

# Master 2 ATIAM

## Estimation de fréquences fondamentales multiples

Geoffroy Peeters

Note: ce cours est très largement inspiré de celui de Gael Richard

LTCI, Télécom Paris, IP Paris



## **1. Introduction**

1.1 What is pitch ?

## **2. Utilisation de modèles de signal**

2.1 Modèle de signal (son quasi-périodique)

2.2 Méthodes temporelles

2.3 Méthodes fréquentielles

2.4 Méthodes combinées

## **3. Utilisation de modèles de perception de hauteur**

3.1 Système auditif humain, deux parties

3.2 Modèle unitaire

3.3 Méthodes temporelles

3.4 Méthodes fréquentielles

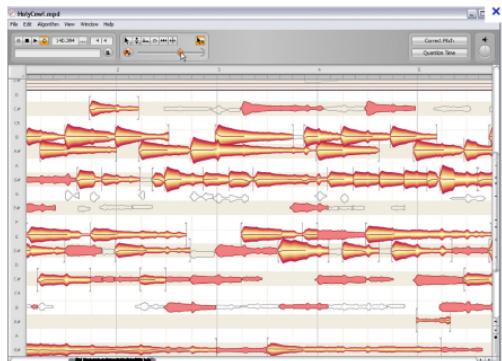
## **4. Utilisation de méthodes de décomposition du signal**

4.1 Séparation de source : méthode NMF

4.2 Probabilistic Latent Component Analysis (PLCA)

## Différents problèmes :

- Estimation de quoi ?
  - des fréquences fondamentales multiples existants à chaque instant :  $f_0(t) = 440 \text{ Hz}$ ,
  - des hauteurs, début et fin de notes : note=A4,
  - de l'instrument ayant joué la note
- Estimation sur quel horizon
  - par trame,
  - globalement sur toute la durée du signal
- Estimation reposant sur
  - les modèles de signaux, la morphologie du spectre, la répétition (en temps/ en fréquence),
  - la modélisation de la perception,
  - la décomposition du signal en sources



## Différentes applications :

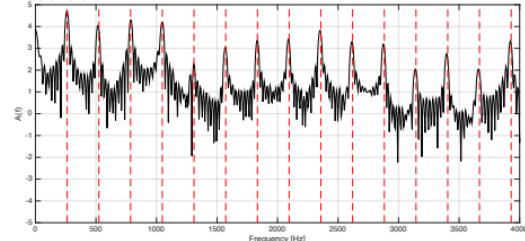
- Codage/ synthèse/ transformation du son : P-SOLA, synthèse sinusoïdale harmonique, melodyne, Audio2Note
- Séparation de sources
- Transcription



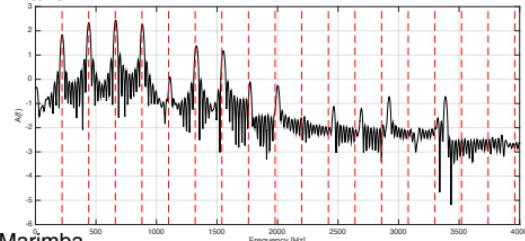
## Plusieurs catégories de sons

- Sons harmoniques
  - $f_h = hf_0$
- Sons légèrement inharmoniques
  - $f_h = hf_0\sqrt{1 + (h^2 - 1)\beta}$
  - Facteur d'inharmonicité  $\beta = 10^{-4}, 10^{-3}$
  - Exemple : piano (inharmonicité due à la raideur des cordes)
- Sons pour lesquels nous percevons un pitch mais qui ne sont pas harmoniques (glockenspiel, vibraphone)
- Sons non-harmoniques

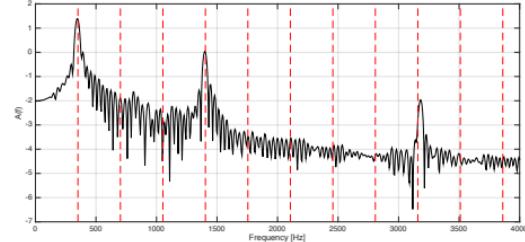
Accordéon



Piano

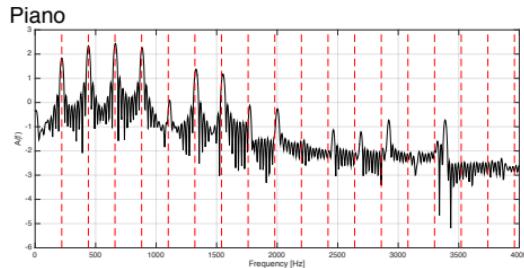


Marimba



## Problématiques source

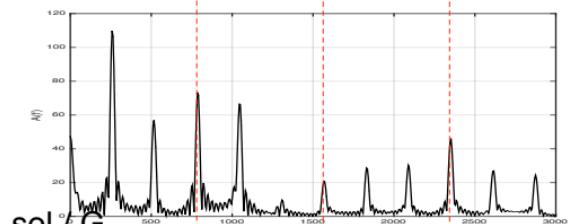
- Les notes peuvent ne pas être parfaitement harmonique



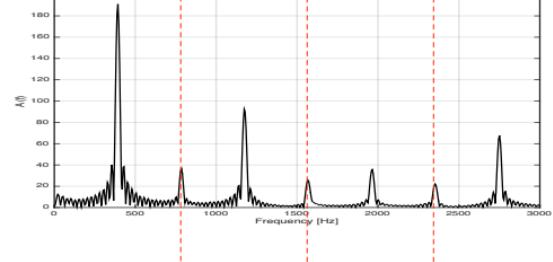
## Problématique f0-multiples

- Les notes peuvent être en rapport harmonique (souvent le cas en musique !)

do / C



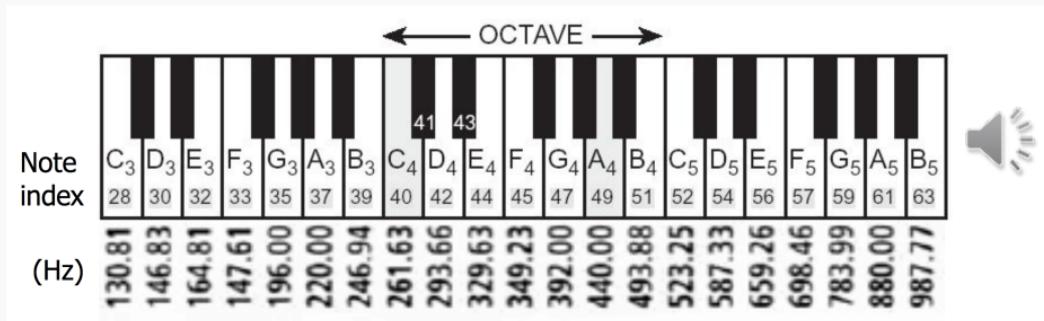
sol / G



# What is pitch ?

## Pitch

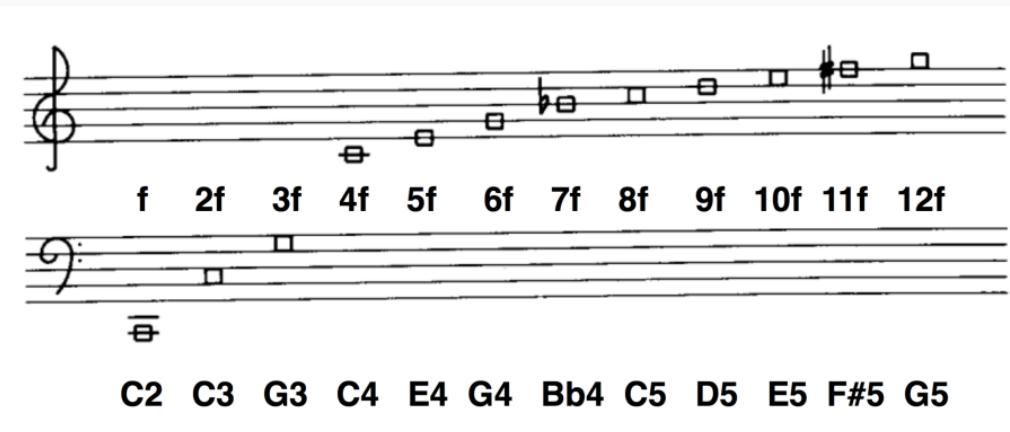
- That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high (ANSI)
- (Operational) A sound has a certain pitch if it can be reliably matched to a sine tone of a given frequency at 40 dB SPL
- People hear pitch in a logarithmic scale



# What is pitch ?

## Harmonics

- Harmonics make tones more pleasant, but may confuse pitch perception, especially in polyphonic settings (octave/harmonic errors)



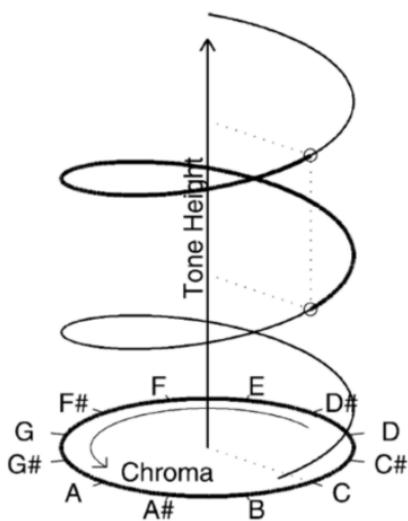
source : Duan, Benetos, 2015

# What is pitch?

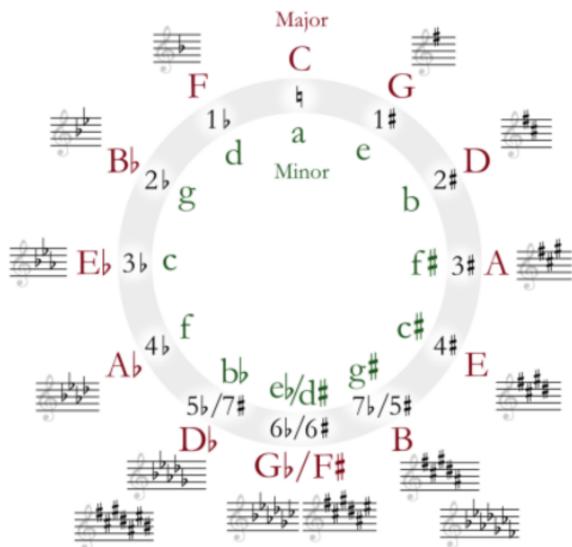
## Multidimensional aspects of pitch

Pitch is not a one-dimensional entity! (low/high)

- Octave similarity – helix representation [Revesz, 1954]
- Pitch distance – circle of fifths representation [Shepard, 1982]



source : Duan, Benetos, 2015

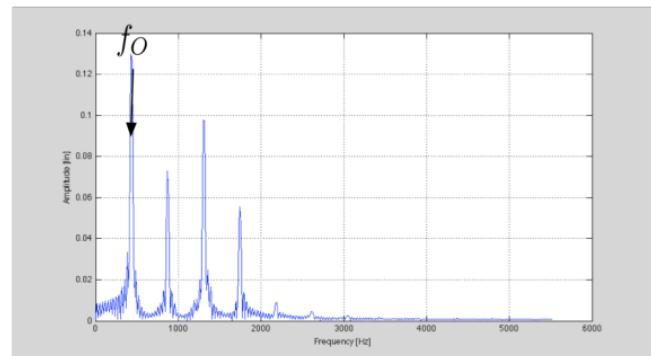
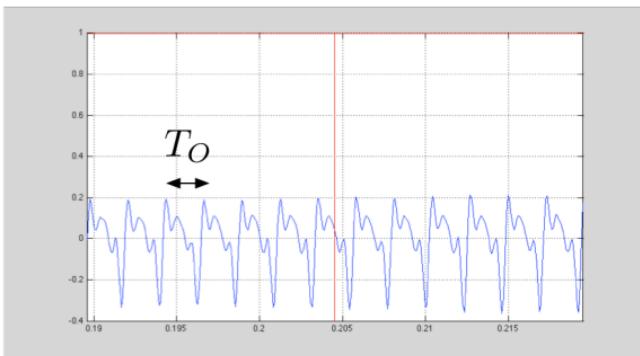


source : Duan, Benetos, 2015



## Période fondamentale $T_0$ ou fréquence fondamentale $f_0$

- $f_0$  : fréquence fondamentale en Hz
  - exemple La3/A4 = 440Hz
- $T_0 = \frac{1}{f_0}$  : période fondamentale en secondes
  - exemple La3/A4 = 0.0023s.



