*Ecole d'été en traitement du signal et des images, Peyresq, Juillet 2011*

# Séparation de sources en musique


Laurent DAUDET

Université Paris Diderot / IUF

Institut Langevin

laurent.daudet@espci.fr



Les sons n’ont pas été inclus dans cette présentation. Pour les obtenir à des fins de recherche, merci de me contacter à laurent.daudet@espci.fr

# Avant propos

**Disclaimer:** Cette présentation n'est en aucune manière

- exhaustive
- représentative de l'état de l'art
- un catalogues de mes contributions au domaine

**Tutoriaux:**

- T. Virtanen, *Unsupervised learning methods for source separation in monaural music signals*, in Signal Processing Methods for Music Transcription, A. Klapuri and M. Davy eds, Springer (2006)
- E. Vincent / N. Ono, *Music Source Separation and its Applications to MIR*, tutorial ISMIR 2010
- M. Müller, D. P. Ellis, A. Klapuri and G. Richard, *Signal Processing for Music Analysis*, IEEE JSTSP (2011)
- E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, M. Davies, *Probabilistic Modeling Paradigms for Audio Source Separation,* in Machine Audition: Principles, Algorithms and systems, W. Weng ed, IGI Global (2010).

# Plan du cours

- Pourquoi ? / Modèles de signaux / modèles de mélanges
- Approches “aveugles”
  - CASA
  - Approches additives
  - Modélisation de l’amplitude du spectrogramme
- Approches “semi-aveugles”
- Séparation de sources informée

Pourquoi ? Quels signaux ?

Quels modèles de mélanges ?

# Utilité de la séparation de sources en audio

- post-production musicale / cinéma
  - re-mixage
  - up-mixing (mono ou 2.0 vers 5.1)
  - suppression de dialogues (pour doublage)
  - suppression de musique

- Music Information Retrieval (indexation automatique)
  - identification des locuteurs (archive TV)
  - transcription de voix chantée et/ou instrument dominant pour identification de *cover song*

- Prothèses auditives (“cocktail party effect”)
- Audioconférences, téléphonie, ...

# Séparation de sources en audio

## Quelques définitions

### Qu'entend-on ici par audio ?

Dans cet exposé nous nous limiterons à un cadre *musical* (i.e. hors parole).

Audio typique = le son d'un CD  (mono ou stereo)

### Qu'est ce qu'une source ?

Il n'existe pas de définition unique pour une source sonore.

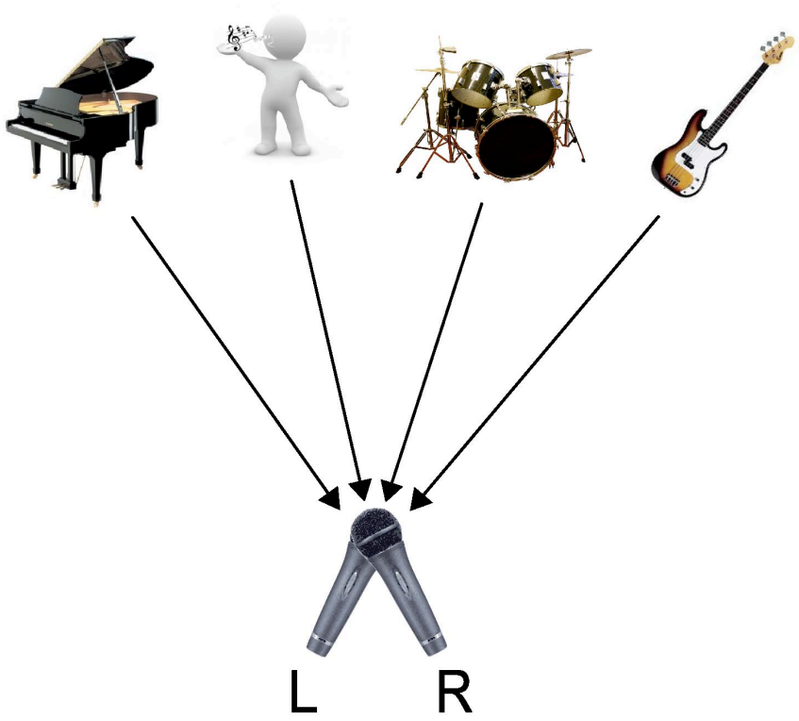
- chaque entité physique vibrante (instrument / partie d'un instrument) ?
- un ensemble de sources associée au même instrument
- un flux sonore *perçu* comme une source individuelle (ex. section des violons dans un orchestre) ?
- une piste entrant dans une table de mixage ?

# Séparation de sources en audio

## Cas 1 du mélange des sources : la prise de son acoustique

### Stereo naturelle / prise de son acoustique

La linéarité des équations de l'acoustique fait que, dans un environnement naturel, les différentes sources sonores s'additionnent dans le signal total capté aux microphones.
Le mix peut varier en fc du temps (sources mobiles)




mélange linéaire instantané

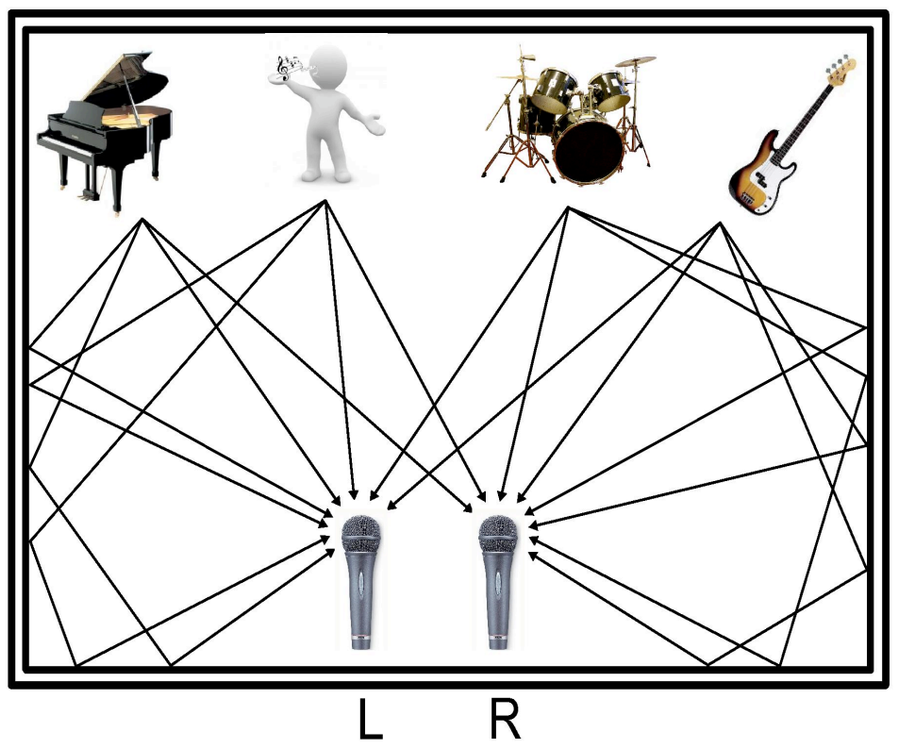
$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \dots \\ x_J(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1I} \\ a_{21} & a_{22} & \dots & a_{2I} \\ \dots & \dots & \dots & \dots \\ a_{J1} & a_{J2} & \dots & a_{JI} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \\ \dots \\ \dots \\ s_I(t) \end{bmatrix}$$

