un filtrage médian à travers les f de $X(f,n)$
- **Filtrage médian** ?
  - remplace chaque entrée par la valeur médiane de son voisinage
- Valeur médiane
  - tel que 50% des valeurs en-dessous et 50% au-dessus







Séparation de la partie percussive et harmonique d'un morceau de musique

- Création d'un **masque harmonique**

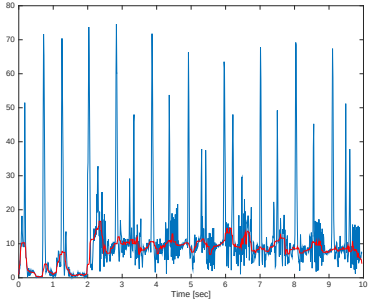
$$M_H(f,n) = \begin{cases} 1 \text{ si } H(f,n) > P(f,n) \\ 0 \text{ sinon} \end{cases} \quad (11)$$

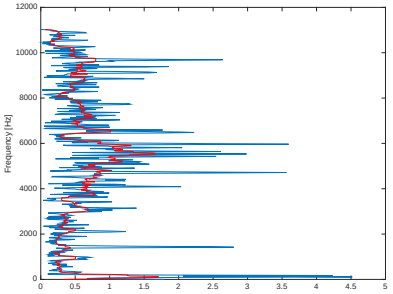
- Création d'un **masque percussif**

$$M_P(f,n) = \begin{cases} 1 \text{ si } P(f,n) > H(f,n) \\ 0 \text{ sinon} \end{cases} \quad (12)$$

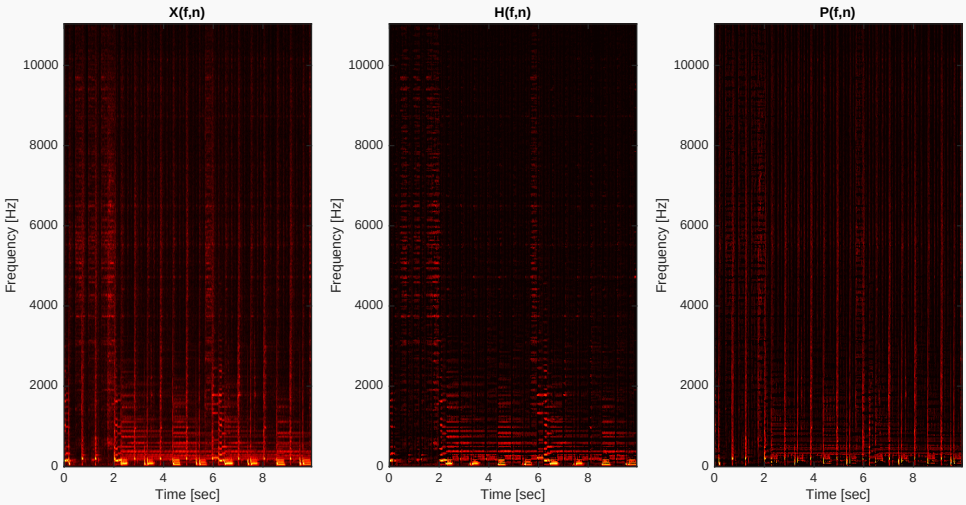
- Re-création de la TFCT

$$\begin{aligned} H(f,n) &= X(f,n) \cdot M_H(f,n) \\ P(f,n) &= X(f,n) \cdot M_P(f,n) \end{aligned} \quad (13)$$



