ENSEA 3ème SyM (2017-2018)
Traitement du signal audio musical
Partie 1: transformation et séparation du son

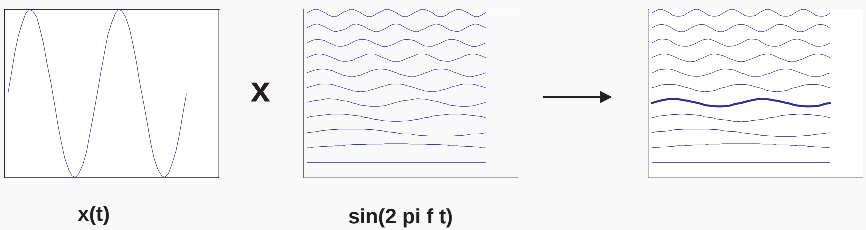
Geoffroy.Peeters@ircam.fr
UMR SMTS 9912 (IRCAM CNRS UPMC)

1- Théorie : Traitement du signal fréquentiel

Transformée de Fourier (temps et fréquences continus)

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

- 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)

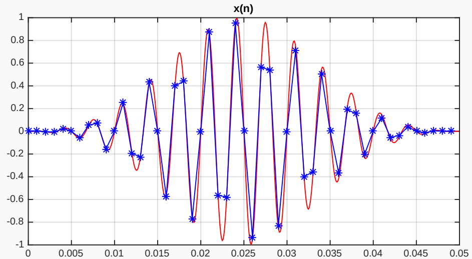
Propriétés	$x(t)$	$X(f)$
Similitude	$x(at)$	$\frac{1}{\lvert a\rvert}X(\frac{f}{\lvert a\rvert})$
Linéarité	$ax(t)+by(t)$	$aX(f)+bY(f)$
Translation	$x(t-t_0)$	$X(f)\exp(-j2\pi f t_0)$
Modulation	$x(t)\exp(j2\pi f_0 t)$	$X(f-f0)$
Convolution	$x(t)\circledast y(t)$	$X(f)Y(f)$
Produit	$x(t)y(t)$	$X(f)\circledast Y(f)$
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
	$x^\star(t)$	$X^\star(f)$

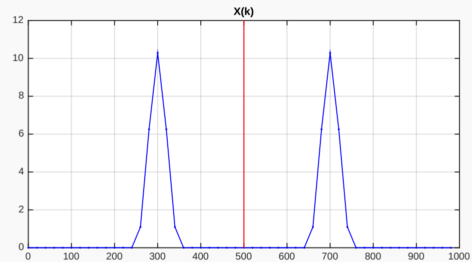
1- Théorie : Traitement du signal fréquentiel
Transformée de Fourier (temps et fréquences discrets)

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

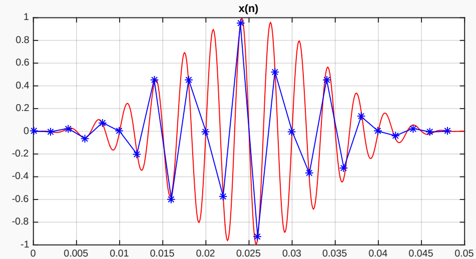
- Variables :
  - m le numéro d'**échantillon**
  - k les **fréquences discrètes**
- Fréquence d'échantillonnage (sampling rate) sr
  - sr définit à quelle fréquence le signal temporel va être échantillonné
  - Exemple :
    - Compact Disc $sr = 44100$ Hz
    - La distance temporelle entre deux échantillons (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** : $f_{Nyquist} = \frac{sr}{2} > f_{\max}$

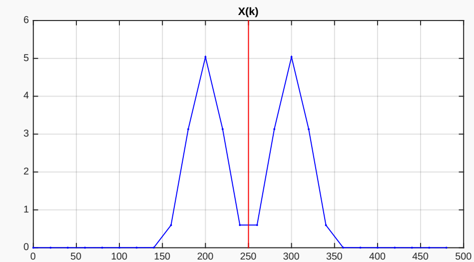
$f_{\max} = 300$, $sr = 1000$





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



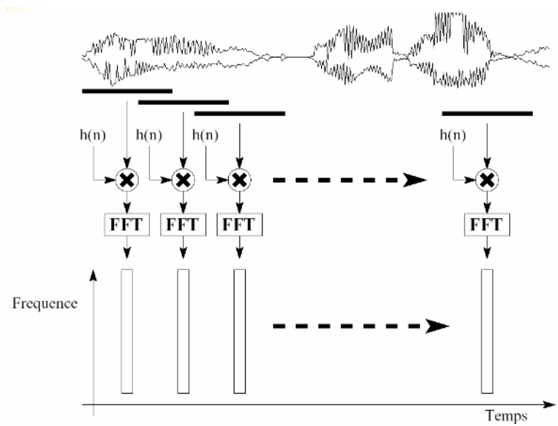


$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \qquad \forall k \in [0,N]$$

- 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

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

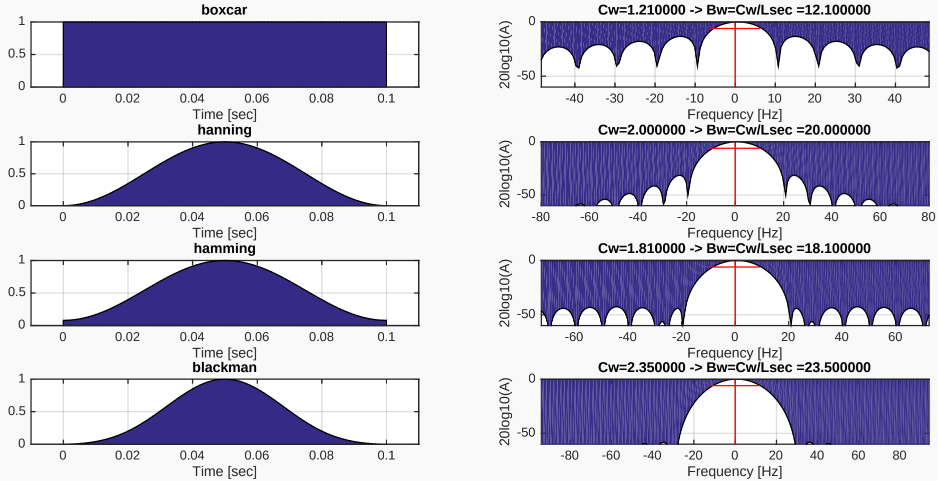
Fenêtre de pondération $w(t)$

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

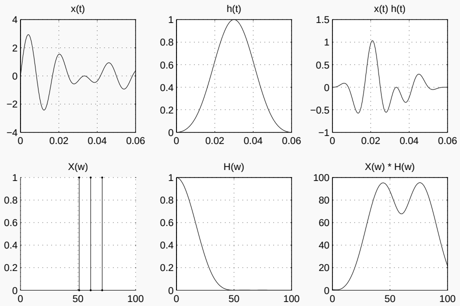
Influence du type de la fonction



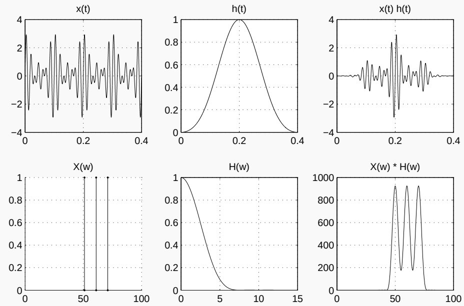
Choix de la longueur temporelle L :

- 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.

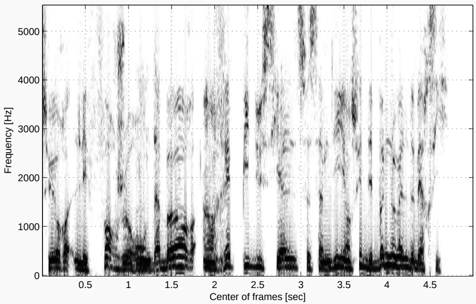
Influence de la longueur temporelle L ($L = 0.06$s.)



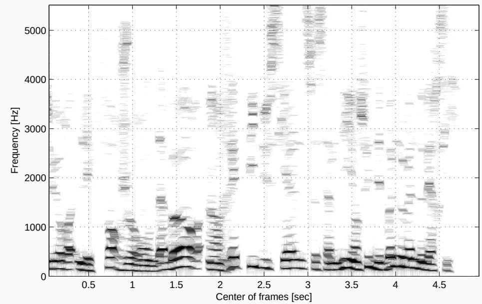
Influence de la longueur temporelle L ($L = 0.4$s.)