# Séparation de sources en audio

## Cas 1 du mélange des sources : la prise de son acoustique

### Stereo naturelle / prise de son acoustique

La linéarité des équations de l'acoustique fait que, dans un environnement naturel, les différentes sources sonores s'additionnent dans le signal total capté aux microphones.
Le mix peut varier en fc du temps (sources mobiles)




mélange convolutif

$$\mathbf{A}(t) = \begin{bmatrix} h_{11}(t) & \dots & h_{1I}(t) \\ \dots & \dots & \dots \\ h_{J1}(t) & \dots & h_{JI}(t) \end{bmatrix}$$

Nombreux principes de prise de son stéréo

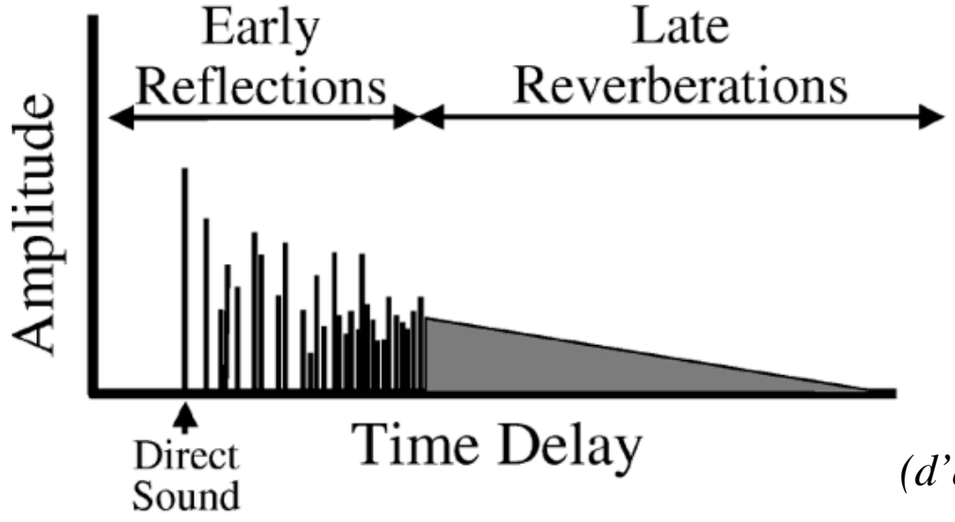
- Couple AB
- Couple XY
- Couple ORTF
- Microphone MS
- Tête artificielle

# Séparation de sources en audio

## Cas 1 du mélange des sources : la prise de son acoustique

### Effets de la réverbération

- Pour des sources ponctuelles, la réverbération agit comme un filtrage (dépendant de l'espace)



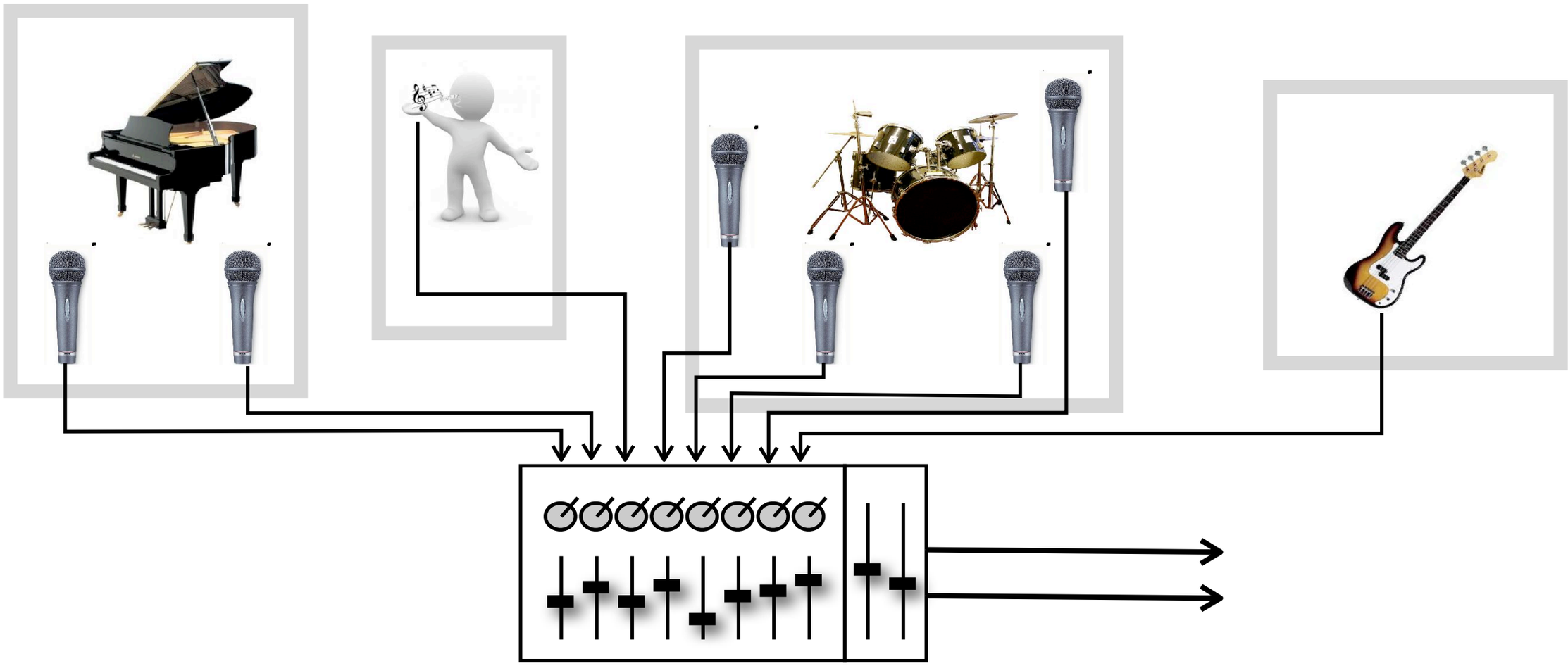

*(d'après E. Vincent et N. Ono)*

dépend des positions
relatives source-récepteur

- Le temps de réverbération peut être très long
- Pour des sources étendues, il faut considérer la somme des contributions de tous les points sources.