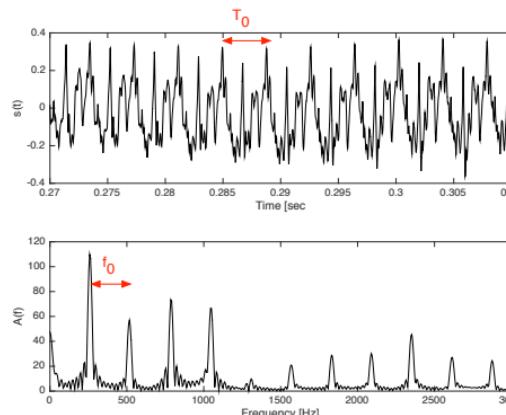
## Modèle de signal (son quasi-périodique)

$$x(n) = \sum_{h=1}^H 2A_h \cos(2\pi h f_0 n + \phi_h) + w(n)$$

- $f_0 = \frac{1}{T_0}$  : fréquence/ période fondamentale
- $H$  est le nombre total d'harmoniques
- $A_h$  sont les amplitudes des harmoniques,  $A_h \geq 0$
- $\phi_h$  sont les phases des harmoniques,  $\phi_h \in [-\pi, \pi]$
- $w(n)$  est un bruit blanc centré de variance  $\sigma^2$

## Auto-covariance

- $x(n)$  est un processeur SSL\* centré d'auto-covariance
  - (\*) SSL : stationnaire au sens large
    - $\mu_x(t) = \mu_x$  et  $P(t, \tau) = P(t - \tau)$
- Auto-covariance :  $r_x(m) = \sum_{h=1}^H [2A_h^2 \cos(2\pi h f_0 m)] + \sigma^2 \delta(m)$



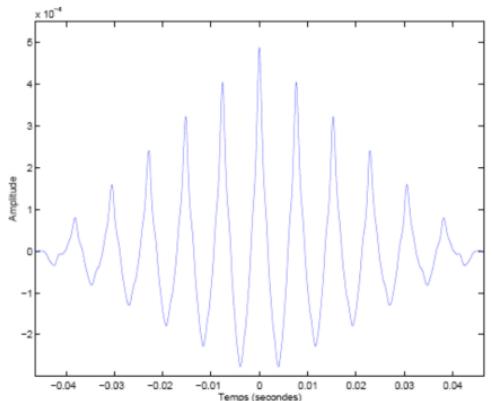
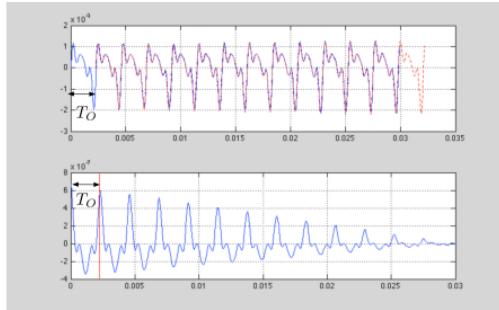
## Auto-corrélation biaisée

$$\hat{r}_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n)x(n+m) \text{ si } m \geq 0$$

- Propriétés :

$$E[\hat{r}_x(m)] = \frac{N - |m|}{N} r_x(m)$$

$$|\hat{r}_x(m)| \leq \hat{r}_x(0)$$



## Auto-corrélation non-biaisée

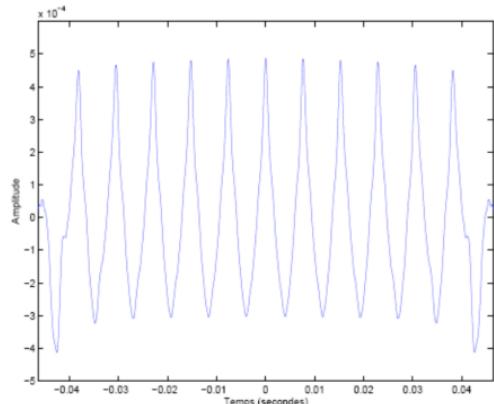
$$\tilde{r}_x(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} x(n)x(n+m) \text{ si } m \geq 0$$

- Propriétés :

$$E[\tilde{r}_x(m)] = r_x(m)$$

$$Var[\tilde{r}_x(m)] = \left(\frac{N}{N-m}\right)^2 Var[\hat{r}_x(m)]$$

$$|\tilde{r}_x(m)| \leq \tilde{r}_x(0)$$



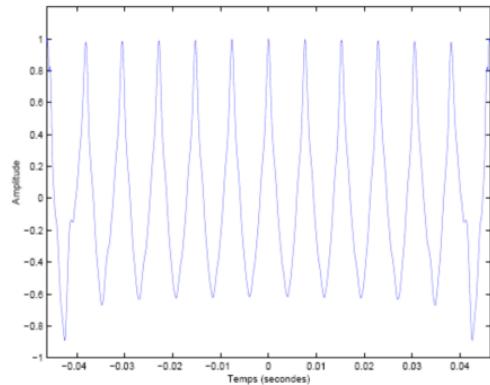
source : Richard, 2012

## Auto-corrélation normalisée

$$\bar{r}_x(m) = \frac{\sum_{n=0}^{N-1-m} x(n)x(n+m)}{\sqrt{\sum_{n=0}^{N-1-m} x(n)^2} \sqrt{\sum_{n=0}^{N-1-m} x(n+m)^2}}$$

- Propriétés :

$|\bar{r}_x(m)| \leq \bar{r}_x(0) = 1$   
 $|\bar{r}_x(m)| = 1$ ssi les vecteurs sont colinéaires



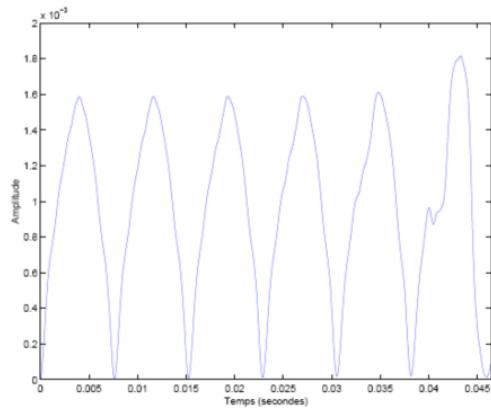
source : Richard, 2012

## Average Square Difference Function (ASDF)

$$ASDF(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} (x(n) - x(n+m))^2$$

- La période  $T_0$  peut être estimée en recherchant le minimum de l'écart quadratique entre les signaux  $x(n)$  et  $x(n+m)$
- Propriétés :

$$ASDF(m) = 0 \text{ ssi } x \text{ est de période } T_0 = m$$
$$E[ASDF(m)] = 2(r_x(0) - r_x(m))$$



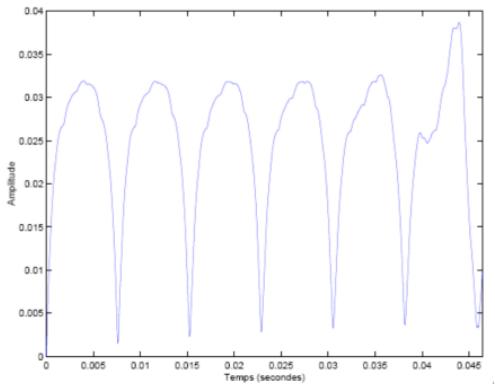
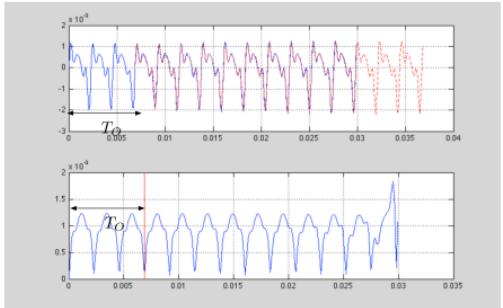
source : Richard, 2012

## Average Magnitude Difference Function (AMDF)

$$AMDF(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x(n) - x(n+m)|$$

- Propriétés :

$$AMDF(m) = 0 \text{ ssi } x \text{ est de période } T_0 = m$$



## Algorithme Yin

[A. de Cheveigné, H. Kawahara, YIN, a fundamental frequency estimator for speech and music, JASA, 2002]

- Point de départ : méthode de l'auto-corrélation
- Améliorations
  - 1) Utilisation de l'ASDF
  - 2) Normalisation
  - 3) Seuillage
  - 4) Interpolation
  - 5) Minimisation locale en temps

Version	Gross error (%)
Step 1	10.0
Step 2	1.95
Step 3	1.69
Step 4	0.78
Step 5	0.77
Step 6	0.50

source : Richard, 2012

## Algorithme Yin

- 1) Utilisation de l'ASDF

- $d_t(\tau) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2$
- lien avec l'auto-corrélation  $d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau)$
- Gain ?
  - l'ASDF est beaucoup moins sensible aux variations des amplitudes relatives que l'ACF (qui est sensible, par exemple, à l'accentuation des partiels d'ordre pair)

- 2) Normalisation

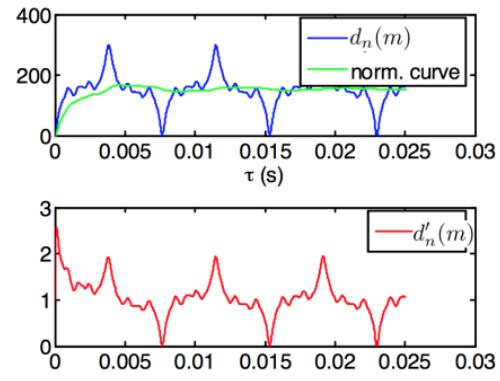
- Normalisation par la "moyenne cumulée"

$$d'_t(\tau) = 1 \quad \text{si } \tau = 0$$

$$= \frac{d_t(\tau)}{\frac{1}{\tau} \sum_{j=1}^{\tau} d_t(j)} \quad \text{sinon}$$

- Gain ?

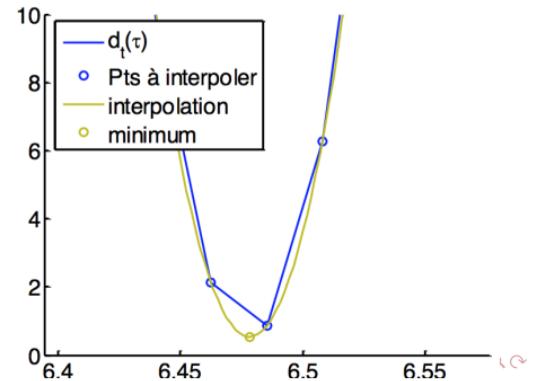
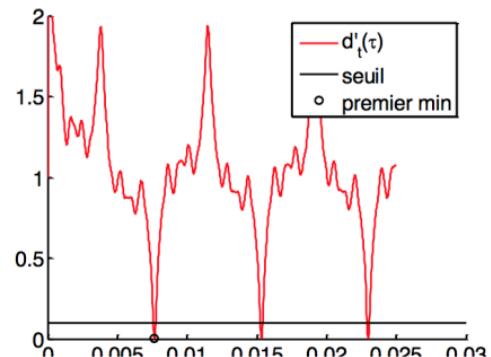
- permet d'éviter les erreurs pour les F0 élevées (suppression du lobe en 0)



source : Richard, 2012

## Algorithme Yin

- 3) Seuillage absolu
  - La plus petite période inférieure au seuil est choisie
  - Si aucune période n'est inférieure au seuil, alors le minimum global est choisi
- 4) Interpolation parabolique autour du minimum
  - Réalisée sur  $d_t(\tau)$  (i.e avant normalisation)
  - Gain : meilleure précision sur la valeur de  $F_0$
- 5) Minimisation locale en temps
  - si on note  $T_t$  la période estimée au temps  $t$
  - pour un temps  $t$ ,
    - on cherche pour  $\theta \in [t - T_{\max}/2, t + T_{\max}/2]$  ( $T_{\max}$  est la période la plus grande considérée, 25 ms)
    - le minima de  $d_\theta(T_\theta)$
  - on réitère avec cette nouvelle estimation et intervalle de recherche de  $\pm 20\%$
  - Gain : effet de lissage en cas de fluctuations de l'estimation
- Autres méthodes possibles pour le lissage : filtre médian, programmation dynamique



## Algorithme Yin

Evaluation sur quatre bases de données de parole

- annotées automatiquement (par YIN, à partir du laryngographe) puis vérifiées et triées à la main

Method	Gross error (%)					
	DB1	DB2	DB3	DB4	Average	(low/high)
pda	10.3	19.0	17.3	27.0	<b>16.8</b>	(14.2/2.6)
fxac	13.3	16.8	17.1	16.3	<b>15.2</b>	(14.2/1.0)
fxcep	4.6	15.8	5.4	6.8	<b>6.0</b>	(5.0/1.0)
ac	2.7	9.2	3.0	10.3	<b>5.1</b>	(4.1/1.0)
cc	3.4	6.8	2.9	7.5	<b>4.5</b>	(3.4/1.1)
shs	7.8	12.8	8.2	10.2	<b>8.7</b>	(8.6/0.18)
acf	0.45	1.9	7.1	11.7	<b>5.0</b>	(0.23/4.8)
nacf	0.43	1.7	6.7	11.4	<b>4.8</b>	(0.16/4.7)
additive	2.4	3.6	3.9	3.4	<b>3.1</b>	(2.5/0.55)
TEMPO	1.0	3.2	8.7	2.6	<b>3.4</b>	(0.53/2.9)
YIN	0.30	1.4	2.0	1.3	<b>1.03</b>	(0.37/0.66)

source : Richard, 2012

## Cepstre réel

- **Auto-correlation** du signal temporel  $\hat{r}(\tau)$  :

$$\hat{r}(\tau) = \int_t x^*(t)x(t + \tau)dt$$

- Sa Transformée de Fourier  $\Gamma(\omega)$  :

$$\Gamma(\omega) = \int_{\tau} \left( \int_t x^*(t)x(t + \tau)dt \right) e^{-j\omega\tau} d\tau$$

$$\Gamma(\omega) = |X(j\omega)|^2$$

- Donc **Auto-correlation** du signal temporel :

$$\hat{r}(I) = \frac{1}{N-I} \sum_k |X(k)|^2 \cos \left( 2\pi k \frac{I}{N} \right)$$

- **Cepstre réel** du signal temporel :

$$\hat{c}(I) = \frac{1}{N-I} \sum_k \log(|X(k)|) \cos \left( 2\pi k \frac{I}{N} \right)$$

- Relation avec le modèle source/filtre :