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

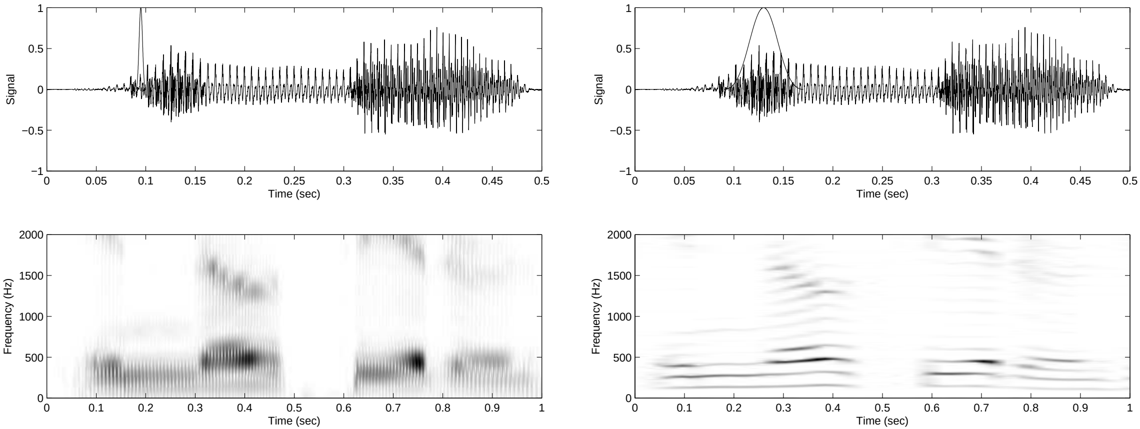


Influence de la longueur temporelle L ($L = 0.1s.$)



Paradoxe temps/ fréquence

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



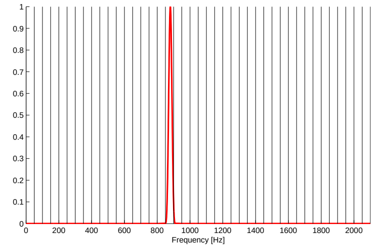
- Comme résoudre ce problème ?
  - Utiliser d'autres tranformées que celle de Fourier

Transformée à Q-Constant (CQT)

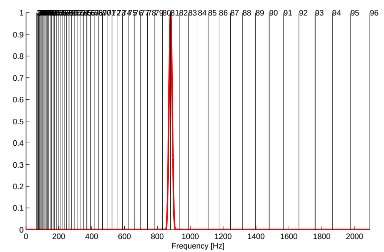
- La DFT
  - Définition : **La précision fréquentielle** : $\Delta f = \frac{sr}{N}$
    - c'est le pas d'échantillonnage du spectre
    - elle dépend de la taille de la DFT : N
    - on peut l'augmenter en augmentant N
  - Définition : **La résolution fréquentielle** : $Bw = \frac{Cw}{L}$
    - c'est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

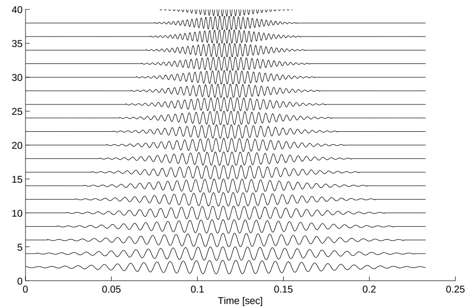
Transformée à Q-Constant

[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

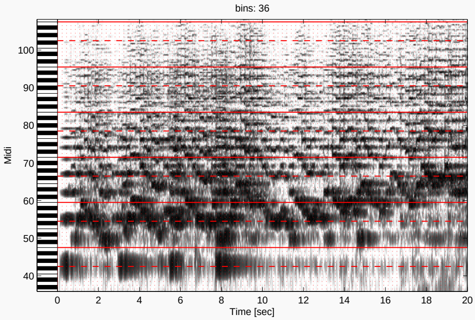
$$Q = \frac{f_k}{Bw} = \frac{f_k}{Cw/L} = \frac{f_k \cdot L}{Cw}$$

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

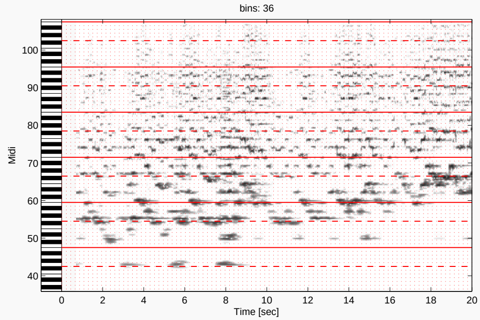


1- Théorie : Traitement du signal fréquentiel
Transformée à Q-Constant (CQT)

Exemples (en utilisant la DFT)

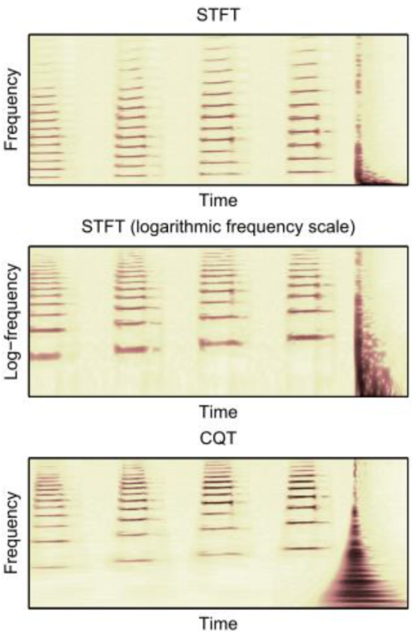


Exemples (en utilisant la CQT)



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

1- Théorie : Traitement du signal fréquentiel

1.5- Deux interprétations de la TFCT

Deux interprétations de la TFCT

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

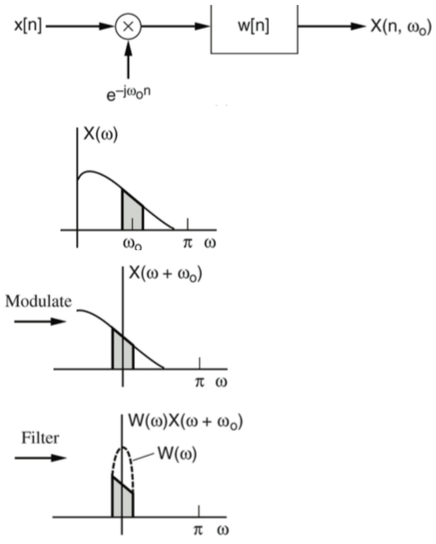
$$X(f,n) = \sum_{m=-\inf}^{+\inf} x(m)w(n-m)e^{-j2\pi fm}$$

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

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

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

- 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 $w(m)$



source : Patrick J. Wolfe, 2009

Déduction du pas d'avancement

Deux interprétations de la TFCT

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

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

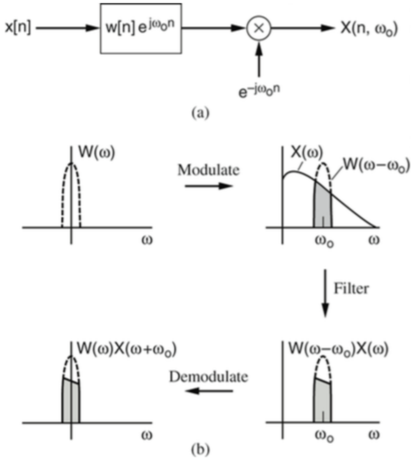
$$X(f_0,n) = \sum_{m=-\inf}^{+\inf} x(m)\underbrace{w(n-m)e^{j2\pi f_0(n-m)}}_{w_0(n-m)} e^{-j2\pi f_0 n}$$

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

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

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

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



source : Patrick J. Wolfe, 2009

1.6- Reconstruction du signal par addition/ recouvrement (TFTC inverse)

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

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

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

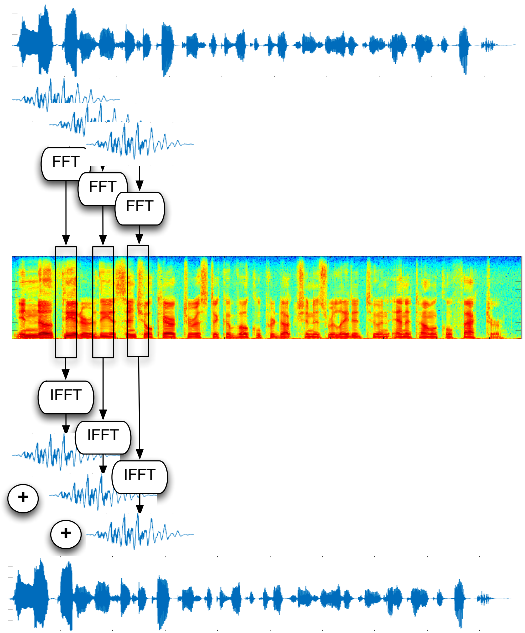
$$\frac{1}{N}\sum_{k=0}^{N-1} X(k,n)e^{+j2\pi \frac{k}{N} m} = x(m)w(n-m)$$