# Séparation de sources en audio

## Cas 2 du mélange des sources : stereo a la table de mixage



Chaque instrument (“source”) est enregistré séparément
En général le mix N pistes stereo -> 1 piste stereo ne varie pas en fc du temps, mais pour chaque piste l’ “upmix” mono -> stereo peut inclure des effets de spatialistation (“autopan”)
A tout cela s’ajoute une égalisation + compression de dynamique (“post non-linéaire !”).

# Séparation de sources en musique

En pratique nous serons souvent dans le cas 2 :

- Nombre de sources limité (cas typique musique ‘pop’ 3 à 6, en pratique jamais > 10)
- Source à prendre au sens d’une piste dans une table de mixage (ou un mix de pistes représentant le même instrument) incluant les effets individuels (réverb, panning / autopan, incluant compression du mix), juste avant mélange instantané.
- Forget about independence ...
- Nombre de canaux : 1 ou 2 (presque toujours sous-déterminé)

# Séparation de sources en audio

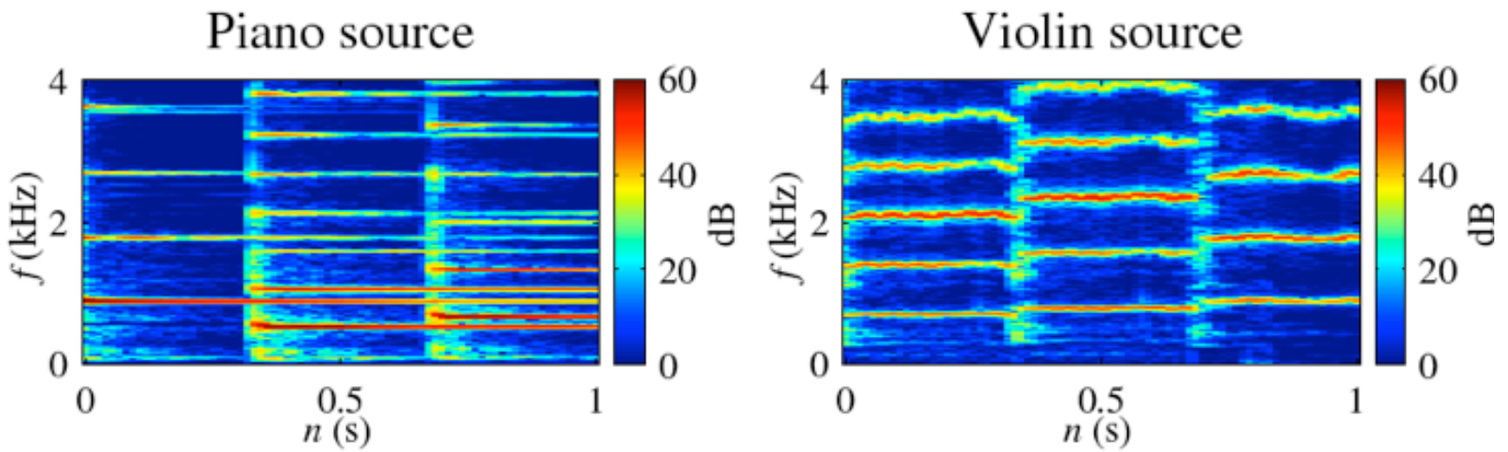
Dans cet exposé, nous ne parlerons pas de mélange convolutif, mais cela reste un problème d'actualité :

( cf par ex. travaux INRIA équipe Gribonval / Vincent - thèses P. Sudhakar -filtres parcimonieux - et Alexis Benichoux - estimation filtres si sources connues - )

Difficultés en particulier du fait de la taille des filtres (RT60 typique 0,1-0,5 s)

# Exemple de source: l'instrument de musique acoustique

- Son créé par la mise en vibration d'un résonateur
  (modèle source-filtre)
- En fonction des caractéristiques de l'excitation et du résonateur on va avoir différentes caractéristiques
  - polyphonique vs monophonique
  - forme temporelle : son entretenu vs son décroissant
  - forme spectrale
  - partie déterministe vs partie aléatoire