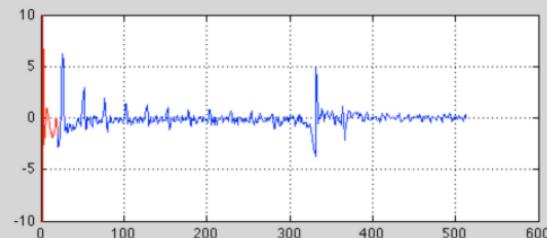
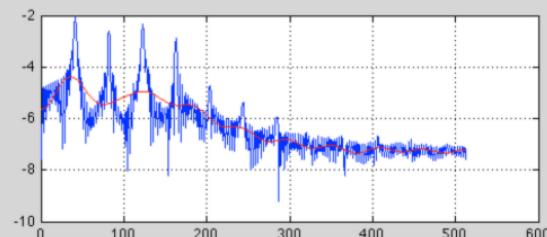
$$x(t) = e(t) \circledast g(t)$$

$$X(\omega) = E(\omega) \cdot G(\omega)$$

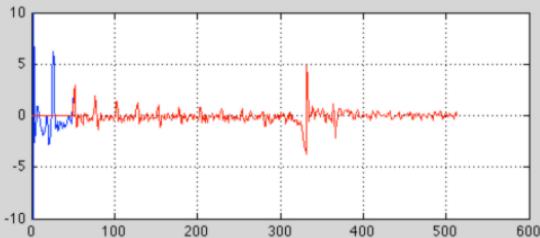
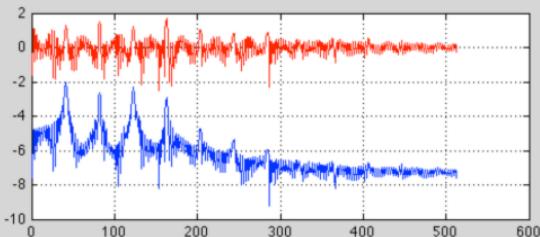
$$\log(X(\omega)) = \log(E(\omega)) + \log(G(\omega))$$

## Cepstre réel

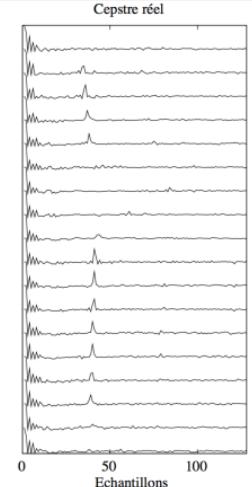
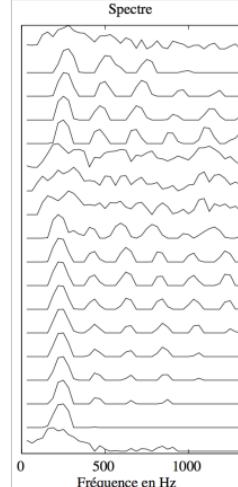
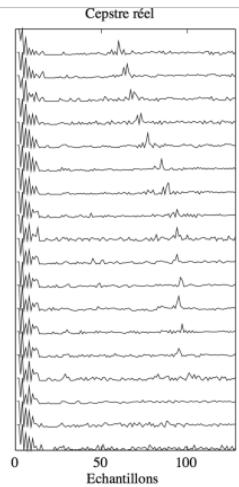
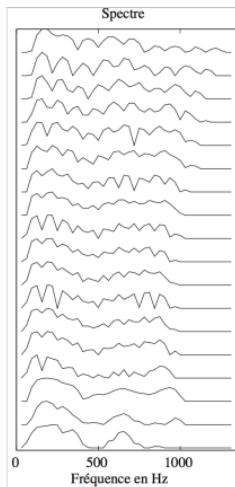
- Le cepstre permet de séparer
  - l'enveloppe spectrale
    - ce qui varie lentement
    - basse fréquence de la  $TF^{-1}$



- la fréquence fondamentale
  - ce qui varie rapidement
  - haute fréquence de la  $TF^{-1}$



## Cepstre réel



source : voix d'homme, Laroche, 1995

source : voix de femme, Laroche, 1995

## Approche par le maximum de vraisemblance

- Modèle de signal :  $x(n) = a(n) + w(n)$ 
  - $a$  est un signal périodique de période  $T_0$
  - $w$  est un bruit blanc gaussien de variance  $\sigma^2$
- vraisemblance des observations

$$p(x|T_0, a, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x(n) - a(n))^2} \quad (1)$$

- log-vraisemblance

$$L(T_0, a, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x(n) - a(n))^2 \quad (2)$$

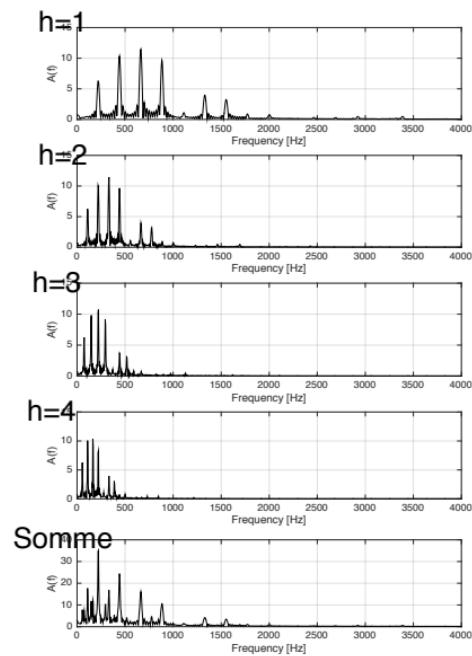
- Méthode :
  - maximiser successivement  $L$  par rapport à  $a$ , puis  $\sigma^2$  et enfin  $T_0$

## Somme spectrale

- On peut montrer que la maximisation de  $L$  par rapport à  $F_0 = \frac{m}{N}$  revient à maximiser la somme spectrale

$$S(e^{j2\pi \frac{m}{N}}) = \sum_{h=1}^H \hat{R}_x(e^{j2\pi \frac{m}{N} \cdot h})$$

$$S(\omega) = \sum_{h=1}^H |X(e^{j\omega \cdot h})|^2 \text{ pour } \omega < \frac{\pi}{H}$$

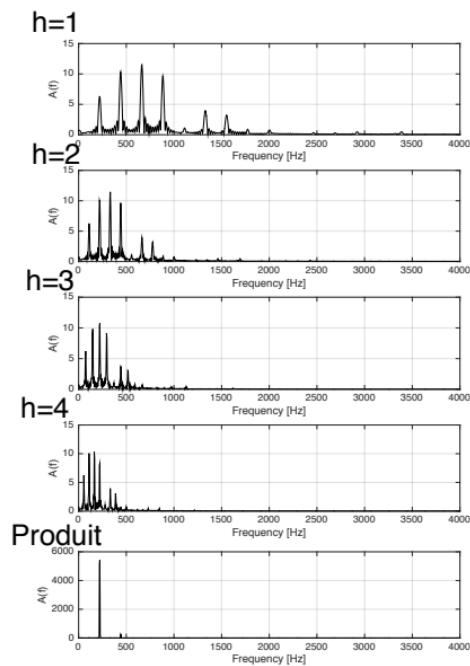


## Produit spectral

- Par similitude avec la somme spectrale on peut définir le produit spectral (souvent plus robuste)

$$P(e^{j2\pi \frac{m}{N}}) = \prod_{h=1}^H \hat{R}_x(e^{j2\pi \frac{m}{N} \cdot h})$$

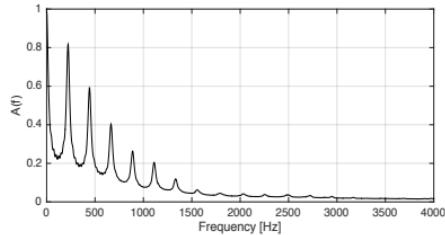
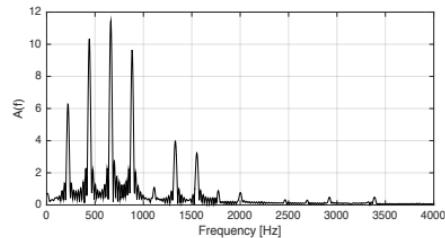
$$P(\omega) = \prod_{h=1}^H |X(e^{j\omega \cdot h})|^2 \text{ pour } \omega < \frac{\pi}{H}$$



## Auto-corrélation du spectre d'amplitude

- Mesure de la périodicité de l'espacement entre les harmoniques
  - ne fait pas l'hypothèse qu'il existe de l'énergie à la fréquence  $f_0$

$$\hat{R}(k) = \frac{1}{N-k} \sum_{\kappa=0}^{N-k-1} |X(\kappa)| |X(\kappa+k)|$$

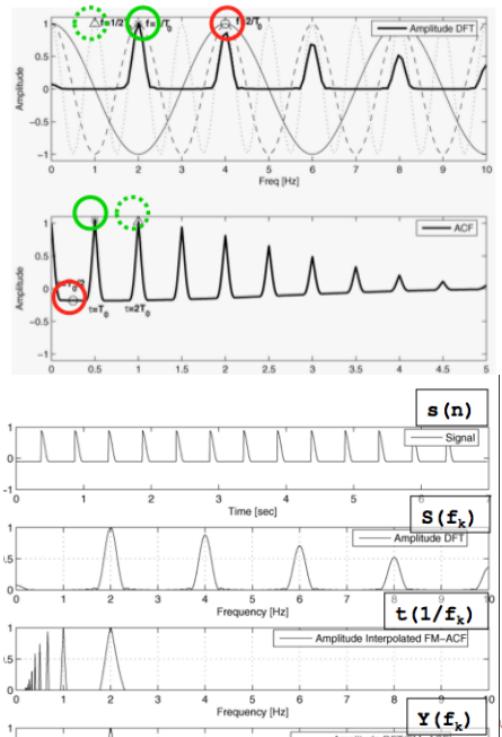


# Méthodes combinées

## Combinaison de la DFT et de l'ACF

[G. Peeters, Music pitch representation by periodicity measures based on combined representations, IEEE ICASSP, 2006]

- Méthodes temporelles  $T(\tau_I)$ 
  - Auto-correlation du signal temporel
  - Cesptre réel du signal temporel
- Méthodes fréquentielles  $S(f_k)$ 
  - Spectre d'amplitude (réassigné fréquentiellement)
  - Auto-correlation du spectra d'amplitude (réassigné fréquentiellement)
- Principe
  - Les erreurs ponctuelles d'octave sont dans des directions opposées
  - Combiner les deux représentations
- Méthode
  - Calculé les valeurs de la représentation temporelle aux fréquences  $f_k$ 
    - interpolation) de  $T(\tau_I)$  à  $f_k : T(1/f_k)$
  - Calculé le produit :
    - $P(f_k) = S(f_k)T(1/f_k)$

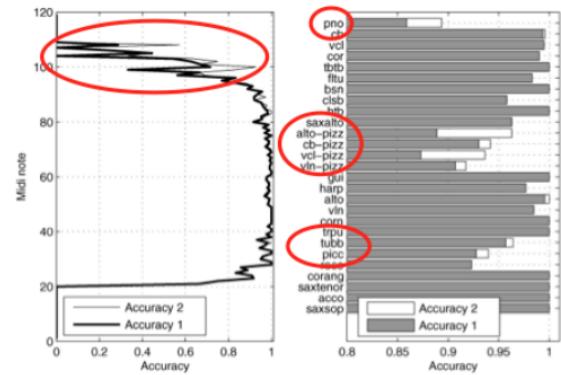


## Combinaison de la DFT et de l'ACF

- Résultats

	accuracy 1	accuracy 2
DFT / ACF	81,6	91,7
DFT / CEP	91,4	95,8
ACFofDFT / ACF	95	96,1
ACFofDFT / CEP	97	97,6
ACFofREAS / CEP	97	97,3
Yin	94,9	95,5

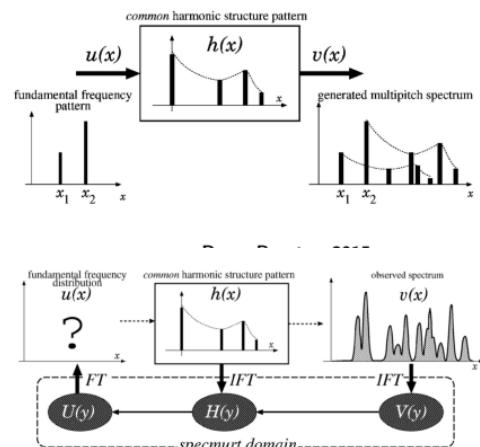
- Résultats



## SpecMurt

[S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama. Specmurt analysis of polyphonic music signals. IEEE TASLP, 2008]

- La transposition d'un son devient une translation sur l'axe de log-fréquence
- Analyse Specmurt :
  - On suppose le spectre formé de la convolution des notes  $u(x)$  et d'une structure harmonique  $h(x)$ 
    - suppose une structure harmonique commune à toutes les notes à la même trame
    - mais ne suppose pas le partage à des trames différentes (en contraste avec d'autres méthodes comme la NMF)
  - Power-spectrum en log-fréquence :
    - $v(x) = u(x) \circledast h(x)$
  - IFFT
    - $V(y) = U(y) \cdot H(y)$
  - Méthode ?
    - Estimation itérative de  $u(x)$  et  $h(x)$



source : Duan, Benetos, 2015



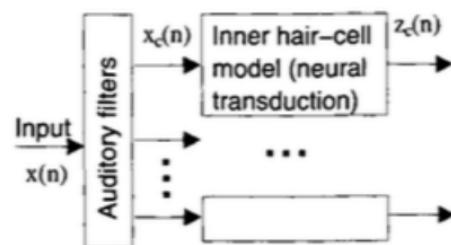
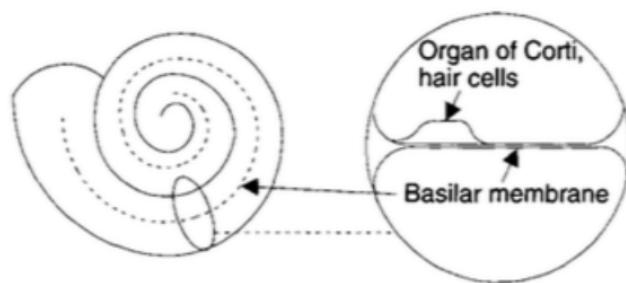
### 3- Utilisation de modèles de perception de hauteur