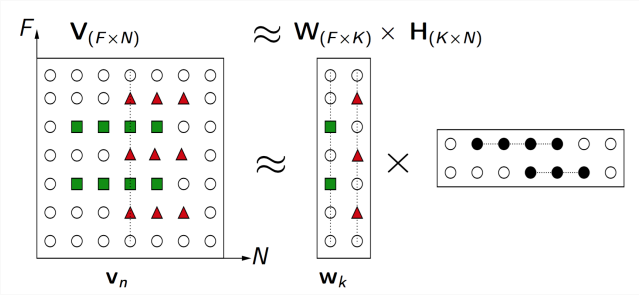


Séparation de la partie percussive et harmonique d'un morceau de musique



Introduction

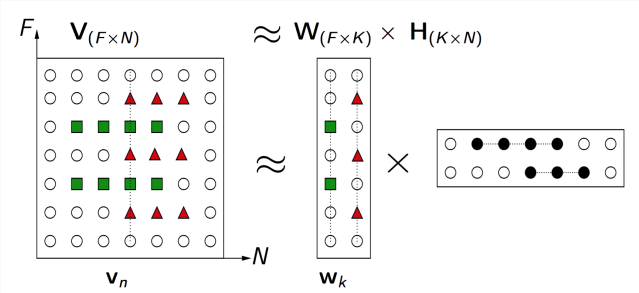
[D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 1999.]



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



source : Cédric Févotte

- Chaque trame n est reconstituée comme l'activation H d'un certain nombre de bases H
  - $V_{(1:F,n)} \simeq \sum_{k=1}^{K} W_{(1:F,k)} H_{(k,n)}$
- Le signal d'une source k est reconstitué comme
  - $V^k_{(1:F,1:N)} = W_{(1:F,k)} H_{(k,1:N)}$

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 :

$$d_{EUC}(x, y) = (x - y)^2 \tag{14}$$

  - Divergence de Kullback-Leibler :

$$d_{KL}(x, y) = x \log \frac{x}{y} - x + y \tag{15}$$

  - Divergense d'Itakura-Saito :

$$d_{IS}(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{16}$$

Dérivation du critère pour la distance Euclidenne

- Non Negative Matrix Factorization

$$\underset{(f,n)}{V} \simeq \underset{(f,k)}{W}\underset{(k,n)}{H} \tag{17}$$

- 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

$$
\begin{aligned}
SSE &= ||V - WH||_F^2 \\
SSE &= (V - WH)^T (V - WH) \\
&= (V^T - H^T W^T)(V - WH) \\
&= V^T V - V^T WH - H^T W^T V + H^T W^T WH \\
&= V^T V - 2V^T WH + H^T W^T WH \\
\frac{\partial sse}{\partial H} &= -2W^T V + 2W^T WH \\
&= 2W^T (WH - V) \\
\frac{\partial sse}{\partial W} &= -2VH^T + 2WHH^T \\
&= -2(V - WH)H^T
\end{aligned}
\tag{18}
$$

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 gradiant :
  - $\frac{\partial sse}{\partial H} = \underbrace{2W^T WH}_{\nabla_+} - \underbrace{2W^T V}_{\nabla_-}$
- Mise à jour de H

$$H \leftarrow H + \eta \cdot [-\text{gradient}]$$
$$H \leftarrow H + \eta \cdot [\underbrace{W^T V}_{\nabla_-} - \underbrace{W^T WH}_{\nabla_+}] \quad (19)$$

Algorithme de descente de gradient

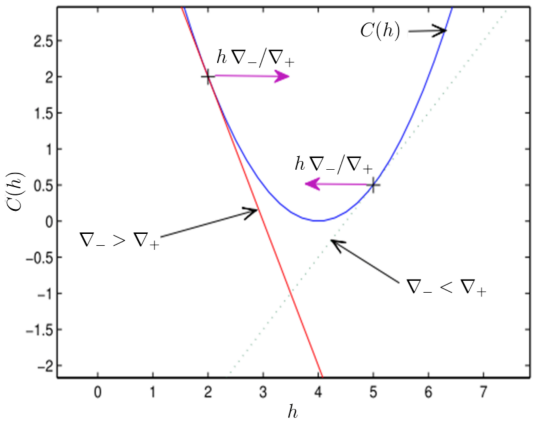
- si on choisit $\eta = \frac{H}{W^T WH}$

$$H \leftarrow H + \frac{H}{W^T WH}(W^T V - W^T WH)$$
$$H \leftarrow H + \frac{H\overbrace{W^T V}^{\nabla_-}}{W^T WH} - H$$
$$H \leftarrow H \cdot \frac{\overbrace{W^T V}^{\nabla_-}}{\underbrace{W^T WH}_{\nabla_+}} \quad (20)$$