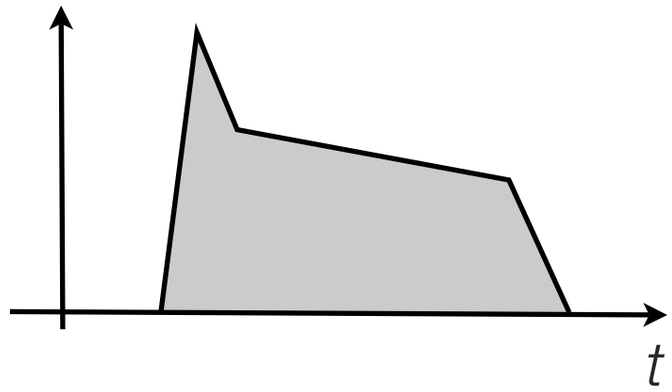


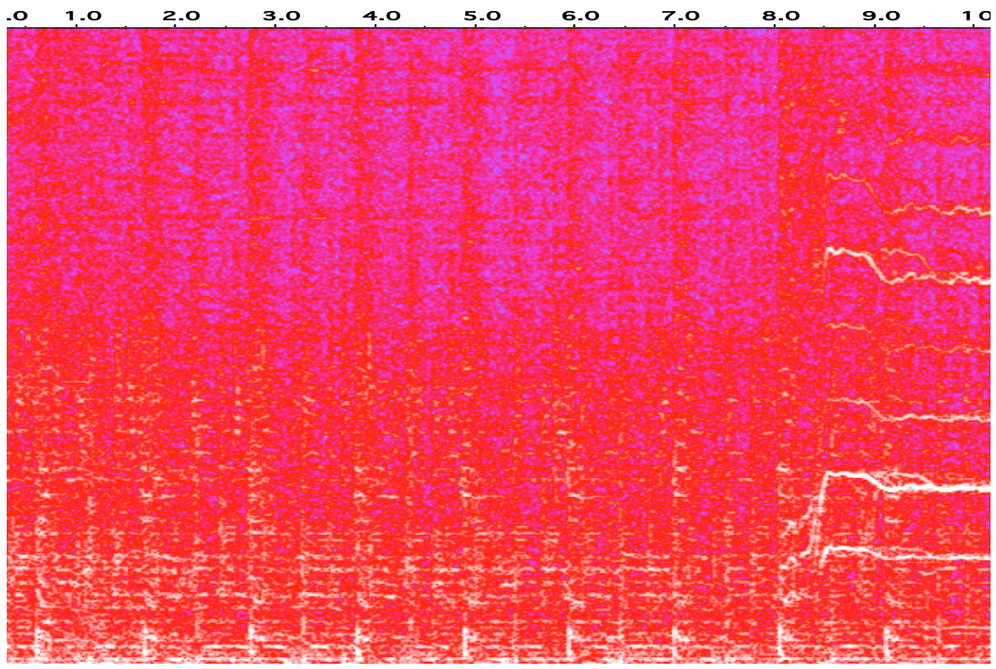

*(d'après E. Vincent et N. Ono)*

# Exemple de source: modèle de note de musique

Composantes d'une note

- amplitude selon modèle ADSR
  (attack - decay - sustain - release)
- modèle sinus + transitoire
- modèle sinus + transitoire + "bruit"



# Evaluation de la séparation

Article E. Vincent, R. Gribonval, C. Févotte, IEEE TSAP 2006

- Paradigme standard : les sources sont estimées à un gain et une permutation près, mais en fonction de l'application les distorsions admissibles peuvent être différentes (permutation / filtrage ...)
- Les critères standard ne permettent pas de distinguer les 3 sources d'erreur
    - interférences des autres sources
    - effet du bruit additif sur l'estimation
    - distorsions du signal reconstruit ("bruit musical")
- Perceptivement ces erreurs ne sont pas équivalentes (à forte distorsion  (c) > (a) > (b) )

# Evaluation de la séparation

un son / 4 distorsions

Original Dist 1 Dist 2 Dist 3 Dist 4

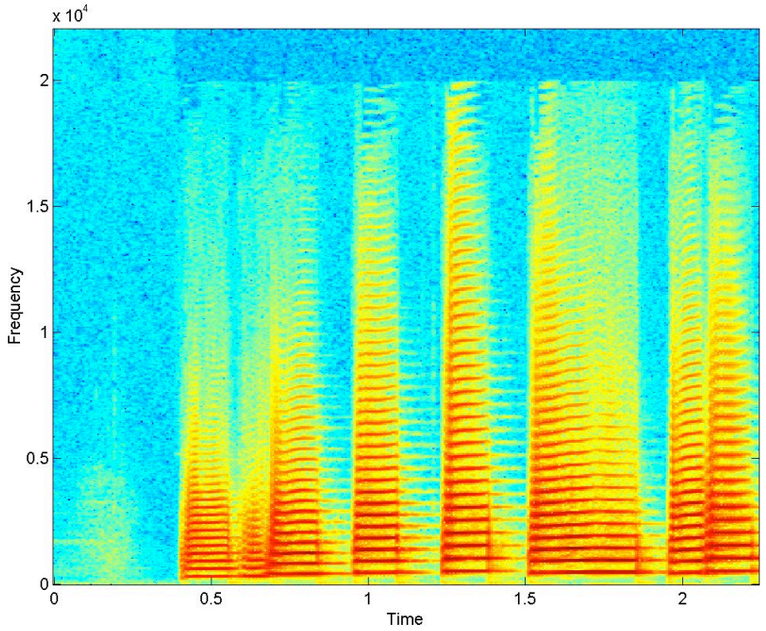
Même SNR (autour de 18dB)