#### 3.1- Système auditif humain, deux parties

- Système périphérique

- oreille externe
  - écoute directionnel
- oreille moyenne
  - écoute directionnel
- oreille interne
  - la cochlée : transforme les variations de pressions en impulsions neuronales dans le nerf auditif
  - membrane basilaire vibre : propagation des ondes : fréquences aiguës (début), fréquences graves (fin)
  - organe de Corti : cellules ciliées (interne ou externe) réceptives à différentes fréquences

- Cortex auditif dans le cerveau



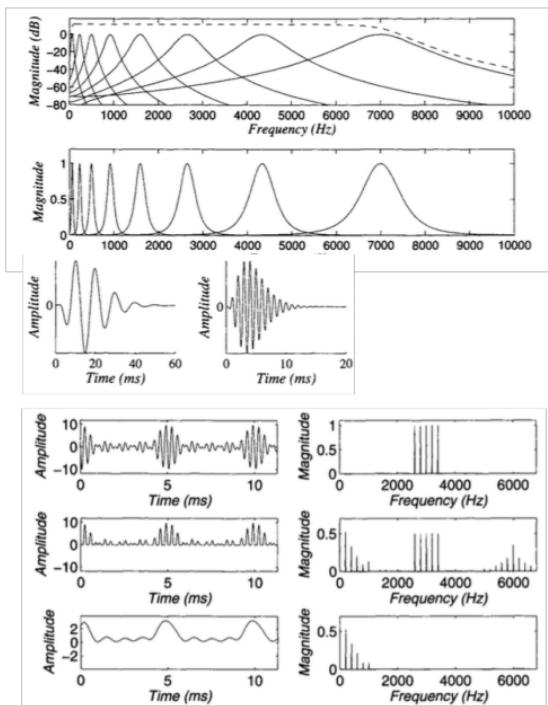
**Fig. 8.3.** An illustration of the cochlea (*left*) and its cross-section (*middle*). The right panel shows a rough computational model of the cochlea.

source : Klapuri, 2006

### 3- Utilisation de modèles de perception de hauteur

#### 3.2- Modèle unitaire

- 1. **Cochlée** : filtrage passe-bande
  - filtre gamma-tone
  - expérience du chat
- 2. **Cellules ciliées interne** :
  - compression/ adaptation-de-niveau,
  - half-wave rectification
  - filtrage passe-bas
  - expérience du chat
- 3. Mesure de la périodicité dans chaque canal
  - ACF ou filtre-résonateur
- 4. Aggrégation des périodicités à travers les bandes
  - sommation ou somme pondérée

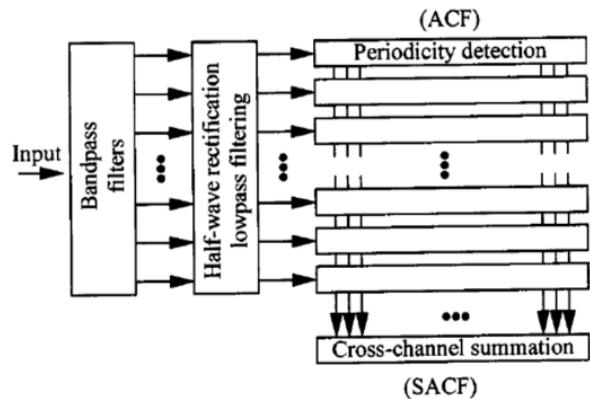


### 3- Utilisation de modèles de perception de hauteur

#### 3.3- Méthodes temporelles

##### Approche par banc de filtres

[R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I : Pitch identification. JASA, 1991.]



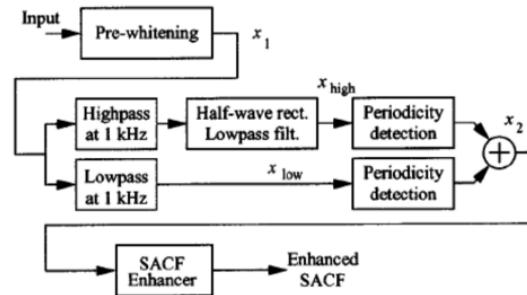
source : Richard, 2012

### 3- Utilisation de modèles de perception de hauteur

#### 3.3- Méthodes temporelles

##### Approche par banc de filtres plus simple

[T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. Speech and Audio Processing, IEEE, 2000.]



source : Richard, 2012

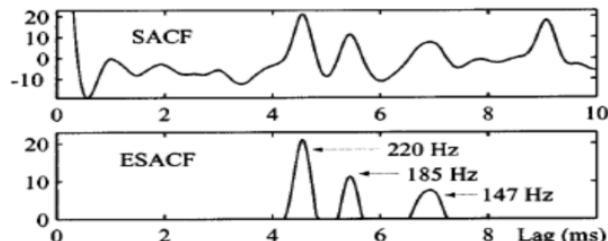
### 3- Utilisation de modèles de perception de hauteur

#### 3.3- Méthodes temporelles

## Enhanced Summary ACF

[T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. Speech and Audio Processing, IEEE, 2000.]

- Plusieurs étapes :
  - Redressement demi-onde
    - On ne conserve que les valeurs positives
  - Ralentie 2 (ou plus) fois puis déduite du SACF redressé
    - Permet de supprimer les pics doubles



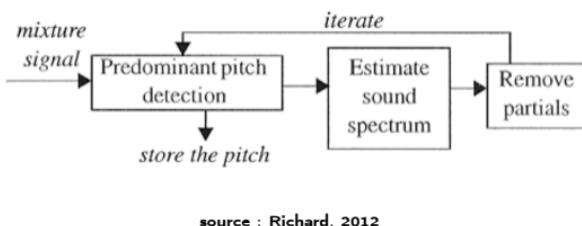
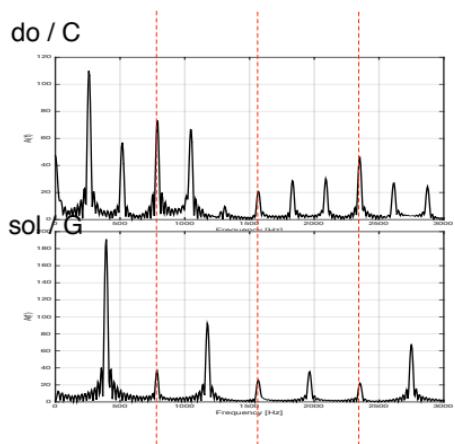
source : Richard, 2012

### 3- Utilisation de modèles de perception de hauteur

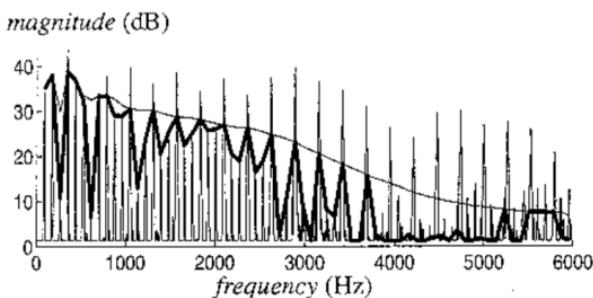
#### 3.4- Méthodes fréquentielles

Estimation fréquences multiples : approche par soustraction itérative

[A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. IEEE TSAP, 2003.]



source : Richard, 2012



source : Richard, 2012

#### Principe de lissage spectral

- $a_h = \min(a_h, m_h)$
- où  $m_h$  est la moyenne sur une fenêtre d'une octave autour du partielle

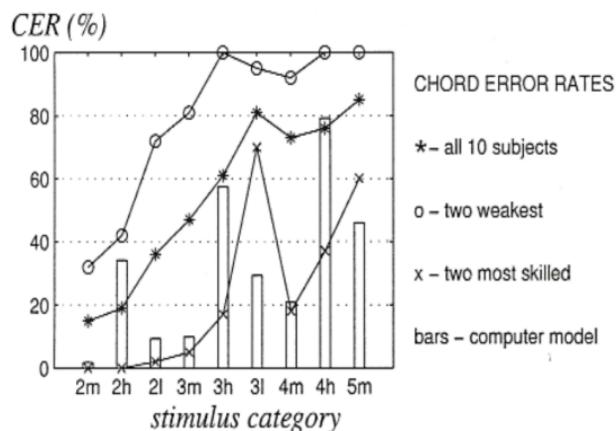
### 3- Utilisation de modèles de perception de hauteur

#### 3.4- Méthodes fréquentielles

Estimation fréquences multiples : approche par soustraction itérative

#### Comparaison des performances algorithmiques aux performances humaines

- (l) Registre bas : 33 à 130 Hz
- (m) Registre médium : 130 à 520 Hz
- (h) Registre haut : 520 à 2100 Hz
- 200 stimuli sonores (20 catégories)
- Sons polyphoniques générés par ordinateur à partir d'échantillons de Piano Steinway provenant du Master samples collection, Mc Gill University
- Personnes ayant participé aux tests :
  - tous sont musiciens
  - dont 2 ont l'oreille absolue (musiciens quasi-professionnels)



source : Richard, 2012

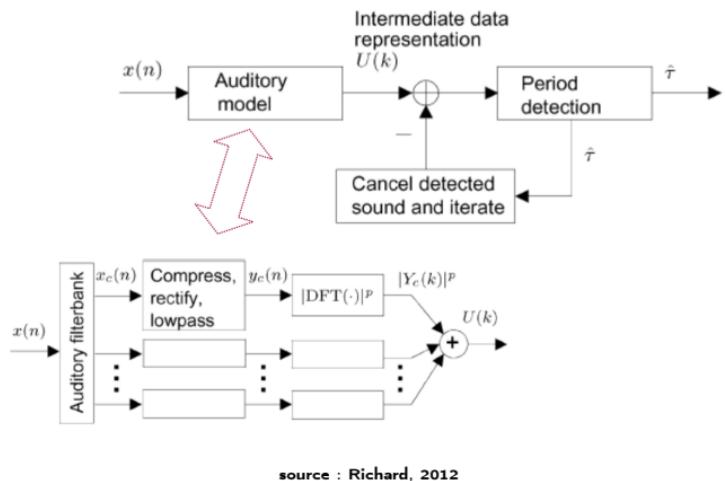
### 3- Utilisation de modèles de perception de hauteur

#### 3.4- Méthodes fréquentielles

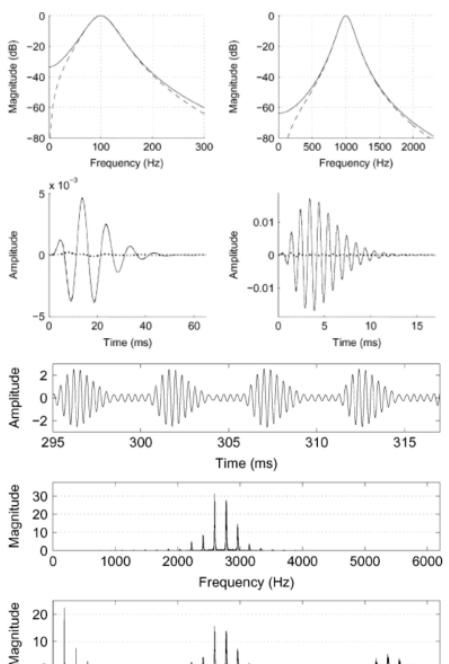
Estimation fréquences multiples : approche par soustraction itérative

#### Amélioration (modèle perceptif)

[A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE TASLP, 2008.]



source : Richard, 2012



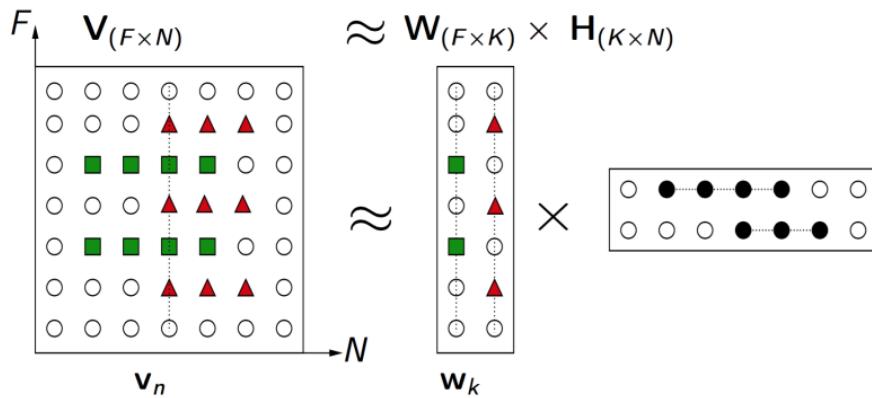




# Séparation de source : méthode NMF

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

- **NMF** : Non-Negative Matrix Factorization
  - Factorisation (décomposition) en matrices non-négatives (NMF)