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

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

- Si on note $n = rl$
  - $r \in \mathbb{N}$ le numéro de la trame d'analyse
  - l le pas d'avancement en échantillons

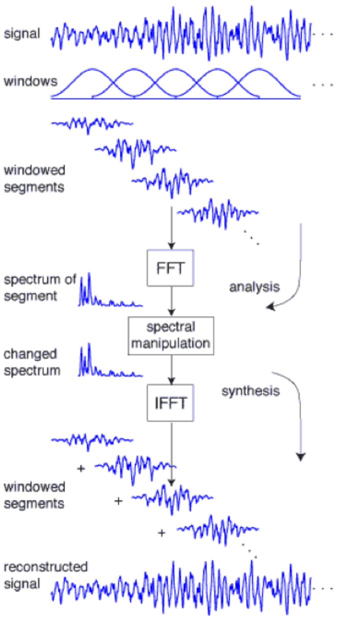
$$
\begin{aligned}
y(m, rl) &= \frac{1}{N} \sum_{k=0}^{N-1} X(k, rl) e^{+j2\pi \frac{k}{N} m} \\
&= x(m) w(rl - m) \\
y(m) &= \sum_r y(m, rl) \\
&= \sum_r x(m) w(rl - m) \\
&= x(m) \sum_r w(rl - m) \\
x(n) &= \frac{\sum_r y(m, rl)}{\sum_r w(rl - n)}
\end{aligned}
$$



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

- Si on note $n = rl$
  - $r \in \mathbb{N}$ le numéro de la trame d'analyse
  - l le pas d'avancement en échantillons

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



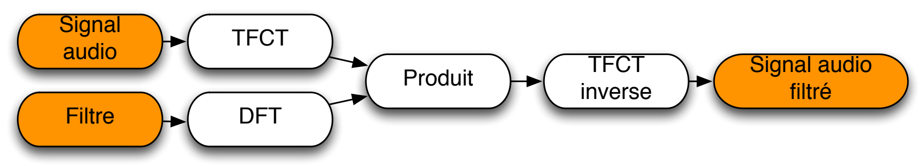
1- Théorie : Traitement du signal fréquentiel

1.7- Application : filtrage constant au cours du temps

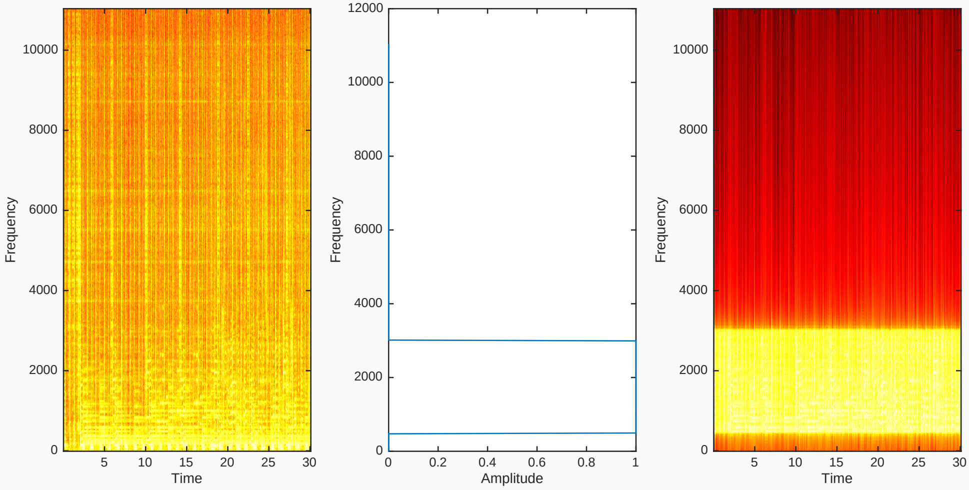
Application : filtrage constant au cours du temps

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

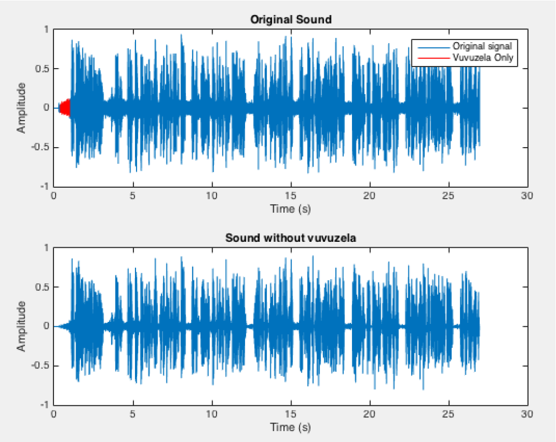


1.8- Application : débruitage par soustraction spectrale

Application : débruitage par soustraction spectrale

[S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. Acoustics, Speech and Signal Processing, IEEE Transactions on, 27(2) :113–120, 1979.]

- 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 : $X(\omega) = S(\omega) + N(\omega)$
- **Méthode**
  - On cherche un filtre fréquentiel $H(\omega)$ permettant de retirer le bruit additif
  - Amplitude de ce filtre
    - $=$ valeur moyenne de $|N(\omega)|^2$ calculée sur un segment ne contenant que du bruit
    - $\mu(\omega) = E\{|N(\omega)|\}$
  - Phase
    - $=$ la phase de X : $\phi_x(\omega)$

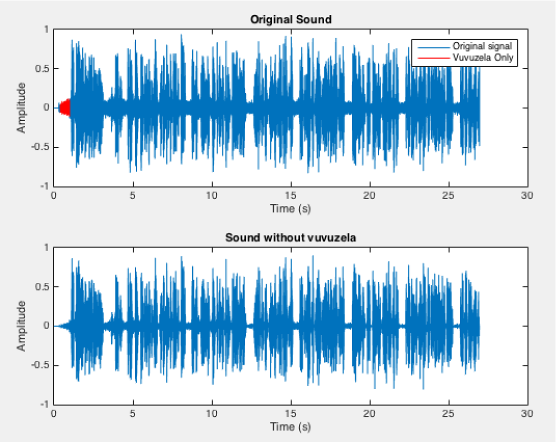


Application : débruitage par soustraction spectrale

- Soustraction :

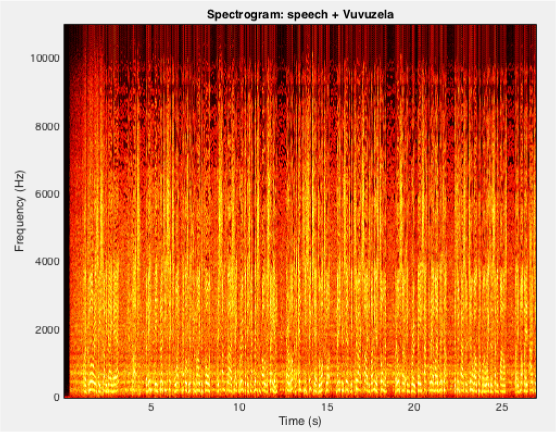
$$\begin{aligned}\hat{S}(\omega) &= [|X(\omega)| - \mu(\omega)]\, e^{j\phi_x} \\ &= \underbrace{\left[1 - \frac{\mu(\omega)}{|X(\omega)|}\right]}_{H(\omega)} |X(\omega)| e^{j\phi_x} \\ &= H(\omega)X(\omega)\end{aligned}$$

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

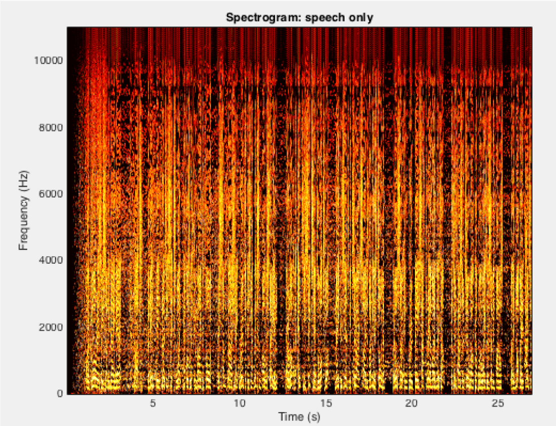


1- Théorie : Traitement du signal fréquentiel
Application : débruitage par soustraction spectrale