Addition de bruit blanc

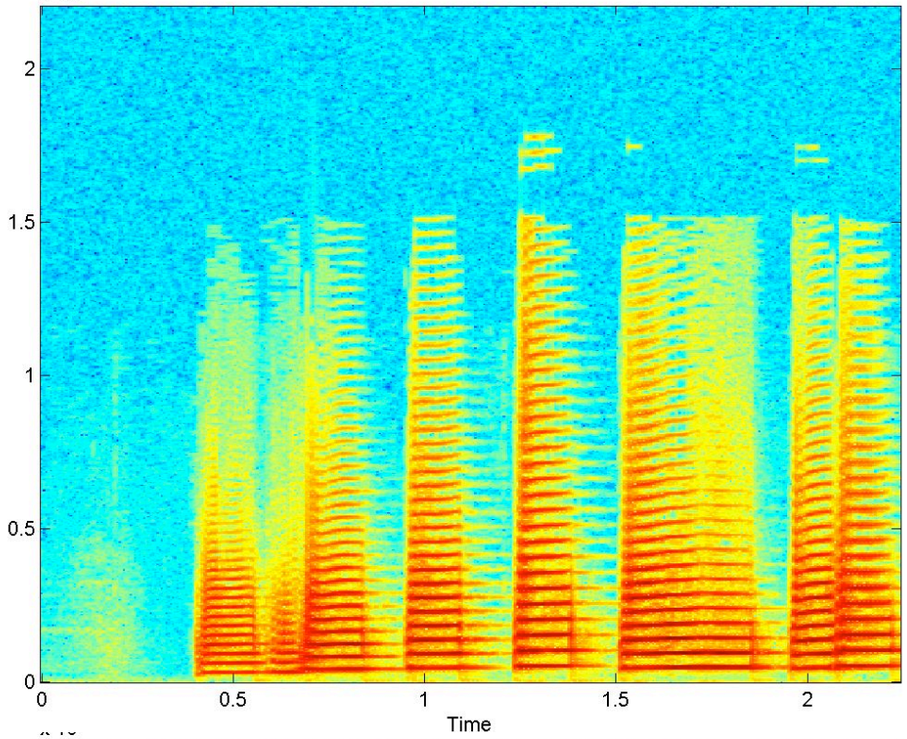
suppression aléatoire bins t-f

Addition de sinus modulé

codage AAC @32kbps

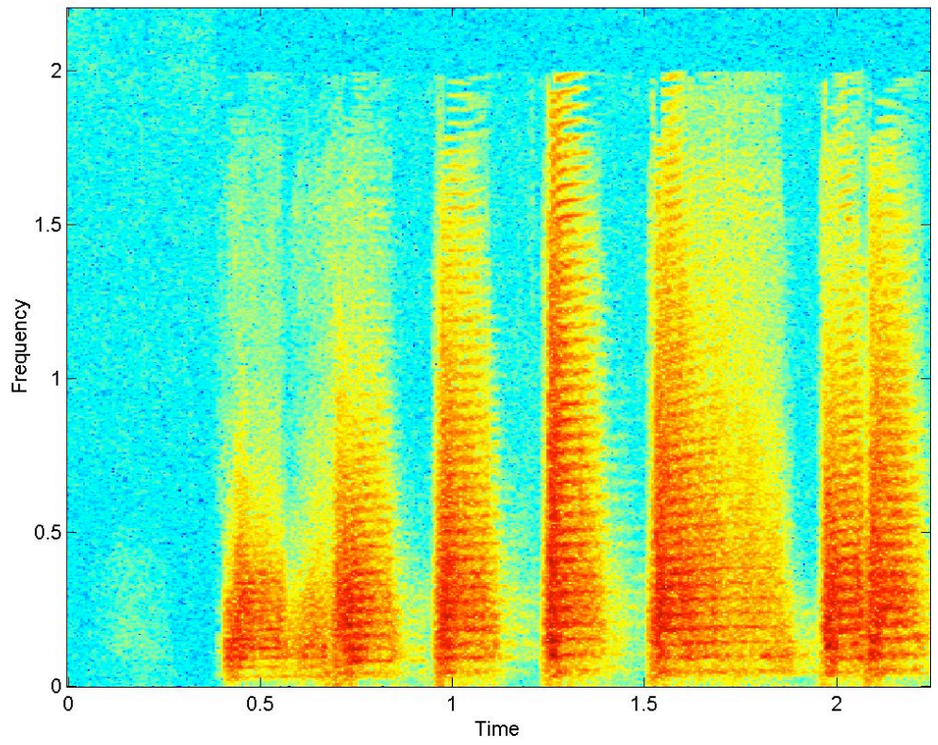



AAC @ 64kbs




original

erreur



# Evaluation de la séparation

- source estimée $\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif}$

- cas linéaire instantané: on définit 3 opérateurs de projection

$$
\begin{aligned}
P_{s_j} &:= \Pi\{s_j\}, \\
P_{\mathbf{s}} &:= \Pi\{(s_{j'})_{1\leq j'\leq n}\}, \\
P_{\mathbf{s,n}} &:= \Pi\{(s_{j'})_{1\leq j'\leq n}, (n_i)_{1\leq i\leq m}\}
\end{aligned}
$$

$$
\begin{aligned}
s_{\text{target}} &:= P_{s_j}\widehat{s}_j, \\
e_{\text{interf}} &:= P_{\mathbf{s}}\widehat{s}_j - P_{s_j}\widehat{s}_j, \\
e_{\text{noise}} &:= P_{\mathbf{s,n}}\widehat{s}_j - P_{\mathbf{s}}\widehat{s}_j, \\
e_{\text{artif}} &:= \widehat{s}_j - P_{\mathbf{s,n}}\widehat{s}_j.
\end{aligned}
$$

# Evaluation de la séparation

- 4 indicateurs de performance

$$\mathrm{SDR} := 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}}+e_{\mathrm{noise}}+e_{\mathrm{artif}}\|^2} \qquad \mathrm{SNR} := 10\log_{10}\frac{\|s_{\mathrm{target}}+e_{\mathrm{interf}}\|^2}{\|e_{\mathrm{noise}}\|^2},$$

$$\mathrm{SIR} := 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}}\|^2}, \qquad \mathrm{SAR} := 10\log_{10}\frac{\|s_{\mathrm{target}}+e_{\mathrm{interf}}+e_{\mathrm{noise}}\|^2}{\|e_{\mathrm{artif}}\|^2}$$

- extension à d'autres indéterminations (filtres, g(t), ...)
- toolbox matlab : BASS
- entension à des critères "perceptifs" (Emiya et al.)
- suppose connue la vérité terrain
- imparfaitement corrélé à perception

# Comparaisons SISEC

http://sisec.wiki.irisa.fr/tiki-index.php?page=Audio+source+separation

# Approches “aveugles”

# Dans quel domaine ?

Une classification possible des méthodes est selon le **domaine** dans lequel s'effectue la transformation

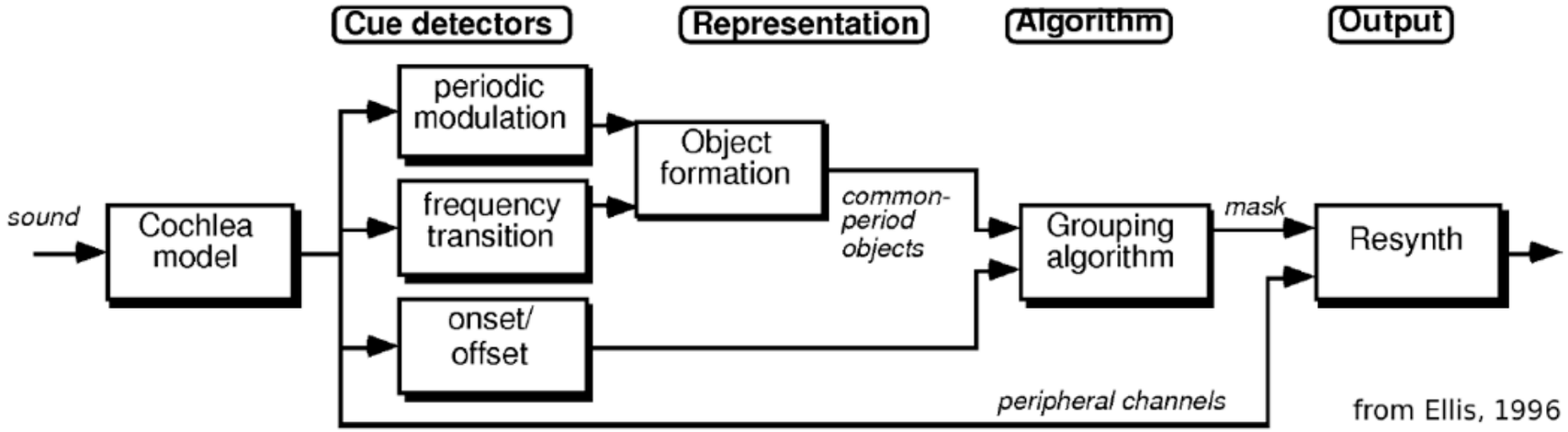
- Modèle "perceptif" (Computational Auditory Scene Analysis)
- Modèle additif (ICA - SCA)
- Modélisation de l'amplitude du spectrogramme