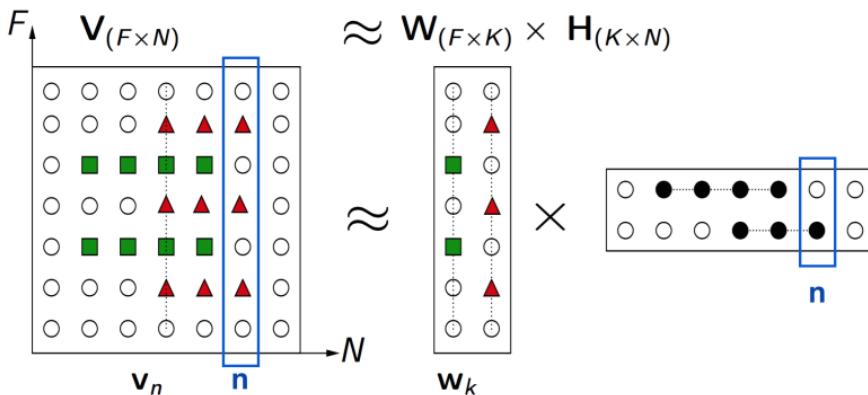


source : Cédric Févotte

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

## Séparation de source : méthode NMF

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

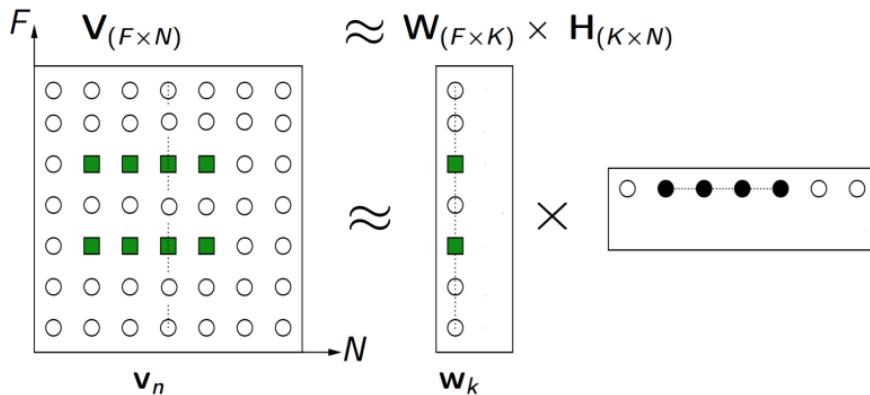


source : Cédric Févotte

## Séparation de source : méthode NMF

- Le signal d'une **source**  $k$  est reconstitué comme

$$\bullet \quad V_{(1:F,1:N)}^k = W_{(1:F,k=1)} H_{(k=1,1:N)}$$

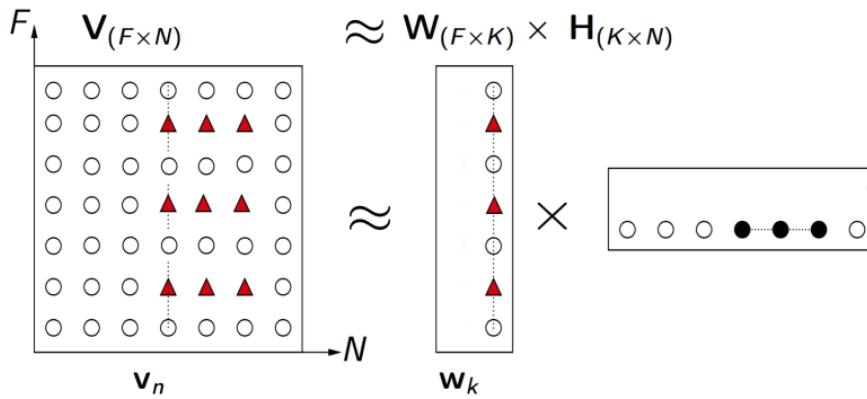


source : Cédric Févotte

## Séparation de source : méthode NMF

- Le signal d'une **source**  $k$  est reconstitué comme

- $V_{(1:F,1:N)}^k = W_{(1:F,k=2)} H_{(k=2,1:N)}$



source : Cédric Févotte

## Estimation des paramètres de la NMF

- $\underline{V}_{(F,N)} \simeq \underline{W}_{(F,K)} \underline{H}_{(K,N)}$
- **Minimisation** de
  - $\min_{\underline{W}, \underline{H} \geq 0} D(\underline{V} | \underline{W}\underline{H})$
  - $\min_{\theta} C(\theta) \stackrel{\text{def}}{=} D(\underline{V} | \underline{W}\underline{H})$  avec  $\theta = \{\underline{W}, \underline{H}\}$
- $D/d$  est une **divergence séparable**
  - $D(\underline{V} | \hat{\underline{V}}) = \sum_{f=1}^F \sum_{n=1}^N d(v_{fn} | \hat{v}_{fn})$
- Choix de  $D/d$  :
  - Distance Euclidienne :
$$d_{EUC}(x, y) = (x - y)^2$$
  - Divergence de Kullback-Leibler :
$$d_{KL}(x, y) = x \log \frac{x}{y} - x + y$$
  - Divergence d'Itakura-Saito :
$$d_{IS}(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

## Estimation des paramètres de la NMF

- **Dérivation du critère pour la distance Euclidienne**

- Non Negative Matrix Factorization

$$V_{(f,n)} \simeq W_{(f,k)} H_{(k,n)}$$

- 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 :  $\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$

## Estimation des paramètres de la NMF

- Déivation du critère pour la distance Euclidienne

$$SSE = \|V - WH\|_F^2$$

$$\begin{aligned} SSE &= (V - WH)^T(V - WH) \\ &= (V^T - H^T W^T)(V - WH) \\ &= V^T V - V^T WH - H^T W^T V + H^T W^T WH \\ &= V^T V - 2V^T WH + H^T W^T WH \end{aligned}$$

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

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

- Propriétés utilisées (Matrix CookBook)

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

## Estimation des paramètres de la NMF

- **Algorithme de descente de gradient**

- déplacement dans la direction opposée au gradient, de manière à faire décroître la fonction

- Le gradient :  $\frac{\partial \text{sse}}{\partial H} = 2\underbrace{W^T WH}_{\nabla_+} - 2\underbrace{W^T V}_{\nabla_-}$

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

$$H \leftarrow H + \frac{H}{W^T WH} (W^T V - W^T WH)$$

$$H \leftarrow H + \frac{HW^T V}{W^T WH} - H$$

$$\nabla_-$$

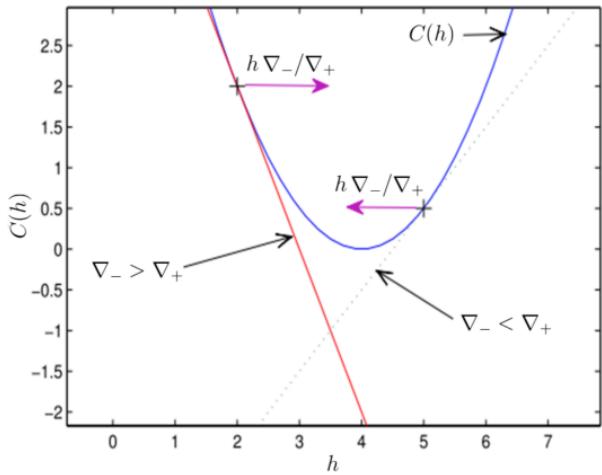
$$H \leftarrow H \cdot \frac{W^T V}{\underbrace{W^T WH}_{\nabla_+}}$$

## Estimation des paramètres de la NMF

- **Mise à jour multiplicative**

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

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



## Estimation des paramètres de la NMF

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

- 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  $H$

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

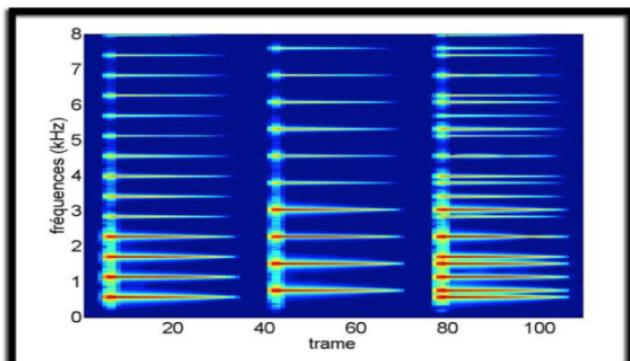
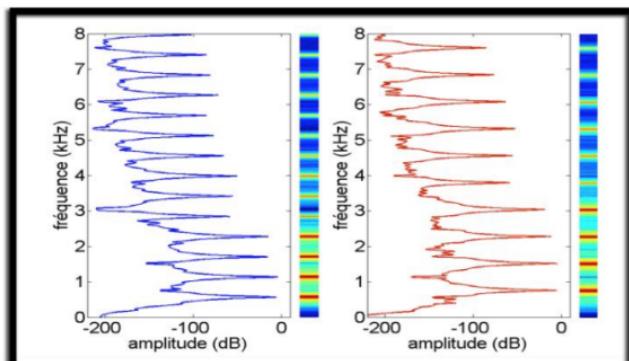
- Mise à jour des activations  $H$  étant donné les bases  $W$