Mise à jour multiplicative

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

$$\nabla_h C(h) = \nabla_+ - \nabla_- \quad (21)$$



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 : $V(f,n) = |X(n,f)'|$
- 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

$$W \leftarrow W \cdot \frac{VH^T}{WHH^T} \tag{22}$$

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

$$H \leftarrow H \cdot \frac{W^T V}{W^T WH} \tag{23}$$

  - 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

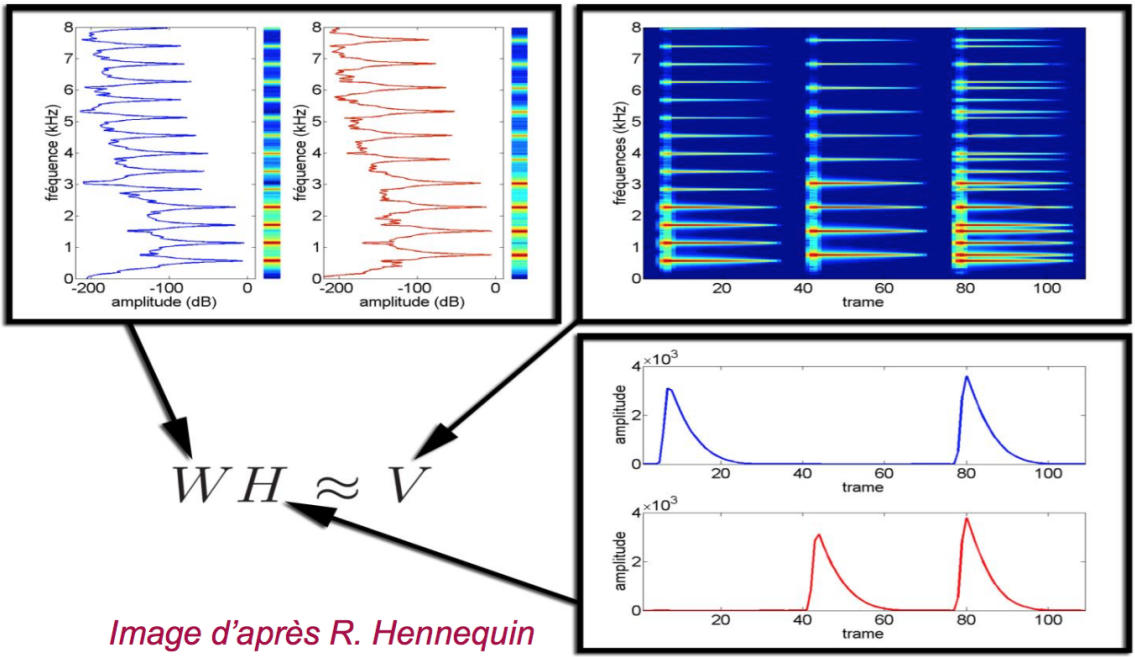
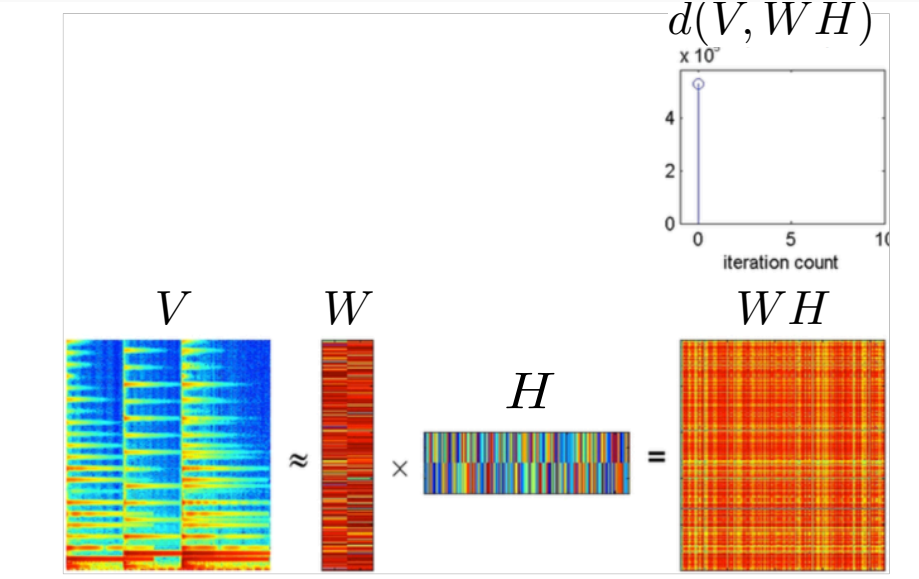