# Approches “aveugles”

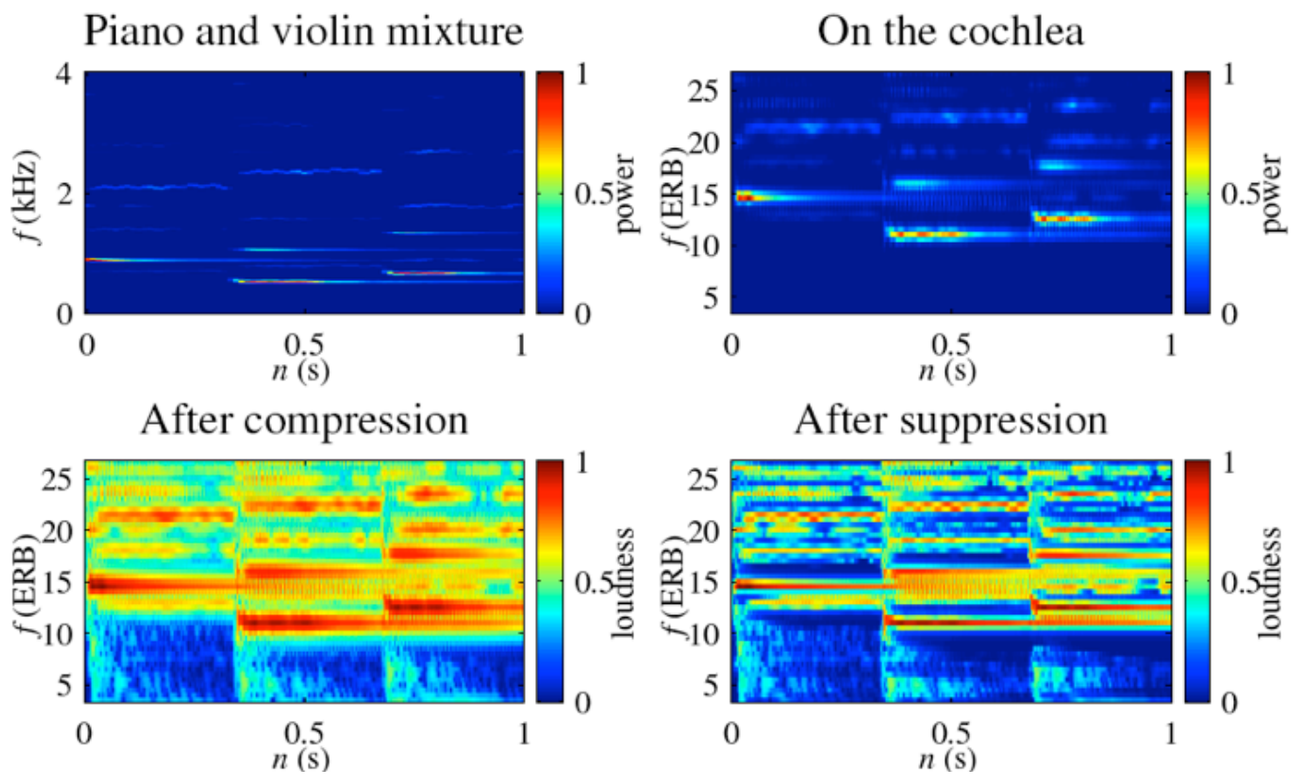
- Modèle “perceptif” (Computational Auditory Scene Analysis)
- Modèle additif (ICA - SCA)
- Modélisation de l’amplitude du spectrogramme

# Computational Auditory Scene Analysis

- Cherche à émuler les mécanismes de perception humaine



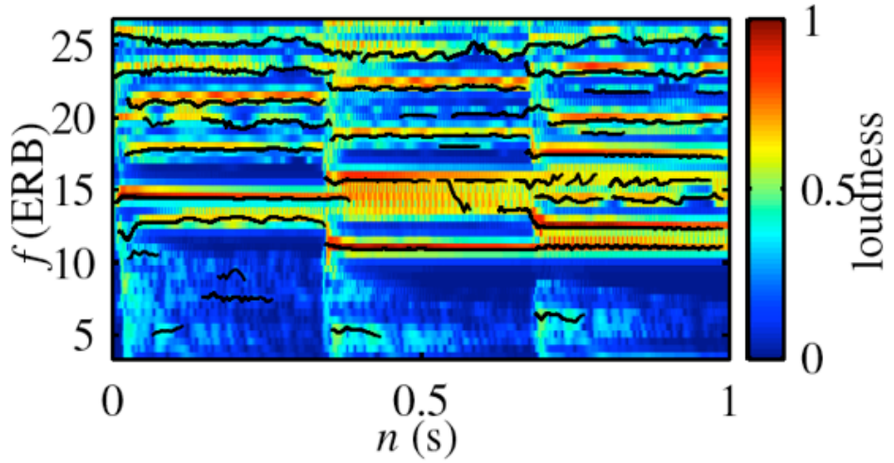

- oreille externe et canal auditif (filtrage)
- banc de filtres (cochlée)
- rectification / compression / inhibition latérale (cellules cillées à la base du nerf)