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

- 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 décroître

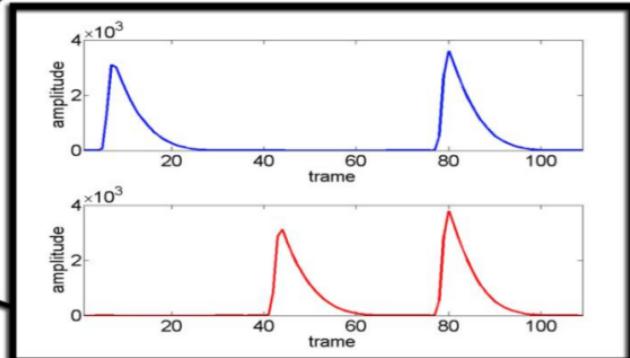
# Séparation de source : méthode NMF

## Exemple

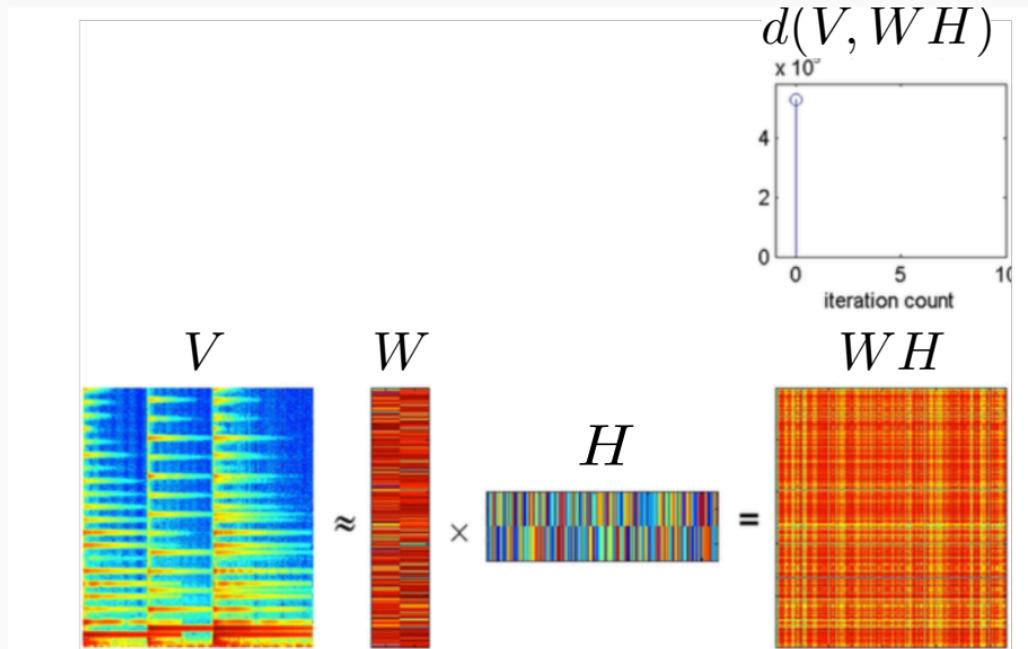


$$W H \approx V$$

*Image d'après R. Hennequin*



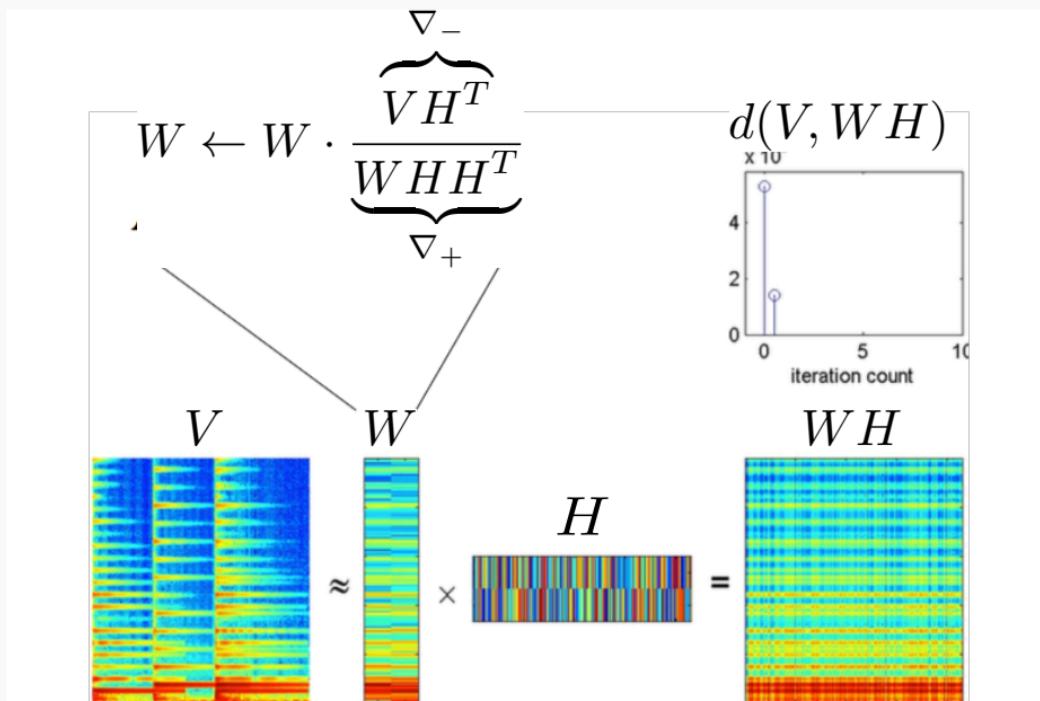
## Initialisation



source : Tuomas Virtanen

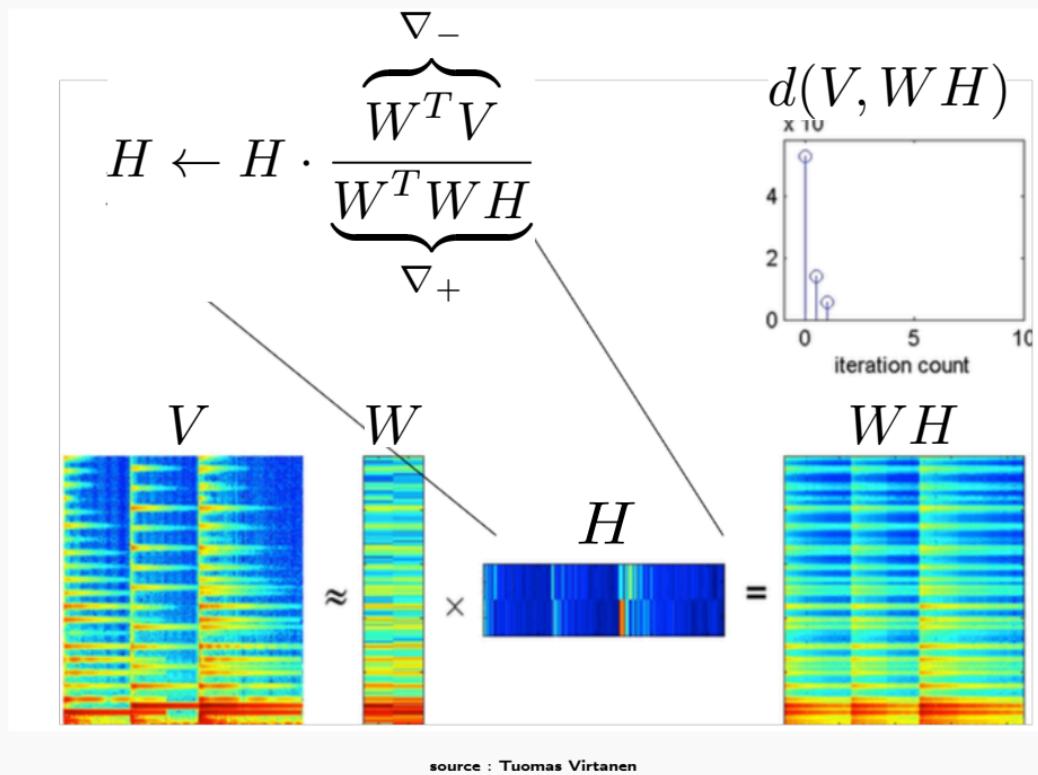
# Séparation de source : méthode NMF

## Iteration 1 : Mise à jour de W



source : Tuomas Virtanen

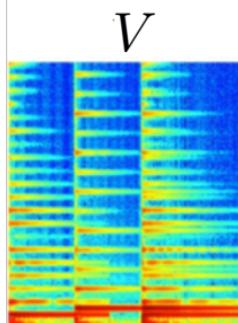
## Iteration 1 : Mise à jour de H



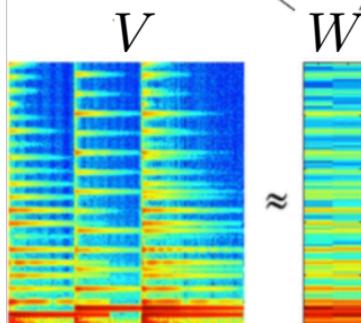
## Séparation de source : méthode NMF

Iteration 2 : Mise à jour de W

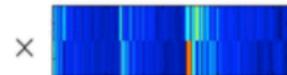
$$W \leftarrow W \cdot \frac{\nabla_-}{\nabla_+} \underbrace{VH^T}_{W H H^T}$$