Image d'après R. Hennequin

source : Romain Hennequin

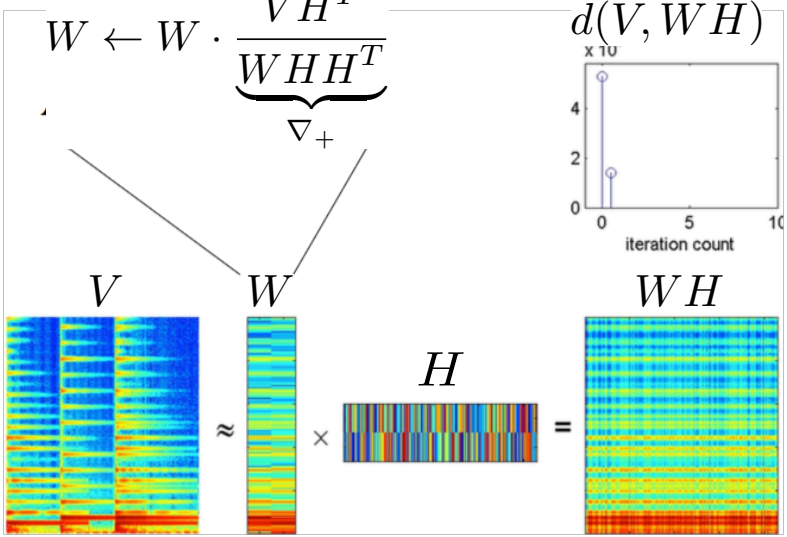
Initialisation



source : Tuomas Virtanen

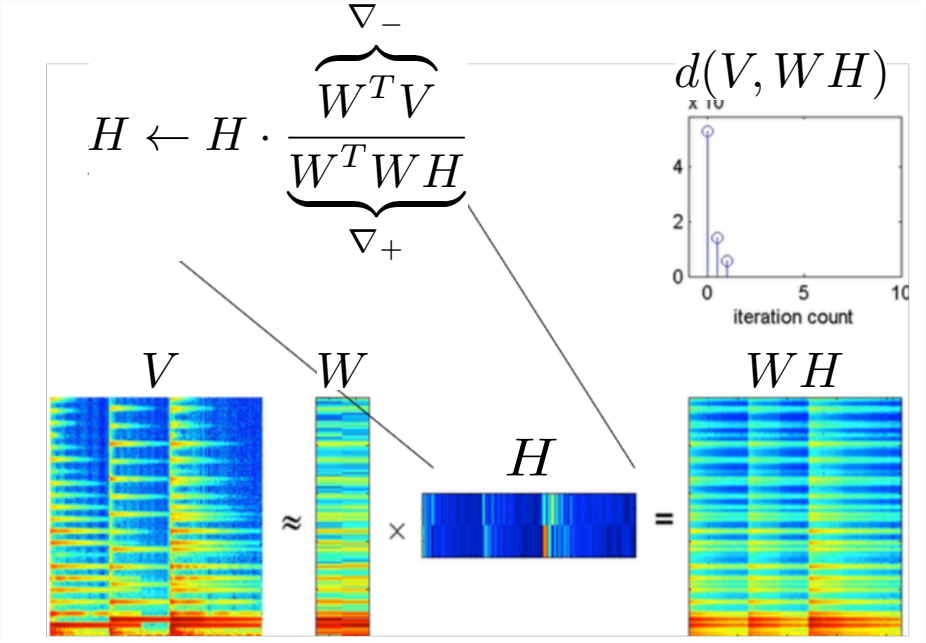
Iteration 1 : Mise à jour de W

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