Spectrogramme speech+noise



Spectrogramme speech



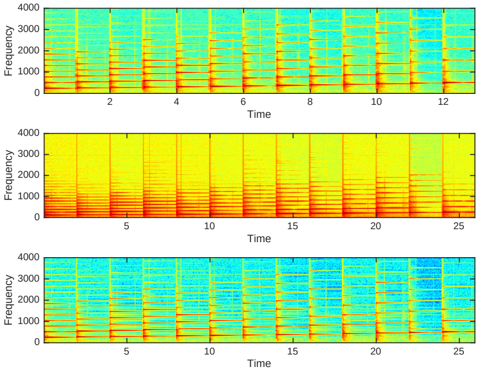
1- Théorie : Traitement du signal fréquentiel

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

1- Théorie : Traitement du signal fréquentiel
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)



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

Objectif

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

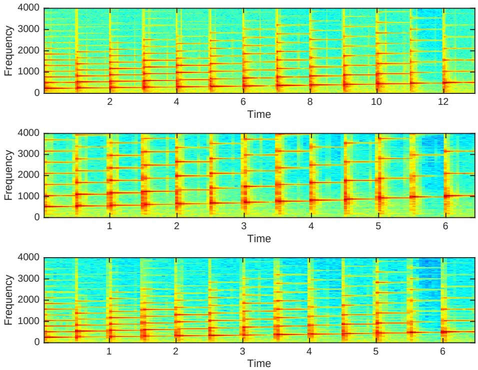
1- Théorie : Traitement du signal fréquentiel
Application : dilatation/ contraction du temps par vocodeur de phase

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**

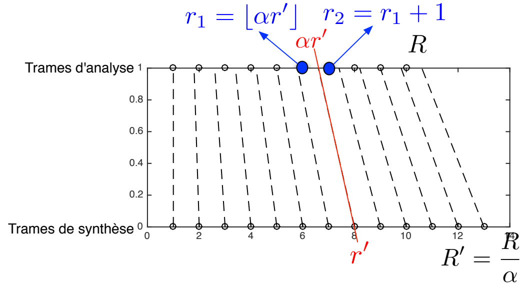


[haut] : signal original, [milieu] $\alpha > 1$ par ré-échantillonnage, [bas] : $\alpha > 1$ par vocodeur de phase

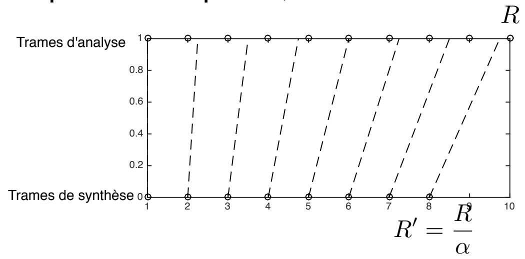
Le vocodeur de phase

- **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
  - $r_2 = r_1 + 1$

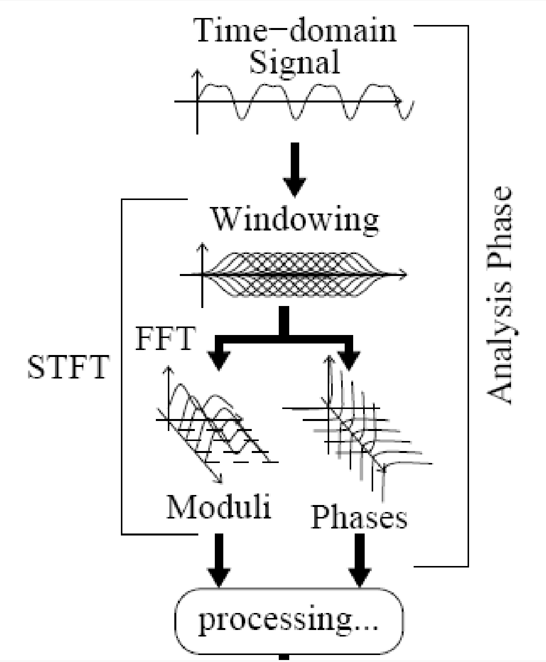
Dilation du temps $\alpha < 1$



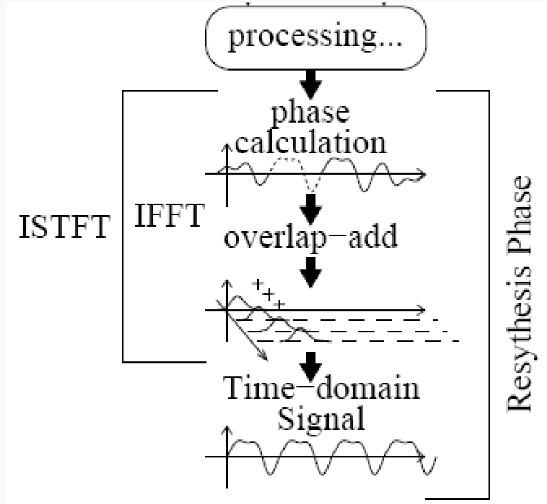
Compression du temps $\alpha > 1$



Le vocoder de phase : analyse



Le vocoder de phase : synthèse



1- Théorie : Traitement du signal fréquentiel
Application : dilatation/ contraction du temps par vocodeur de phase

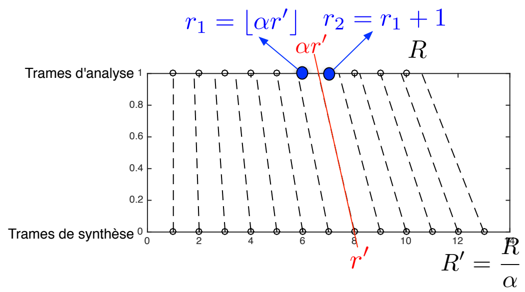
Le vocodeur de phase
1) Spectre d'amplitude

- 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 $\Delta = \alpha r' - r_1$

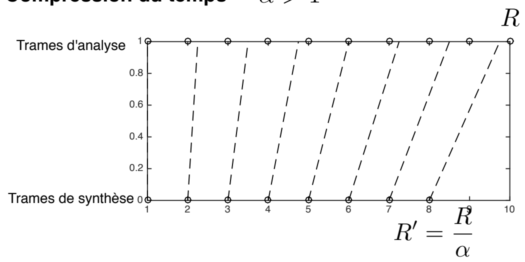
2) Spectre de phase

- **c'est plus compliqué ! ! !**

Dilation du temps $\alpha < 1$

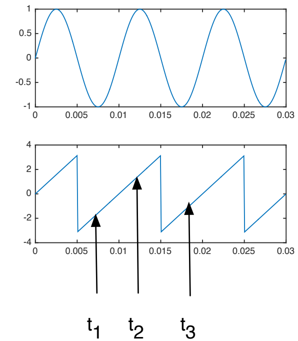


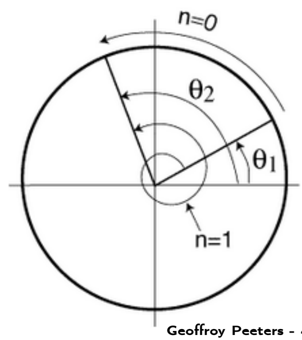
Compression du temps $\alpha > 1$



La phase et la fréquence instantanée

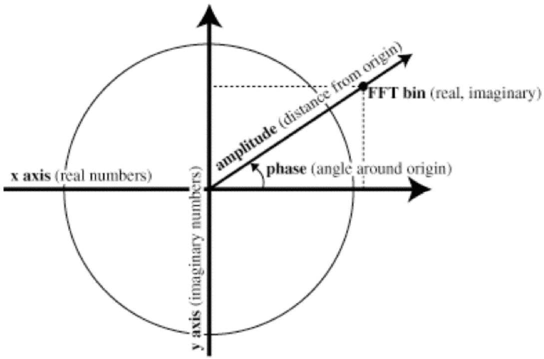
- Considérons un signal formé d'une sinusoide à la fréquence f_0 :
  - $x(t) = \sin(2\pi f_0 t)) = \sin(\phi(t))$
- Entre les instants t_1 et t_2, sa phase a "tournée" de $\phi(t_1)$ à $\phi(t_2)$
- Puisqu'il s'agit d'une sinusoide pure, elle a tournée de
  - $\phi(t_2) = \phi(t_1) + 2\pi f_0 (t_2 - t_1)$
- 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)

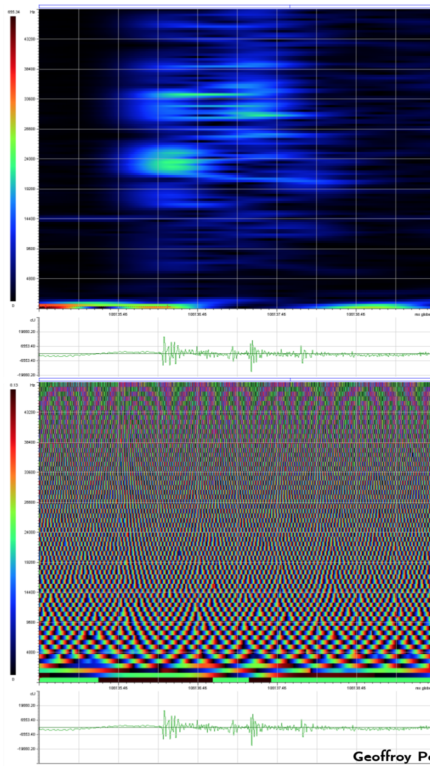
- Pour chaque trame n et fréquence k la TFTC est un nombre complexe
  - $X(k,n) = \sum_m x(m)w(n-m)e^{-j2\pi \frac{k}{N}m}$
- Il peut se décomposer en amplitude (module) et phase :
  - $X(k,n) = A(k,n)e^{j\phi(k,n)}$



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 $(n-1) \rightarrow n$.