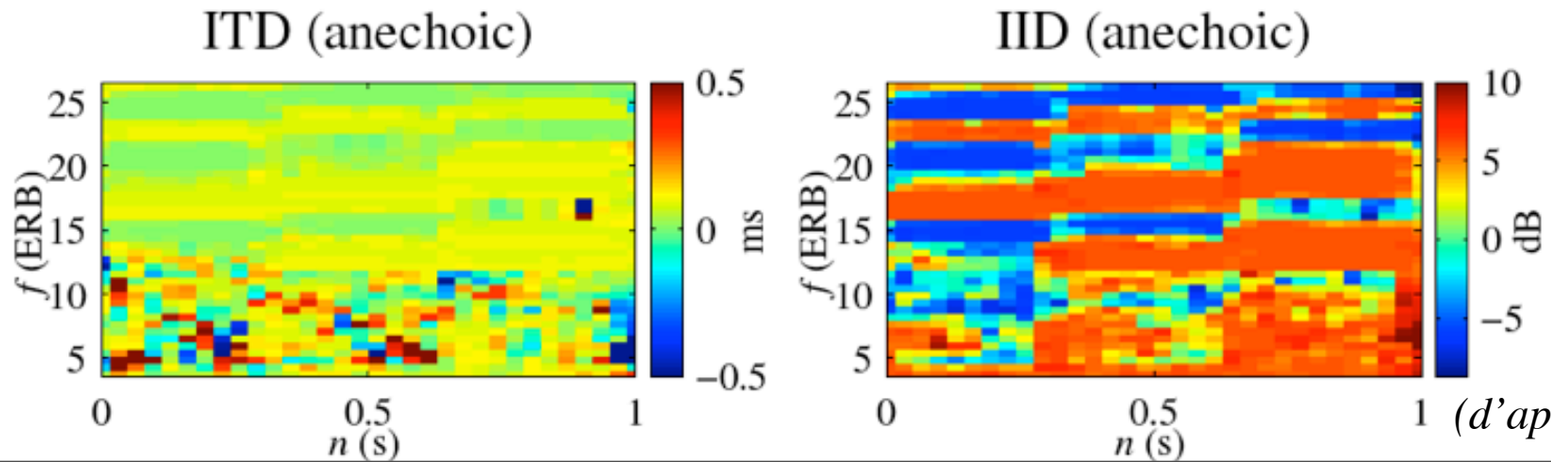

*(d'après E. Vincent et N. Ono)*

# Computational Auditory Scene Analysis

- modèle sinusoïdal



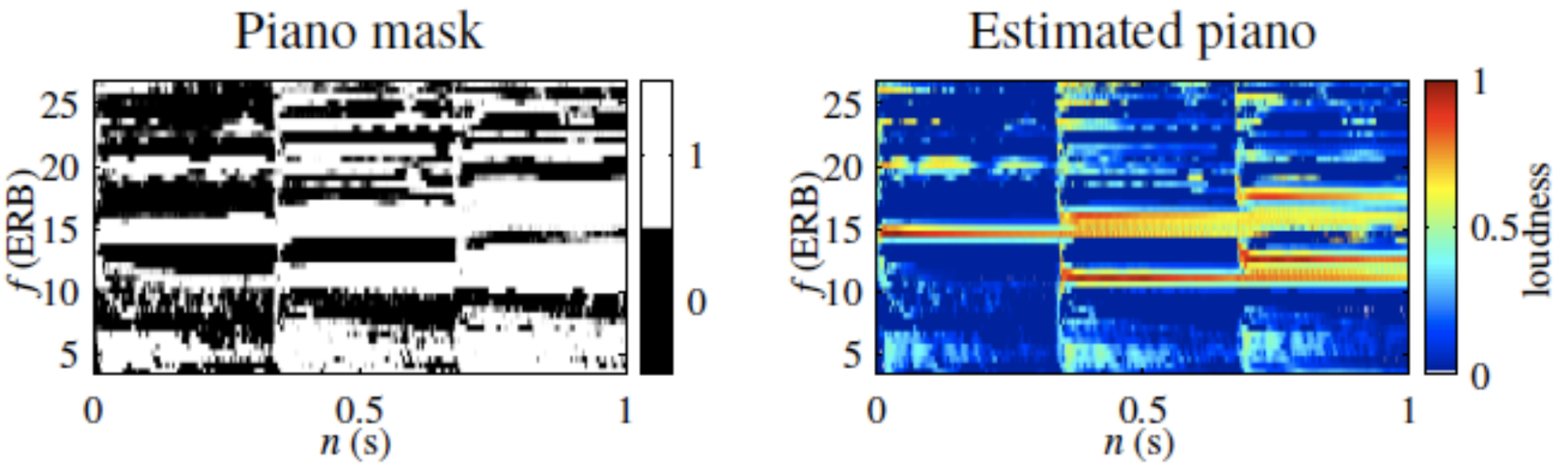

- groupement des composantes des sources selon des formes (Gestalt)
  - proximité
  - continuité
  - similarité
  - fermeture
- groupement spatial selon
  - ITD (Interaural Time Difference)
  - IID (Interaural Intensity Difference)




*(d'après E. Vincent et N. Ono)*

# Computational Auditory Scene Analysis

- séparation par masquage temps-fréquence




- qualité limitée, indices souvent empiriques
- mais qualité croissante des modèles perceptifs

*(d'après E. Vincent et N. Ono)*

# Approches “aveugles”

- Modèle “perceptif” (Computational Auditory Scene Analysis)
- Modèle additif (ICA - parcimonie)
- Modélisation de l’amplitude du spectrogramme

# Modèles parcimonieux

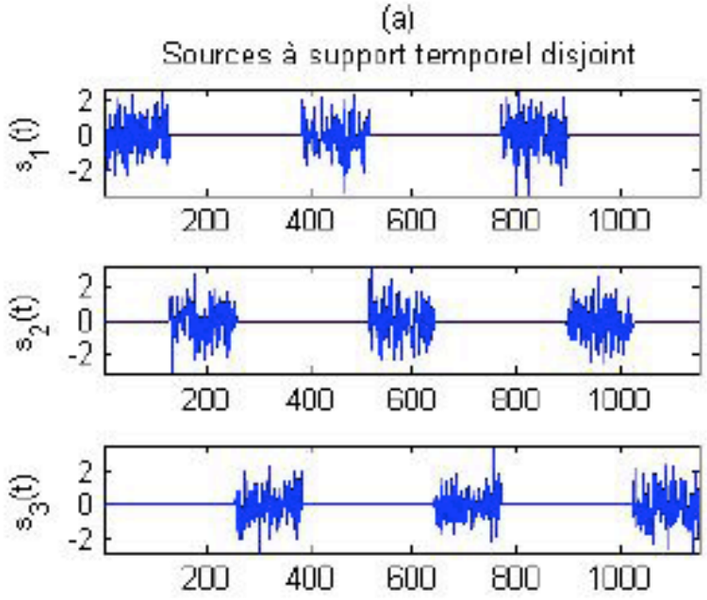
## Estimation MAP sous hypothèse de sources indépendantes Laplaciennes (cf cours Jutten/Comon)