source : Tuomas Virtanen

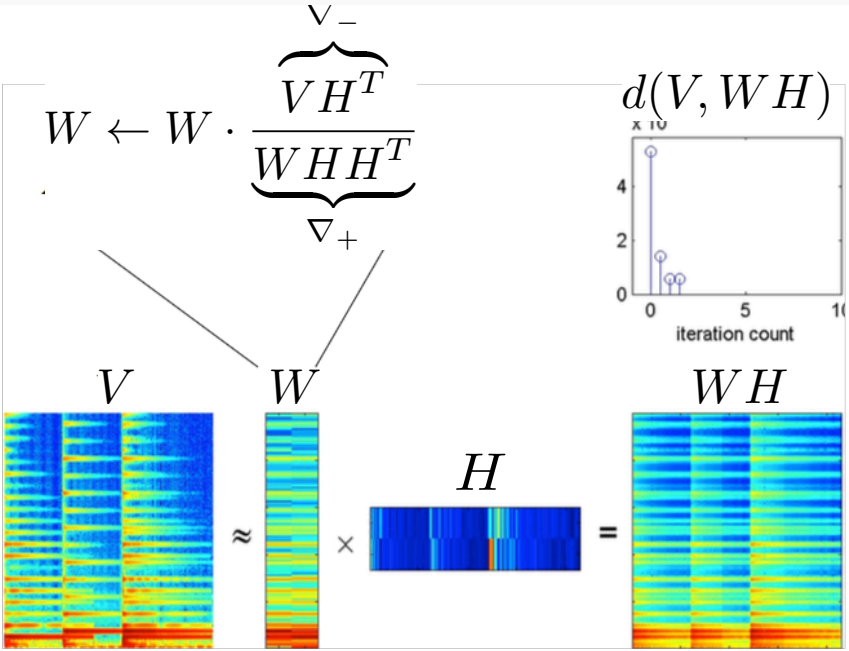
Iteration 1 : Mise à jour de H



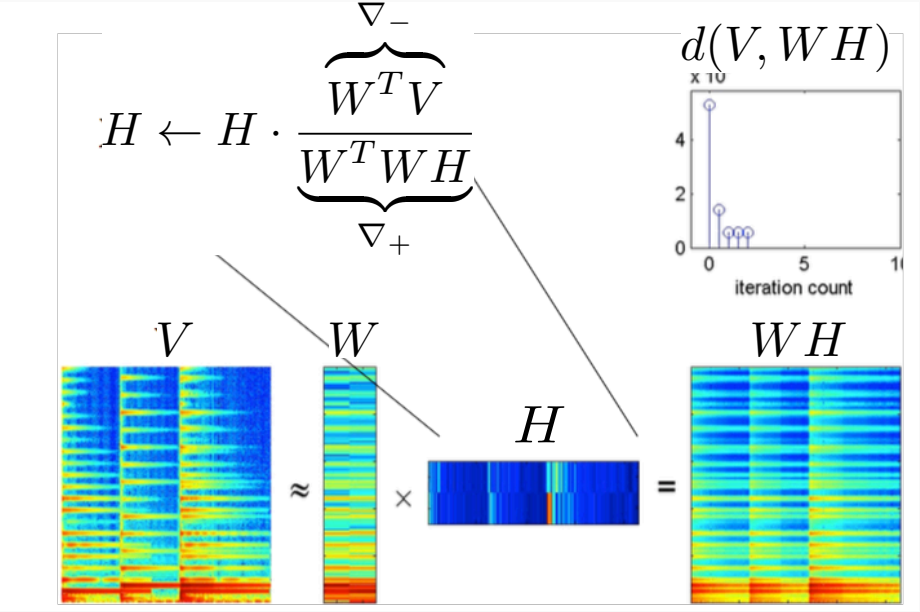
source : Tuomas Virtanen

Iteration 2 : Mise à jour de W

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