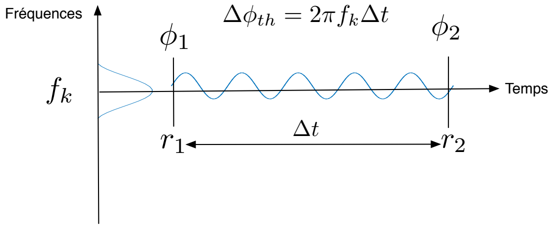


Le vocodeur de phase
2) Spectre de phase

- 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

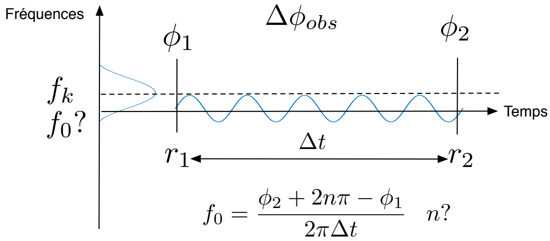
$$\begin{aligned}\phi(k, r') &= \phi(k, r'-1) + \Delta\phi_{th} \\ &= \phi(k, r'-1) + 2\pi f_k \Delta t\end{aligned}$$

  - 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 $f_0 \simeq f_k$

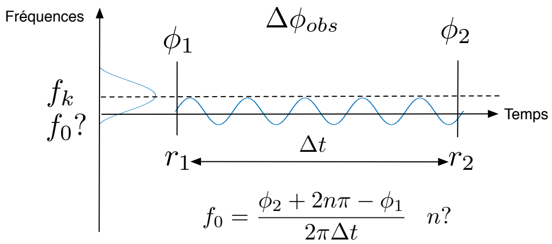
n tel que $\min_n |f_0 - f_k|$

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

$$\min_n |\phi_2 + n2\pi - \phi_1 - 2\pi\Delta t f_k|$$

$$\min_n |\phi_2 + n2\pi - \phi_1 - \Delta\phi_{th}|$$

- 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 $\Delta\phi_{th}$

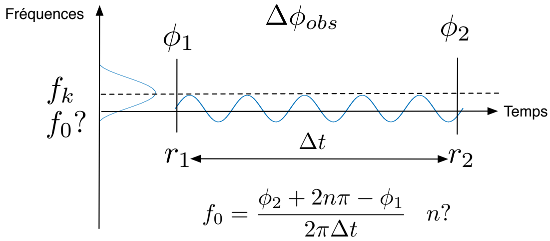


Le vocodeur de phase : spectre de phase

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

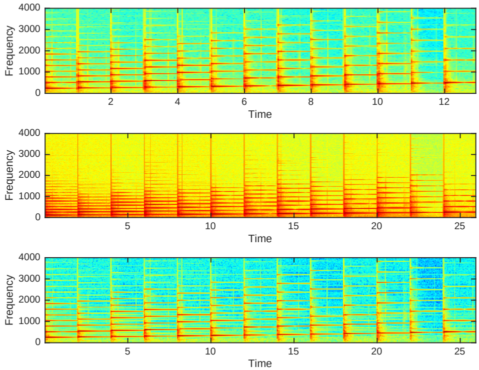
$$\begin{aligned}\phi(k, r') &= \phi(k, r'-1) + 2\pi f_0 \Delta t \\ &= \phi(k, r'-1) + \phi_2 + n2\pi - \phi_1\end{aligned}$$

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

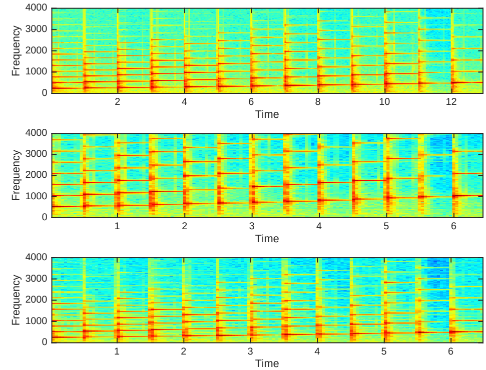


1- Théorie : Traitement du signal fréquentiel

Application : dilatation/ contraction du temps par vocodeur de phase



[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

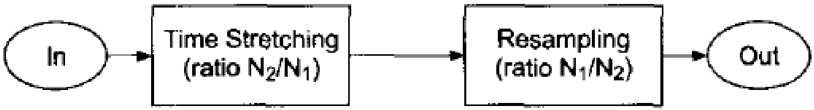


Figure 8.24 Resampling of a time stretching algorithm.

2- Séparation de sources

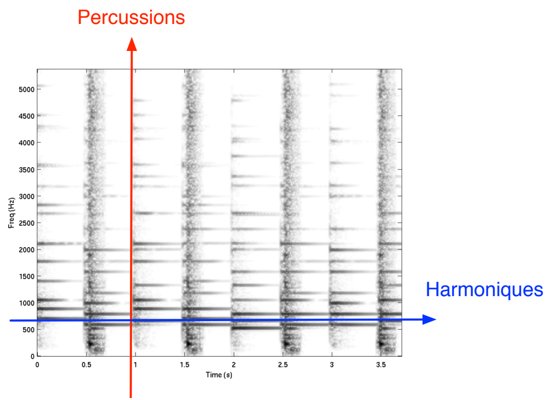
2- Séparation de sources

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

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

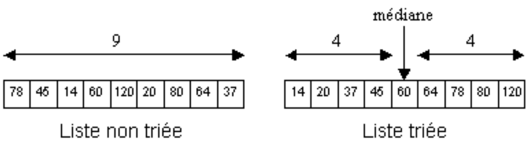
[D. Fitzgerald. Harmonic/percussive separation using median filtering. In Proc. of DAFx, Graz, Austria, 2010.]

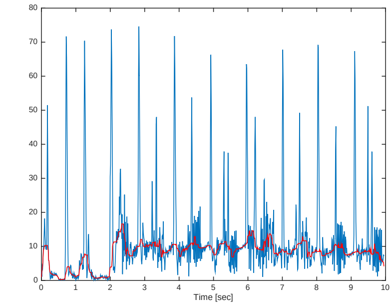
- **HPSS** : Harmonic-Percussive Source Separation
- On considère la TFCT comme le résultat de l'addition de composantes harmoniques et percussives
  - $X(f,n) = H(f,n) + P(f,n)$
- Morphologie en temps/fréquence des instruments de musique
  - composantes harmoniques :
    - lignes horizontales
  - composantes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



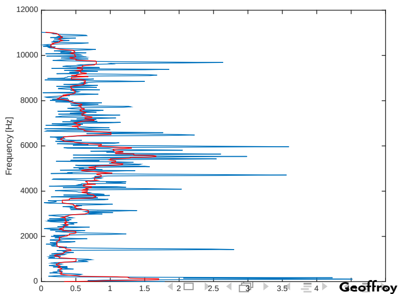
Séparation de la partie harmonique et percussive d'un morceau de musique

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





Filtrage median de $X(f,n)$ à travers les n



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

- Création d'un **masque harmonique**

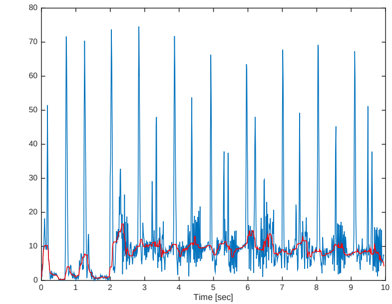
$$M_H(f,n) = \begin{cases} 1 \text{ si } H(f,n) > P(f,n) \\ 0 \text{ sinon} \end{cases}$$