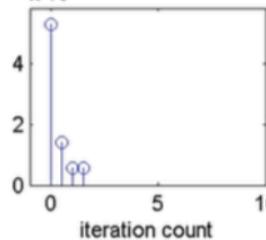
$\approx$



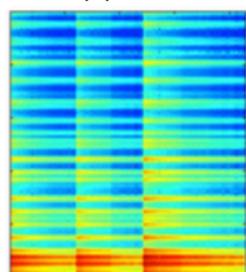
$H$



$$d(V, WH)$$

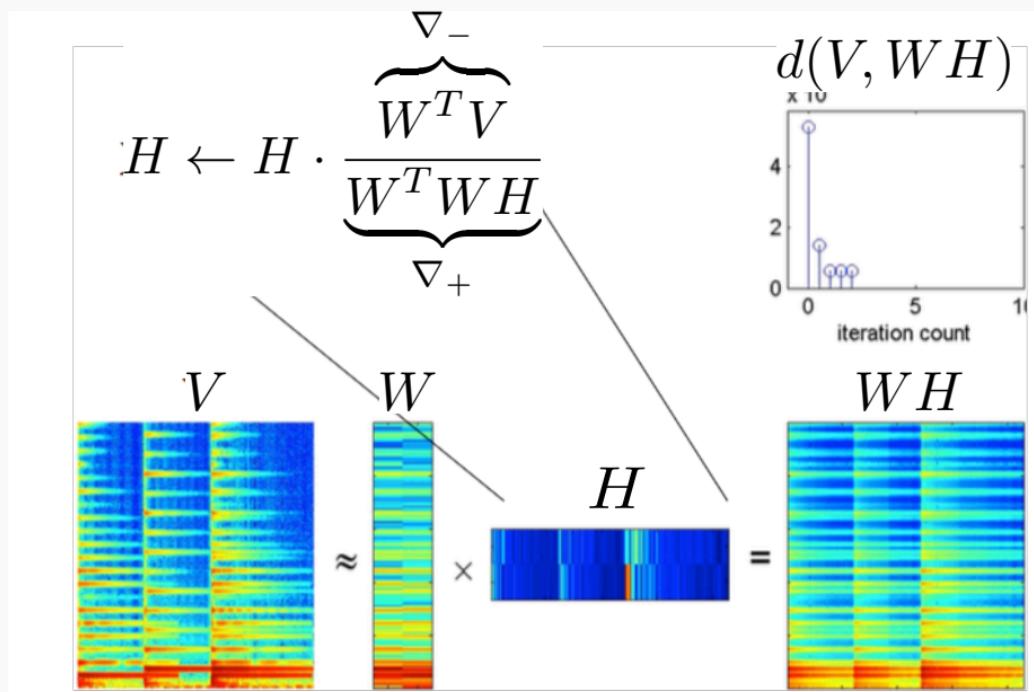


$WH$



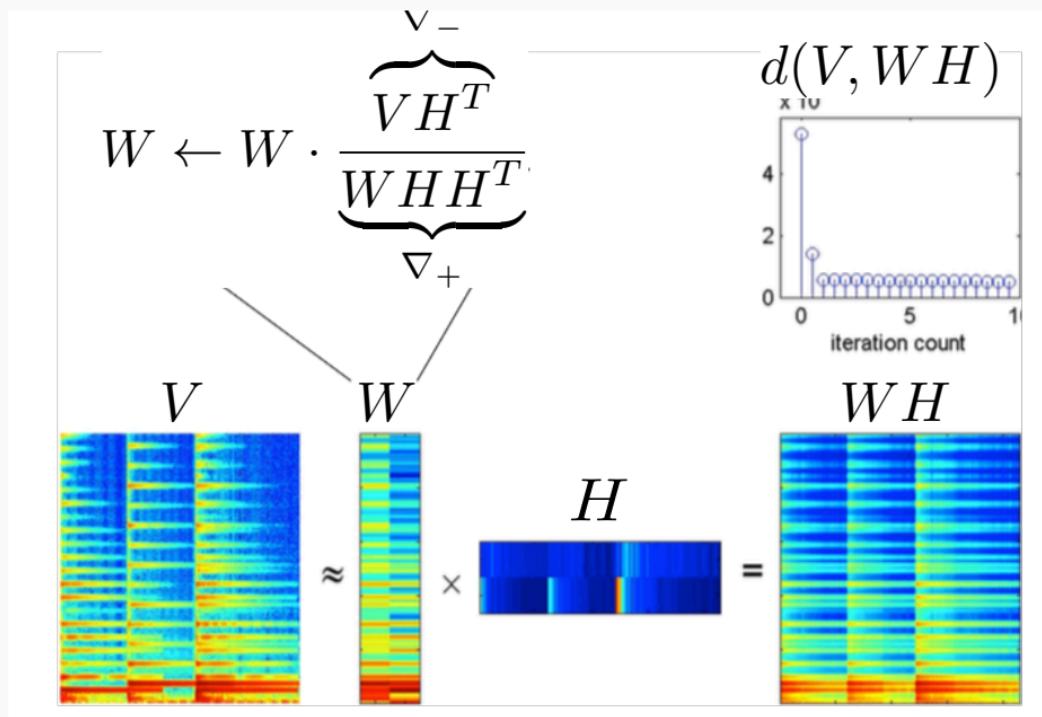
source : Tuomas Virtanen

Iteration 2 : Mise à jour de H



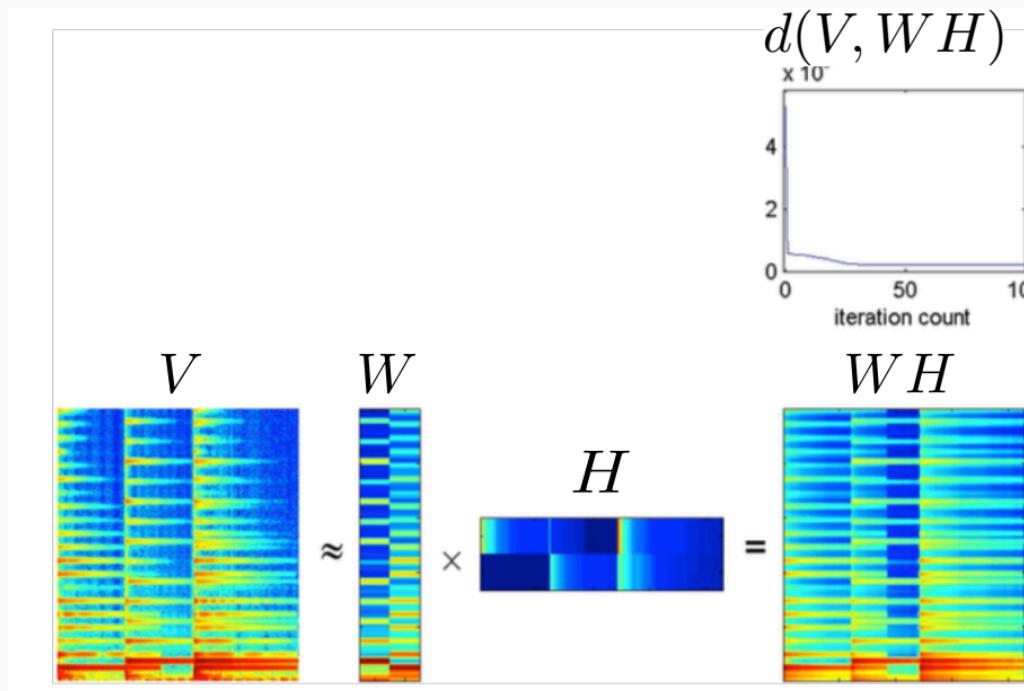
source : Tuomas Virtanen

Iteration 10 : Mise à jour de W



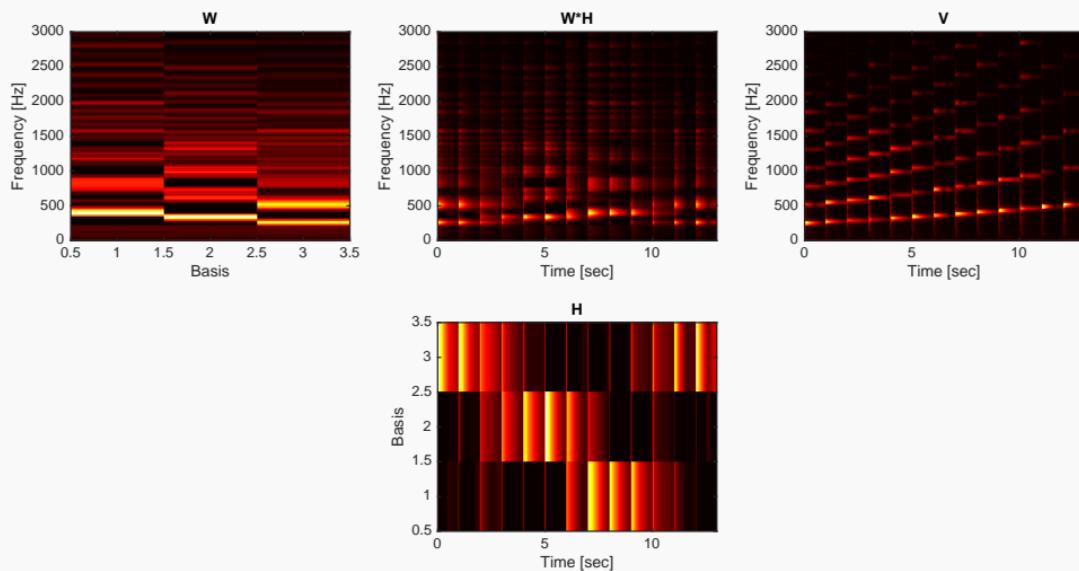
## Séparation de source : méthode NMF

Iteration 100



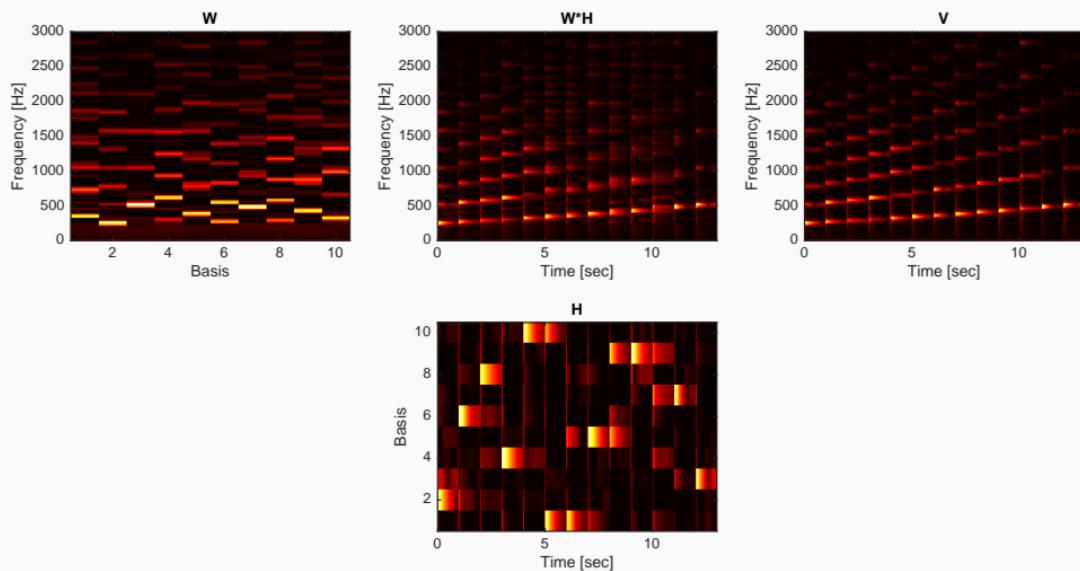
source : Tuomas Virtanen

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

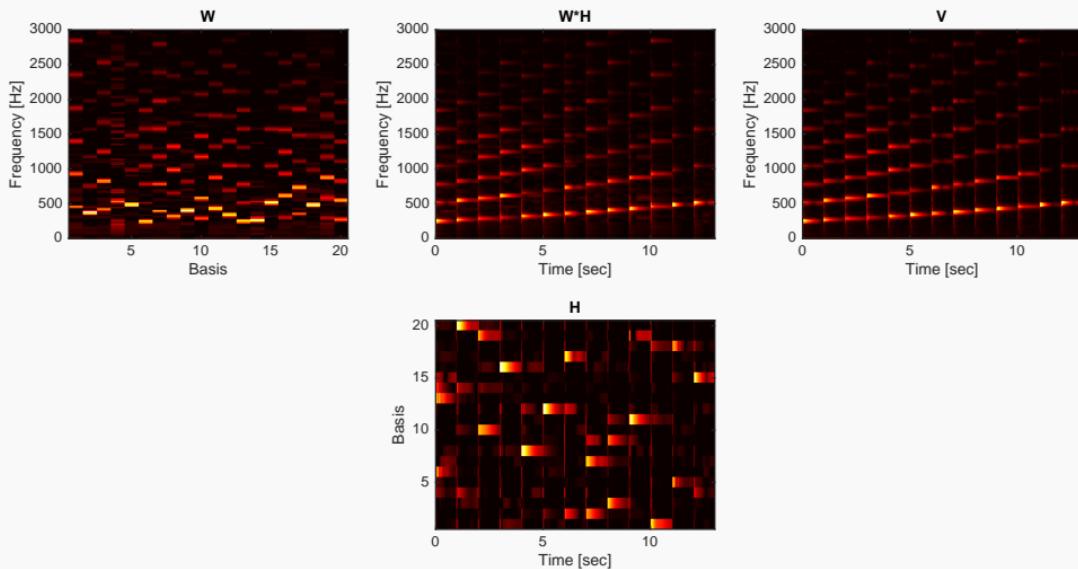


# Séparation de source : méthode NMF

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



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



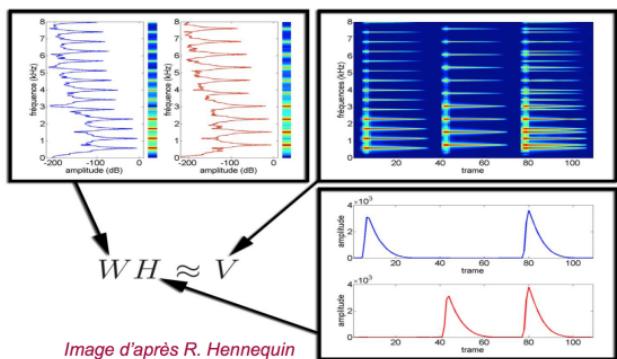
# Utilisation de la NMF pour la détection multi-pitch

- Notations :  $V \simeq WH \rightarrow Y_{ft} \simeq \sum_{i=1}^I A_{it} S_{if}$

## Approche 1

[P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. IEEE WASPAA, 2003]

- Apprendre **simultanément** les activations  $A_{it}$  et les bases  $S_{if}$ 
  - Analyser a posteriori le pitch correspondant à chaque base apprise
  - $Y_{ft}$  est un spectrogramme (amplitude, énergie)



source : Romain Hennequin

# Utilisation de la NMF pour la détection multi-pitch

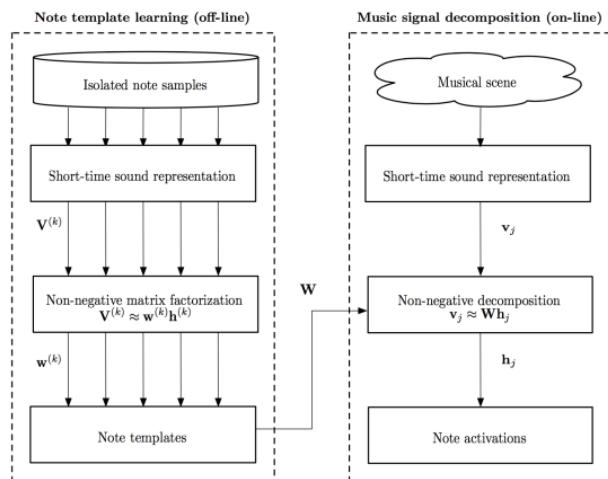
- Notations :  $V \simeq WH \rightarrow Y_{ft} \simeq \sum_{i=1}^I A_{it} S_{if}$

## Approche 2

[A. Dessein, A. Cont, and G. Lemaître. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. ISMIR, 2010.]

- Pré-entraîner les bases  $S_{if}$  sur un ensemble d'entraînement

- permet de contraindre cet apprentissage à l'apprentissage de bases correspondant à des hauteurs connues (fixer les activations)
- Les bases du dictionnaire  $S_{if}$  sont pré-apprises (offline) et représentent les différentes notes d'un piano
  - chaque note est apprise par une NMF de rang  $I = 1$  sur un ensemble de spectre représentant la note
- Utilisation de la NMF avec  $\beta$ -divergence



source : Dessein, 2010

# Utilisation de la NMF pour la détection multi-pitch

- Notations :  $V \simeq WH \rightarrow Y_{ft} \simeq \sum_{i=1}^I A_{it} S_{if}$

## Approche 3

[E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *Audio, Speech and Language Processing, IEEE Transactions on*, 18(3) :528–537, 2010.]

- Construire** le modèle pour que les bases apprises correspondent à des notes
- Modèle générale :  $Y_{ft} = \sum_{i=1}^I A_{it} S_{if}$
- Modèle constraint :
  - $i \rightarrow (p, j)$
  - $Y_{ft} = \sum_{p=p_{low}}^{p_{high}} \sum_{j=1}^{J_p} A_{(p,j),t} S_{(p,j),f}$
  - $S_{(p,j),f}$  : templates ayant le même pitch  $p$  mais différentes enveloppes spectrales  $j$
  - $S_{(p,j),f} = \sum_{k=1}^{K_p} E_{pj,k} N_{pk,f}$ 
    - $N_{pk,f}$  représente la structure fine du spectre associé au pitch  $p$
    - $E_{pj,k}$  représente l'enveloppe spectrale

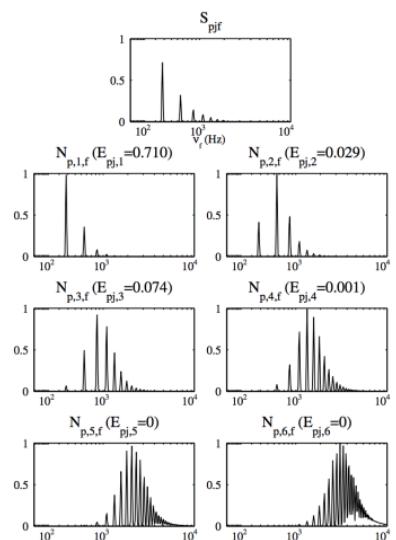


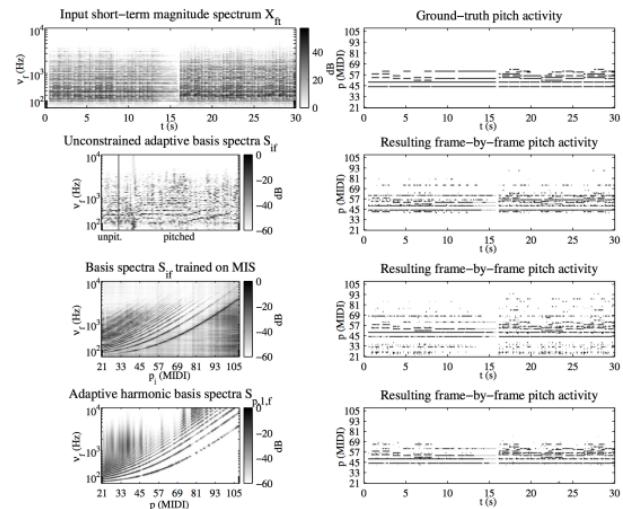
Fig. 2. Basis spectrum  $S_{pjf}$  estimated for the piano excerpt in Fig. 1 given fixed harmonic fine structure spectra  $N_{pkf}$  ( $p = 60$ , gammatone windows of order  $n = 4$ ,  $b = 11/3$  ERB,  $K_{\max} = 6$ ).

# Utilisation de la NMF pour la détection multi-pitch

- Notations :  $V \simeq WH \rightarrow Y_{ft} \simeq \sum_{i=1}^I A_{it} S_{if}$

## Comparaison des trois approches

- 1. Apprendre simultanément les activations  $A_{it}$  et les bases  $S_{if}$
- 2. Pré-entraîner les bases  $S_{if}$  sur une ensemble d'entraînement
- 3. Contraindre le modèle pour que les bases apprises correspondent à des notes



source : Vincent, 2010

## 4- Utilisation de méthodes de décomposition du signal

### 4.2- Probabilistic Latent Component Analysis (PLCA)

# Probabilistic Latent Component Analysis (PLCA)

[M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. Computational intelligence and neuroscience, 2008.]

- Cadre déterministe : **Non-Negative Matrix Factorization (NMF)** :

$$\begin{aligned}\underline{\underline{V}}_{f,t} &= \underline{\underline{W}}_{f,z} \underline{\underline{H}}_{z,t} \\ \underline{\underline{v}}_{:,t} &= \sum_z \underline{\underline{w}}_{:,z} \underline{h}_{zt}\end{aligned}$$

- Cadre probabiliste : **Probabilistic Latent Component Analysis (PLCA)** :

- $P(x, y) = \sum_z P(x, y|z)P(z)$
- si  $x, y$  indépendants étant donné  $z$  :  $P(x, y) = \sum_z P(z)P(x|z)P(y|z) = \sum_z P(x|z)P(z, y)$
- si  $x = f$  et  $y = t$

- $P(f, t) = \sum_z P(f|z)P(z, t)$ 
  - $z$  : variable cachée (latent variable)
  - $P(f|z)$  : probabilité conditionnelle sur la variable cachée

- Estimation : utilisation d'un algorithme Expectation/ Maximization

# Shift-Invariant Probabilistic Latent Component Analysis (SI-PLCA)

[P. Smaragdis and B. Raj. Shift-invariant probabilistic latent component analysis. Journal of Machine Learning Research, 2007.]