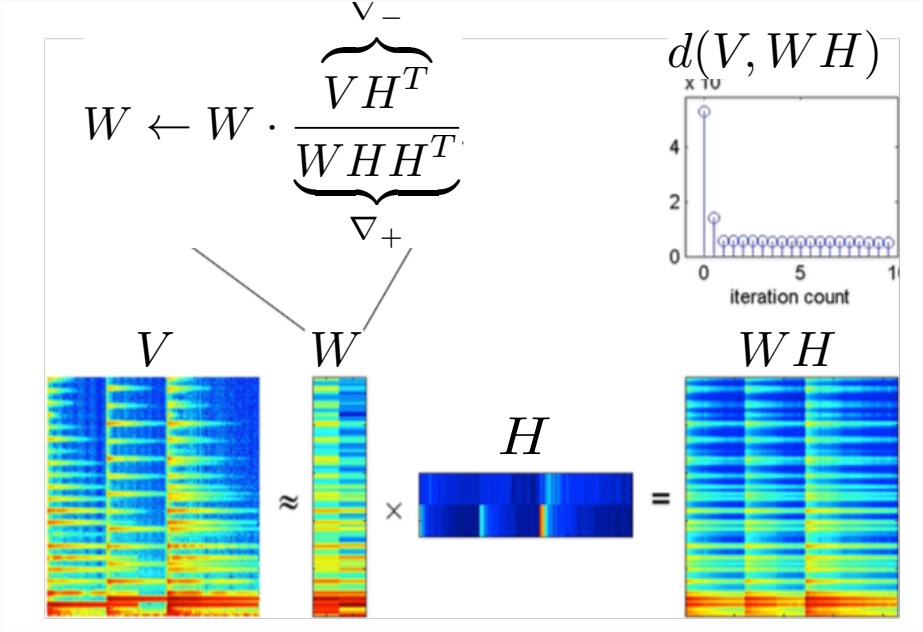


Iteration 2 : Mise à jour de H



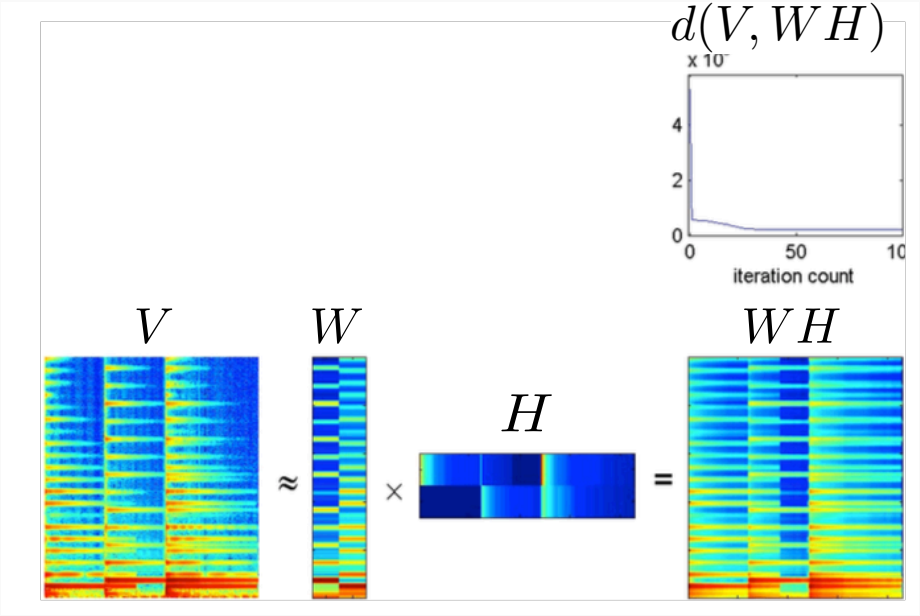
source : Tuomas Virtanen

Iteration 10 : Mise à jour de W



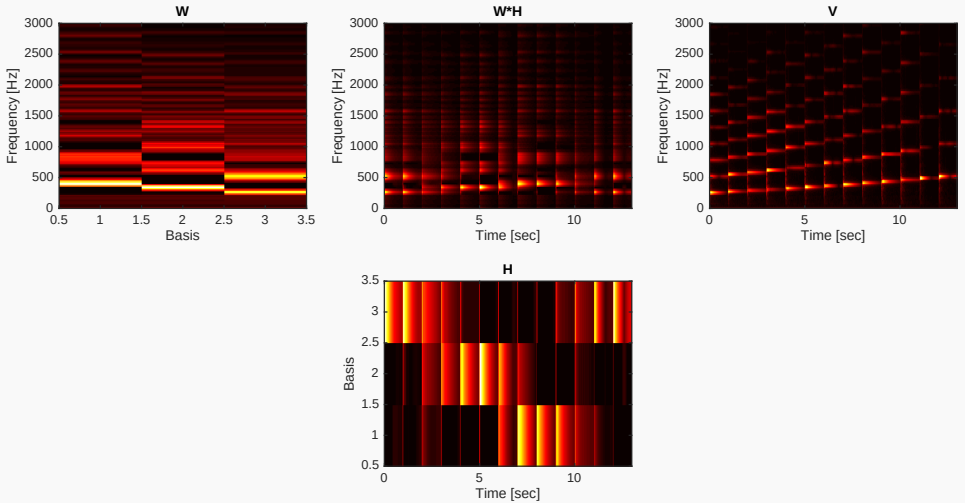