- Création d'un **masque percussif**

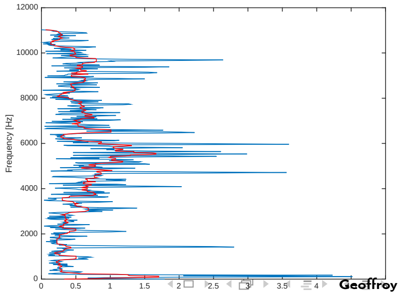
$$M_P(f,n) = \begin{cases} 1 \text{ si } P(f,n) > H(f,n) \\ 0 \text{ sinon} \end{cases}$$

- Re-création de la TFCT

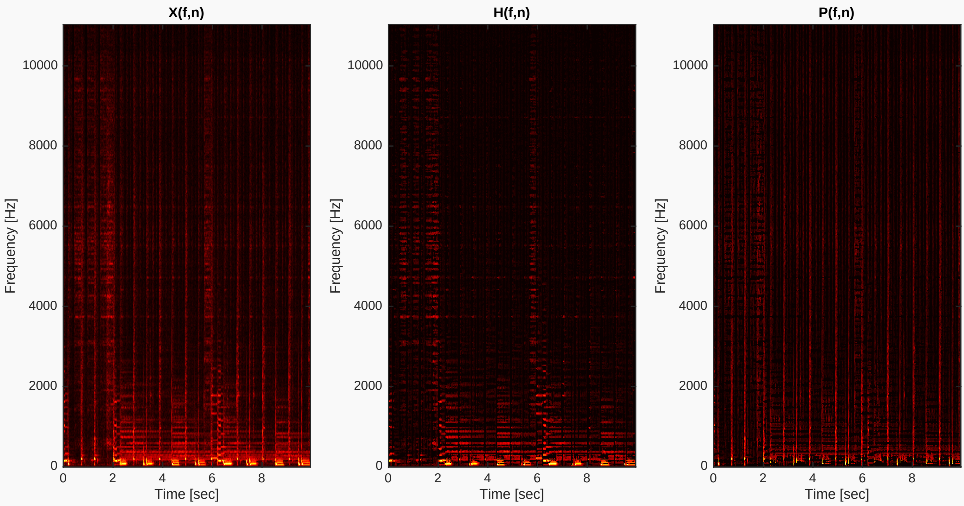
$$H(f,n) = X(f,n) \cdot M_H(f,n)$$
$$P(f,n) = X(f,n) \cdot M_P(f,n)$$



Filtrage median de $X(f,n)$ à travers les n



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



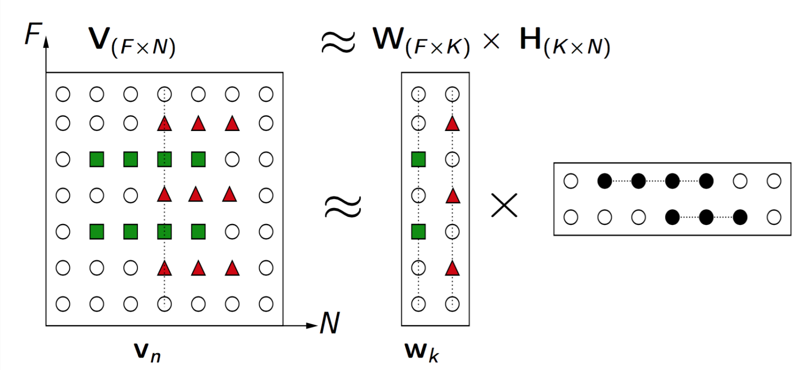
2- Séparation de sources

2.2- 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

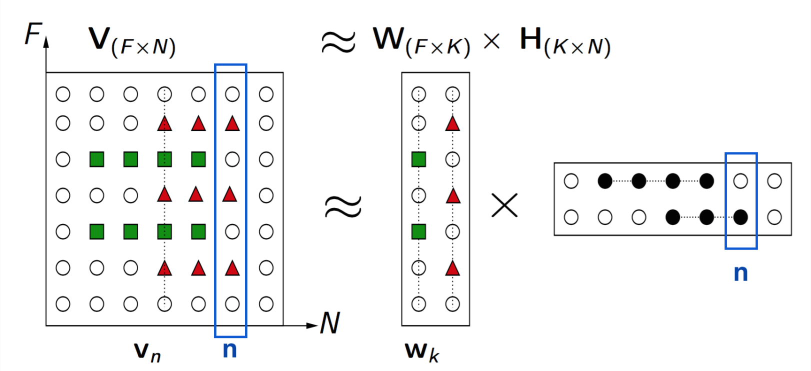


source : Cédric Févotte

- $V_{(F,N)} \simeq W_{(F,K)} H_{(K,N)}$
  - $V_{(F,N)}$: 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

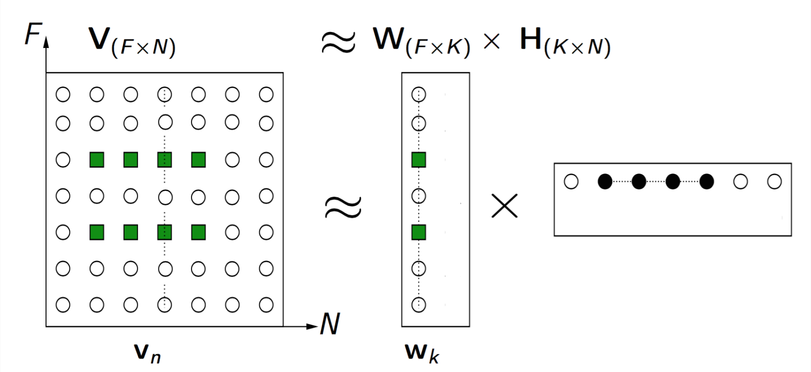
- Chaque trame **n** est reconstituée comme l'**activation** H d'un certain nombre de **bases** W
  - $V_{(1:F,\mathbf{n})} \simeq \sum_{k=1}^{K} W_{(1:F,k)} H_{(k,\mathbf{n})}$



source : Cédric Févotte

Introduction

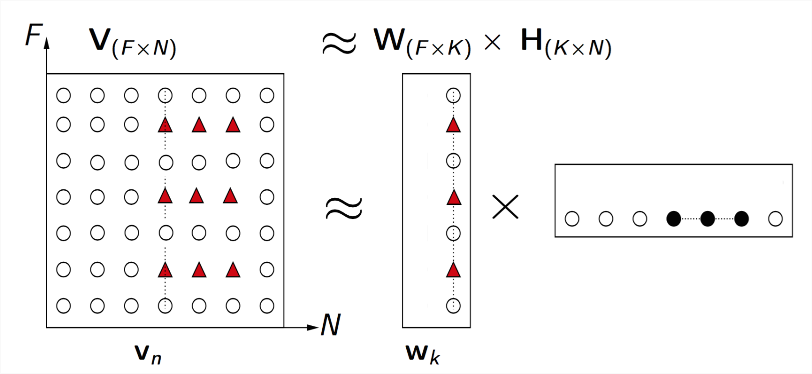
- Le signal d'une **source** k est reconstitué comme
  - $V^{k}_{(1:F,1:N)} = W_{(1:F,\mathbf{k=1})} H_{(\mathbf{k=1},1:N)}$



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)

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

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

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

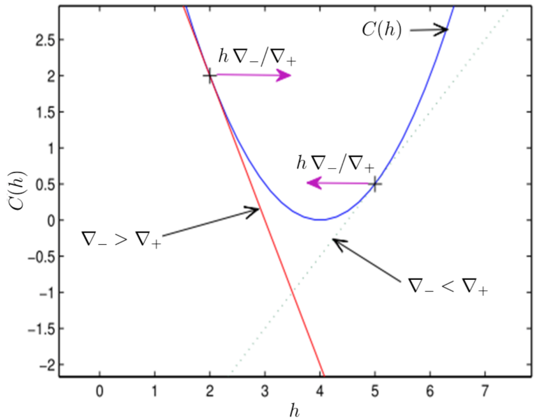
- si on choisit $\eta = \frac{H}{W^T WH}$