- PLCA

- $P(f, t) = \sum_z P(z)P(f|z)P(t|z) = \sum_z P(f|z)P(z, t)$

- **Shift-Invariant** (sur une dimension) PLCA

- $P(f, t) = \sum_z P(z)P(f, t|z)$

- On décompose  $f = f' + \tau$

- $f'$  : fréquence de base

- $\tau$  : variable de transposition

- $P(f, t) = \sum_z P(z) \sum_{f'} P_K(f'|z) P_I(f - f', t|z)$

- $P_K$  est la distribution noyau (Kernel)

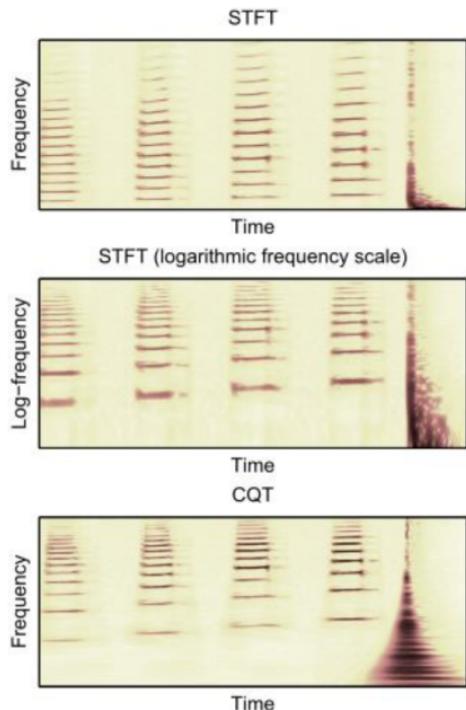
- Il s'agit des différents motifs spectraux qui sont traduits par

- $P_I$  la distribution d'impulsion Impulse)

- Estimation : utilisation d'un algorithme Expectation/ Maximization

# Utilisation de la SI-PLCA pour la détection multi-pitch

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

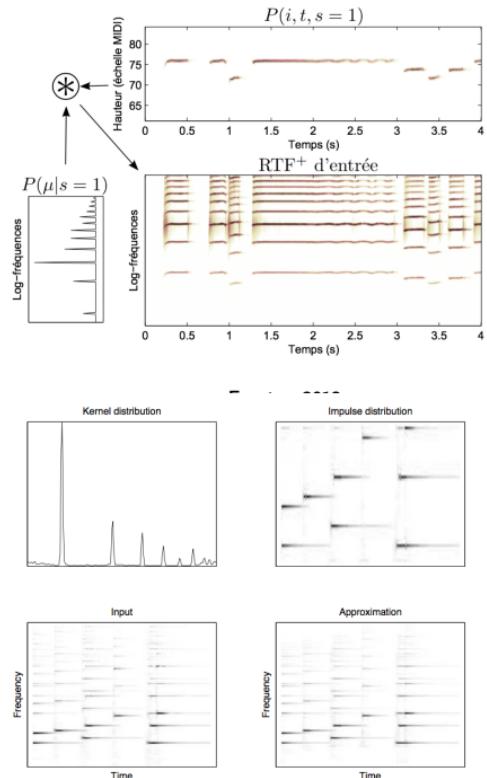


source : Richard, 2012

# Utilisation de la SI-PLCA pour la détection multi-pitch

- Shift invariant PLCA sur une transformée à Q constant

- notation  $i = f$
- $P(f, t) = \sum_z \sum_i P_I(i, t, z) P_K(f' - i | z)$
- peut être ré-écrit comme
  - $P(f, t) = \sum_z P(z) \sum_f P_I(f, t | z) P_K(f' - f | z)$
  - $P(f, t) = \sum_z P(z) \sum_f P_I(f' - f, t | z) P_K(f' | z)$



[B. Fuentes, R. Badeau, and G. Richard. Harmonic adaptive latent component analysis of audio and application to music transcription. IEEE TASLP, 2013.]

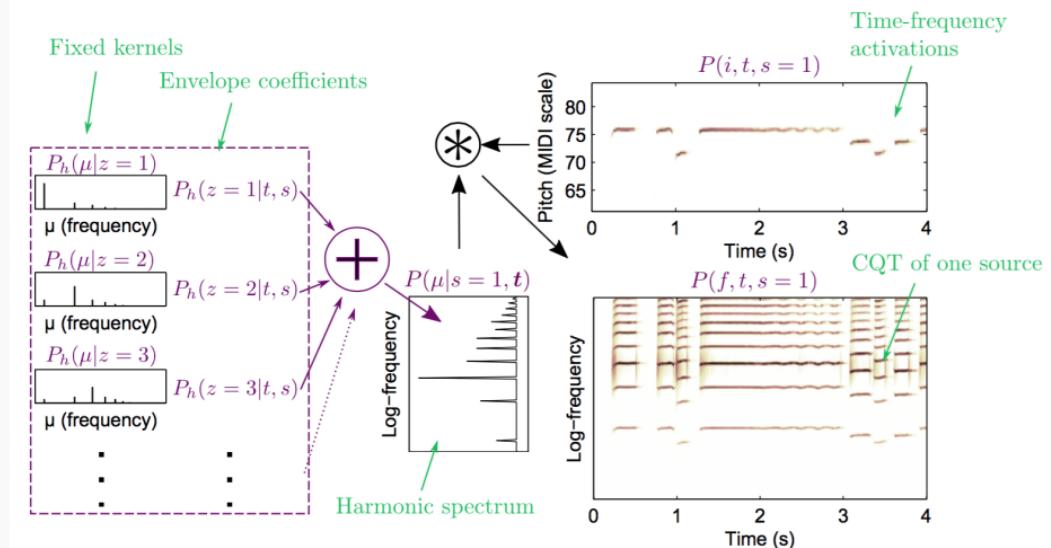
- Subdivision en partie harmonique et bruit

- $P(f, t) = P(c = h)P_h(f, t) + P(c = b)P_b(f, t)$ 
  - $P(c = h)$  : énergie relative de la composante polyphonique harmonique
  - $P(c = b)$  : énergie relative de la composante de bruit

## HALCA Partie harmonique

- Partie harmonique

- $P_h(f, t) = \sum_{s, i, z} P_h(i, t, s) P_h(f - i|z) P_h(z|t, s)$
- $P_h(i, t, s)$  : activation temps-fréquence de chaque source
- $P_h(\mu|z)$  :  $z^{\text{em}}$  noyau harmonique à bande étroite
- $P_h(z|t, s)$  : coefficients de l'enveloppe de la source au temps  $t$



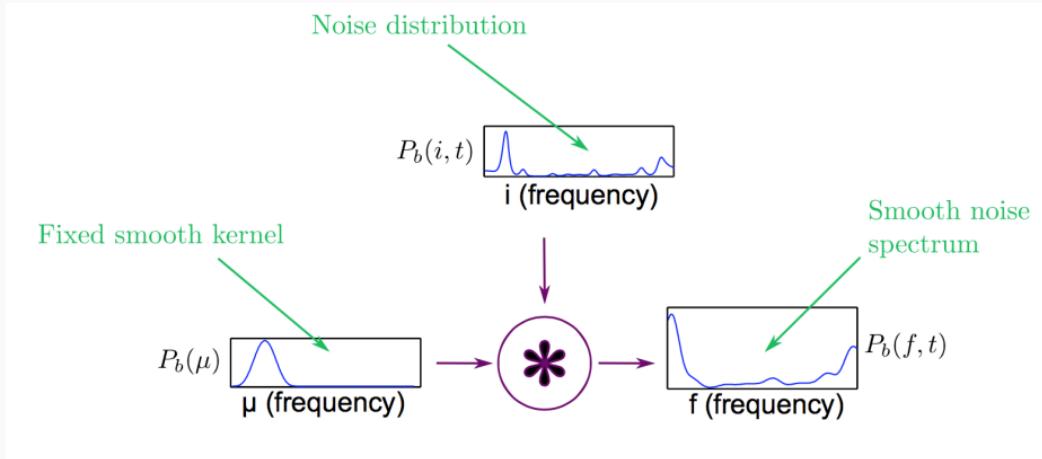
source : Fuentes, 2013

## HALCA Partie bruitée

- Partie bruitée

$$\bullet P_b(f, t) = \sum_i P_b(i, t)P_b(f - i)$$

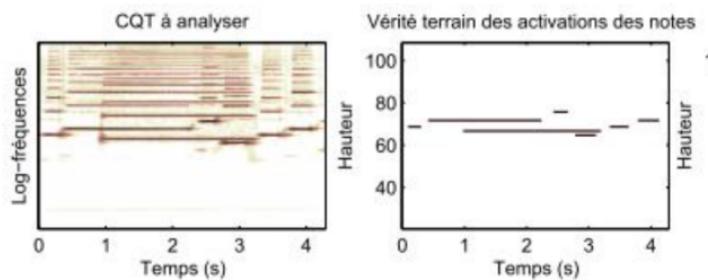
- $P_b(i, t)$  : distribution temps-fréquence du bruit
- $P_b(\mu)$  : noyaux réguliers à bande étroite du bruit



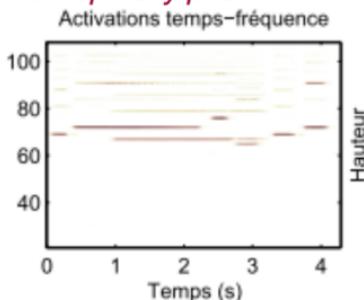
source : Fuentes, 2013

## HALCA contraintes

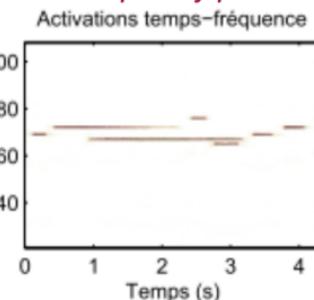
- Utilisation de différents apriori pour les noyaux harmoniques et les activations temporelles



*Without sparsity prior*



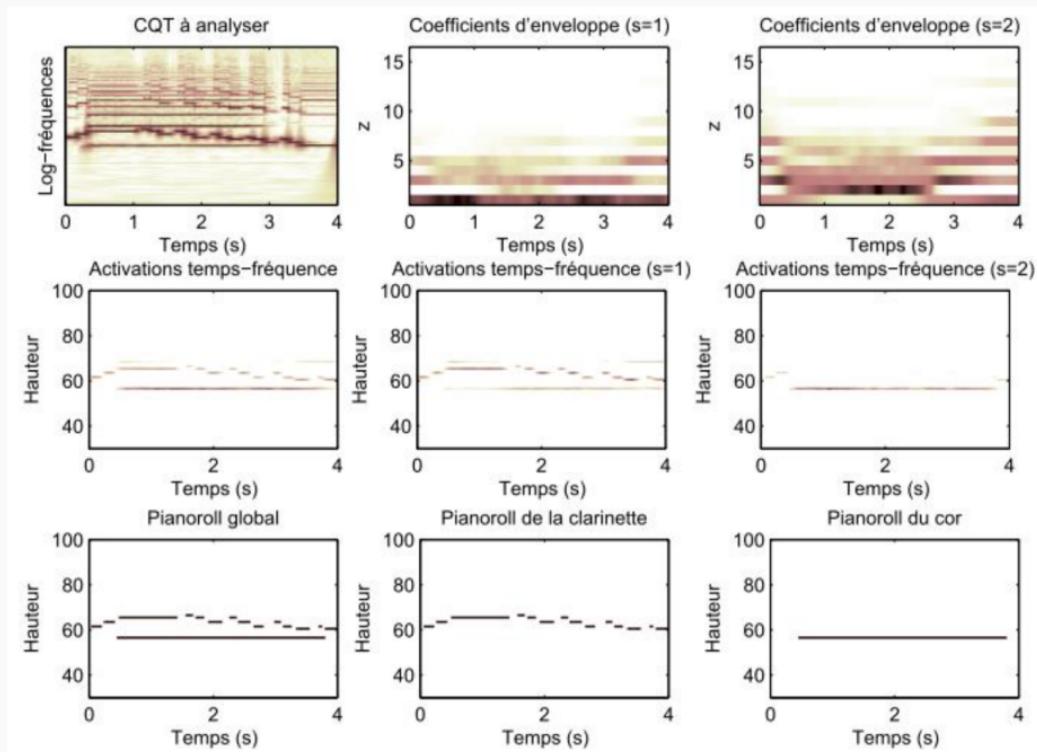
*With sparsity prior*



source : Fuentes, 2013

## HALCA résultats

- Quelques résultats de simulation avec le modèle HALCA



## HALCA résultats

- Résultats de l'évaluation MIREX06 + 6 fichiers Quaero

Algorithm	$\mathcal{F}$ (%)	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	CT ( $\times$ real time)
H	29.9	27.9	37.0	3.4
H-s	31.0	26.6	<b>40.3</b>	4.3
H-st	<b>31.3</b>	27.6	38.6	7.5
Vincent'10	15.8	<b>48.0</b>	10.6	0.9
Dessein'12	16.1	20.1	14.9	0.8

Symbol	Description
H	HALCA model with no prior.
H-t	HALCA model with spectral envelope temporal continuity prior.
H-s	HALCA model with sparseness prior.
H-st	HALCA model with spectral envelope temporal continuity ans sparseness priors.
Vincent'10 [13]	Multiplicative NMF with the Itakura-Saito divergence and harmonicity and spectral smoothness constraints.
Dessein'12 [13]	Spectrogram decomposition on a learned dictionary using $\beta$ -divergence.