source : Tuomas Virtanen

Iteration 100

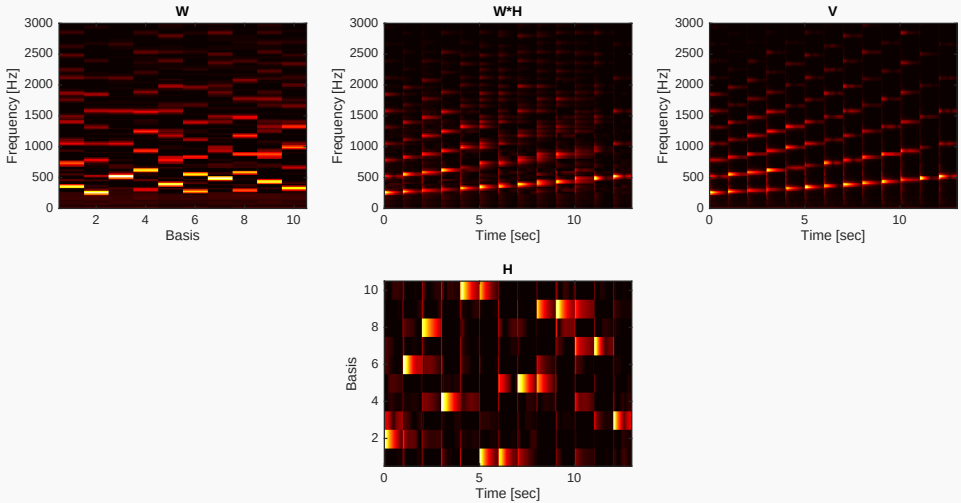


source : Tuomas Virtanen

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



Choix du nombre de bases $K10$ (correcte)



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