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

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

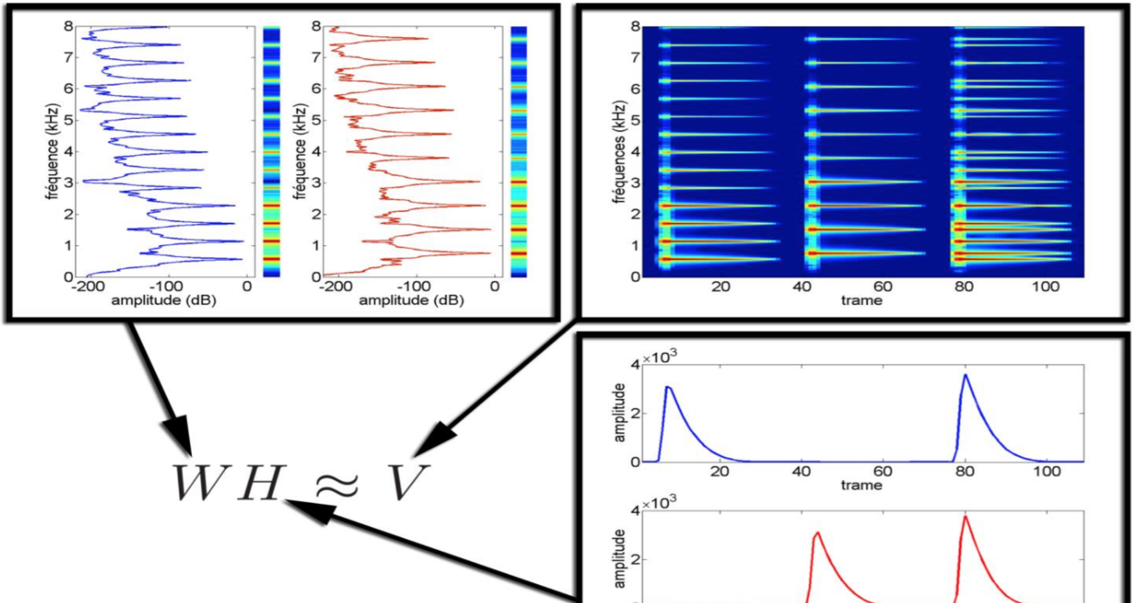
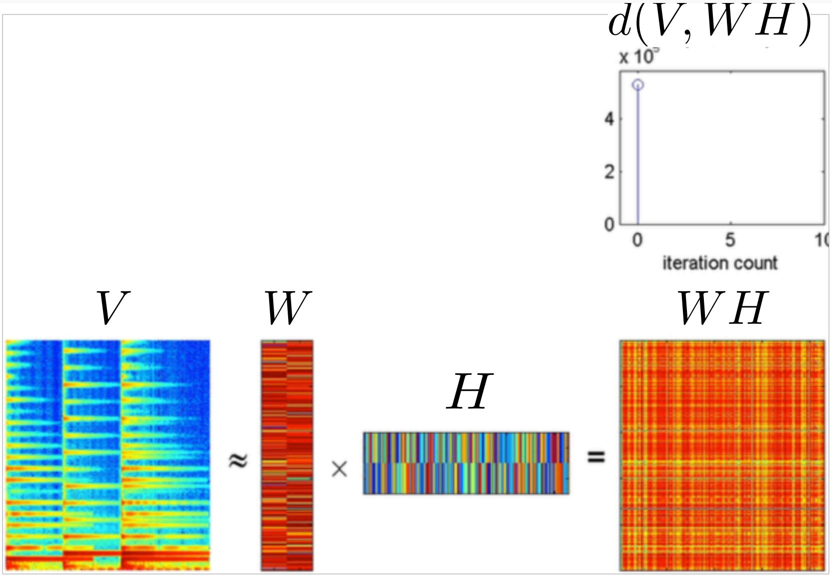


Image d'après R. Hennequin

source : Romain Hennequin

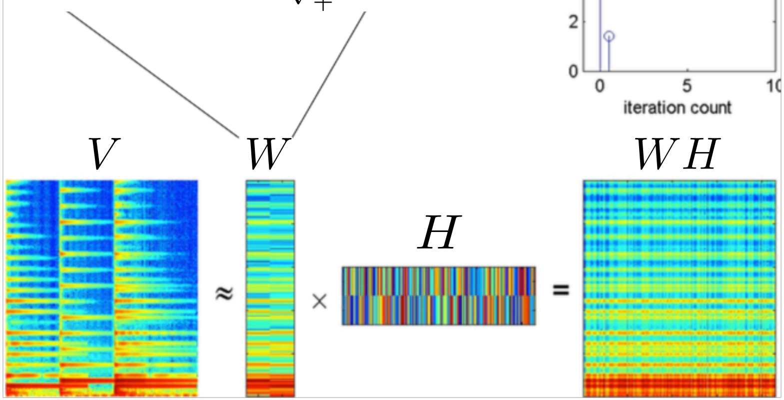
Initialisation



source : Tuomas Virtanen

Iteration 1 : Mise à jour de W

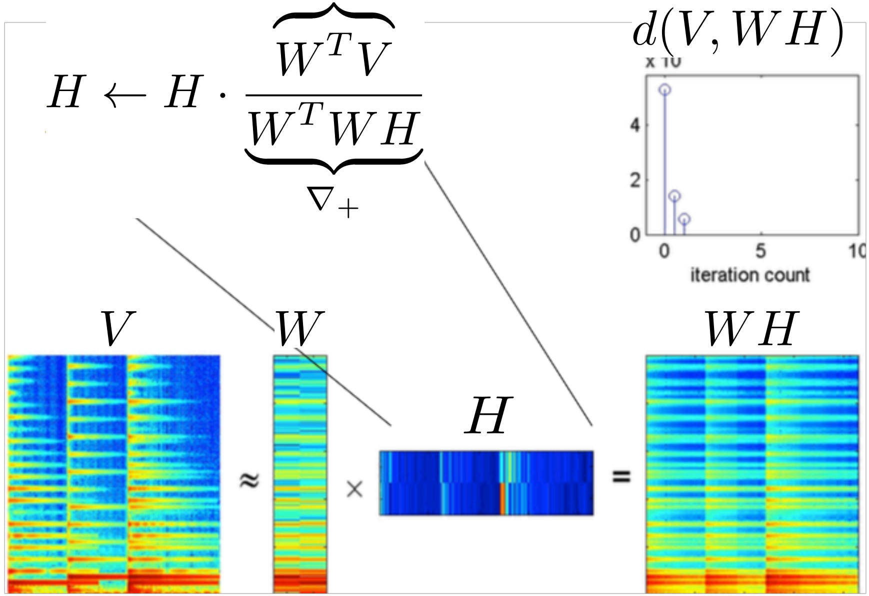
$$W \leftarrow W \cdot \frac{\overbrace{VH^T}^{\nabla_-}}{\underbrace{WHH^T}_{\nabla_+}}$$



source : Tuomas Virtanen

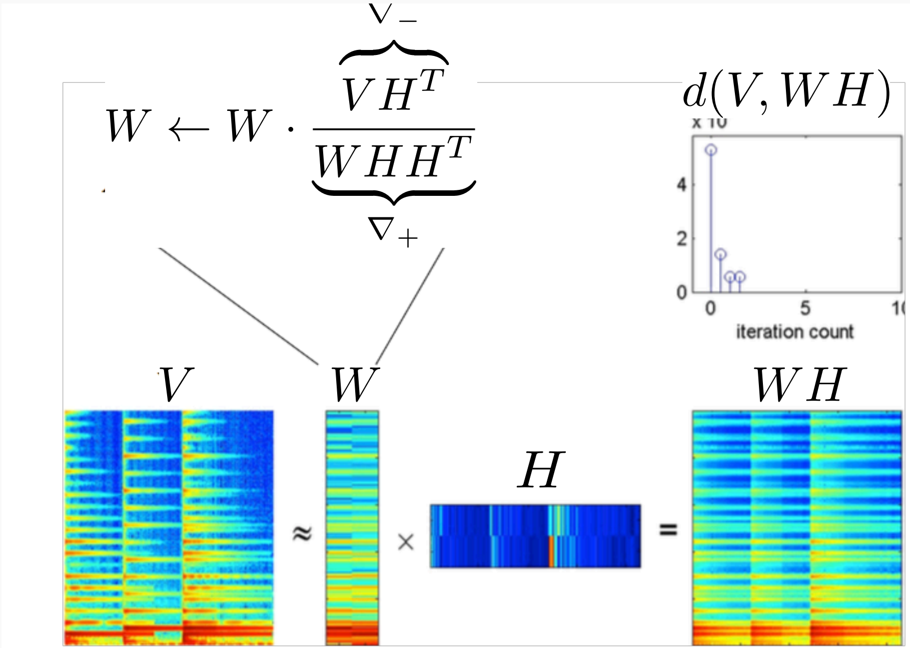
Iteration 1 : Mise à jour de H

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



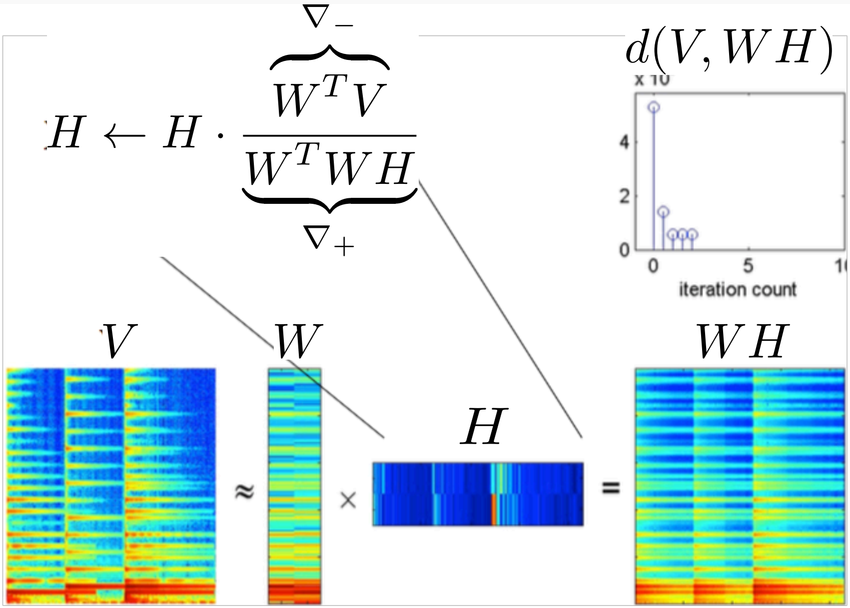
source : Tuomas Virtanen

Iteration 2 : Mise à jour de W



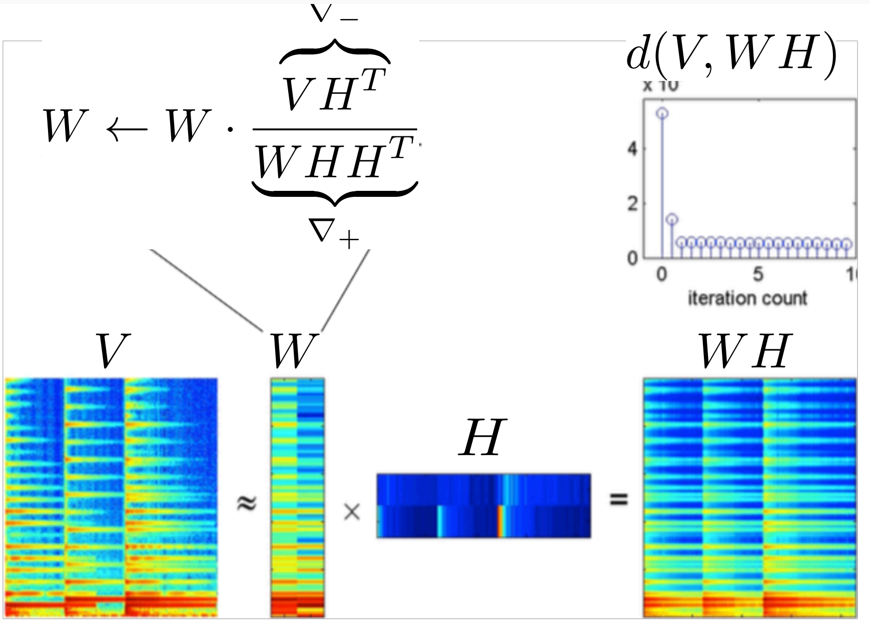
source : Tuomas Virtanen

Iteration 2 : Mise à jour de H



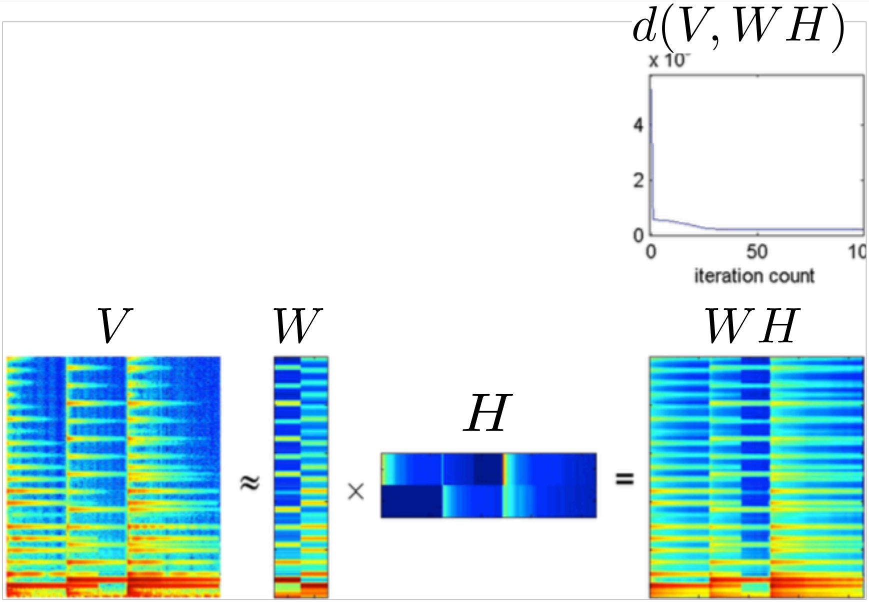
source : Tuomas Virtanen

Iteration 10 : Mise à jour de W



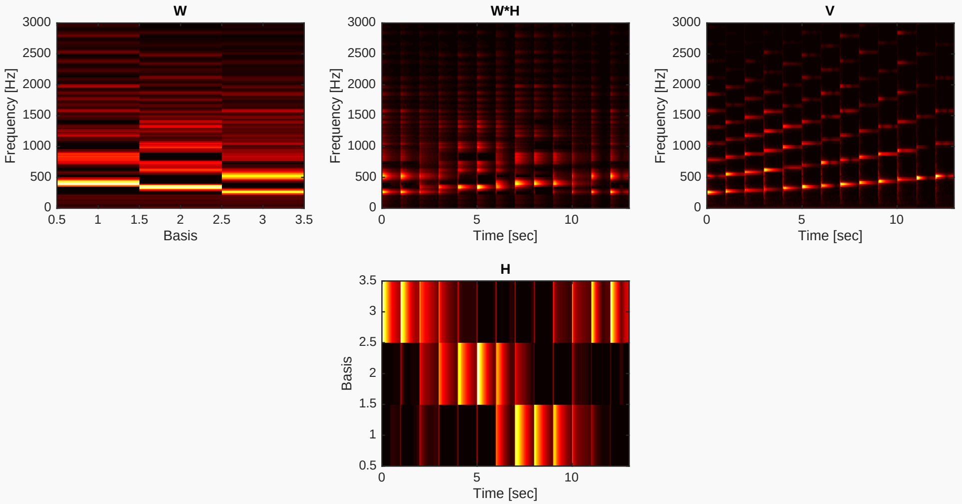
source : Tuomas Virtanen

Iteration 100

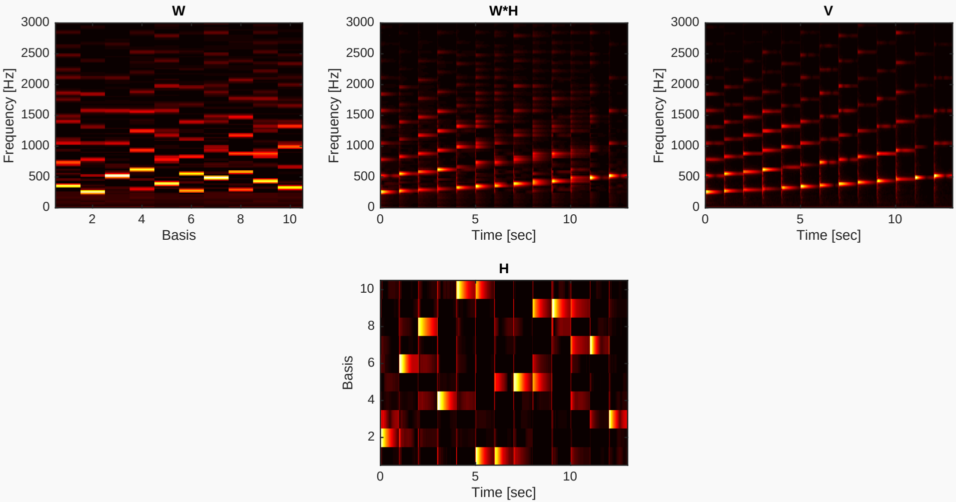


source : Tuomas Virtanen

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)