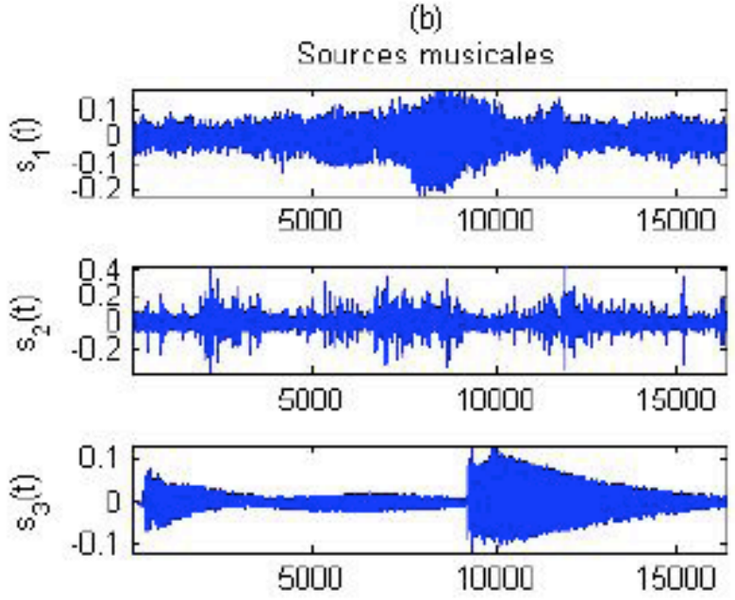
$$\min_{A,S} \frac{1}{2\sigma^2} \|AS - X\|_2^2 + \sum_{j,t} |s_j(t)|$$

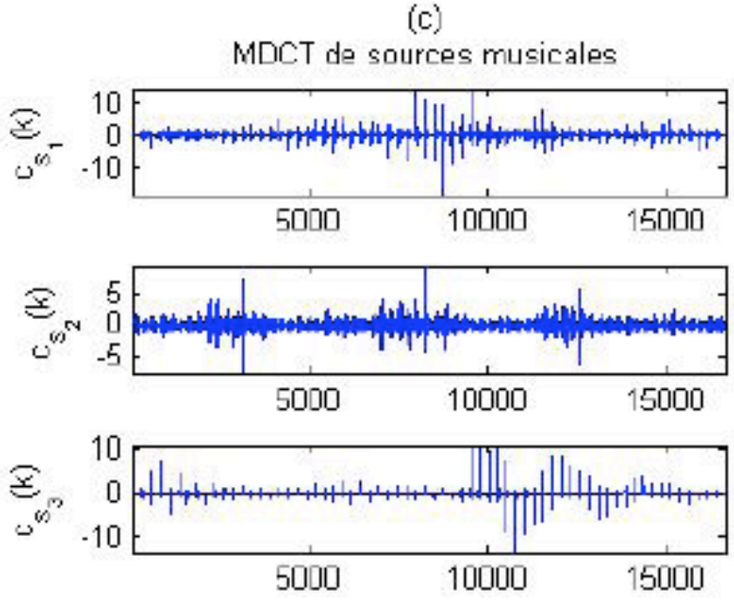
En général (Gribonval, Bofill & Zibulevsky 01), on procède en 2 étapes :

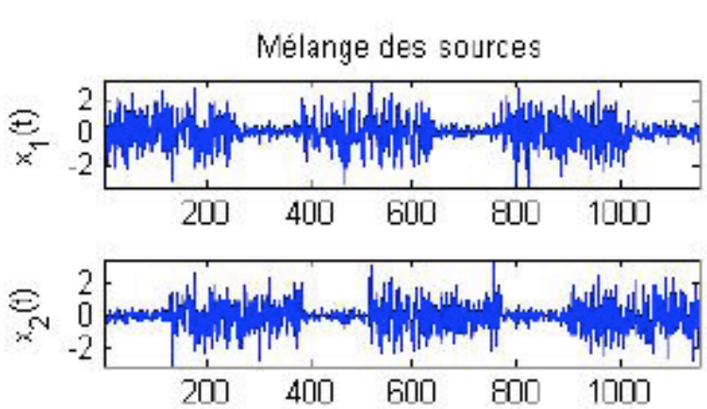
- estimation de A : Â
- estimation de S connaissant Â

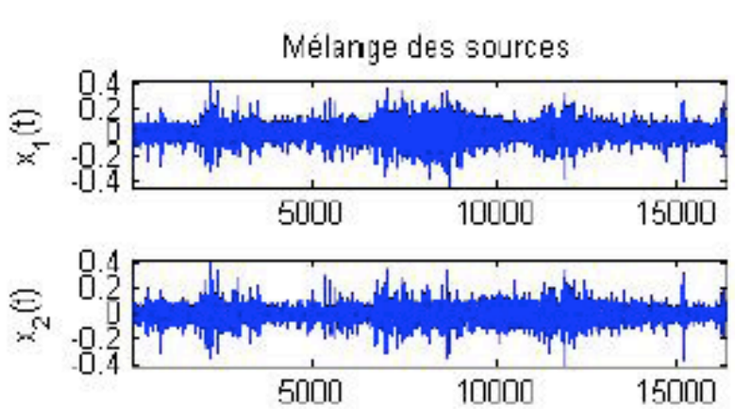
# Modèles parcimonieux

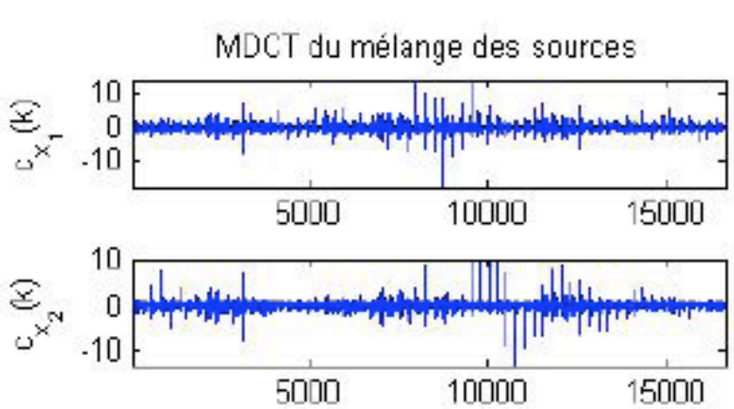


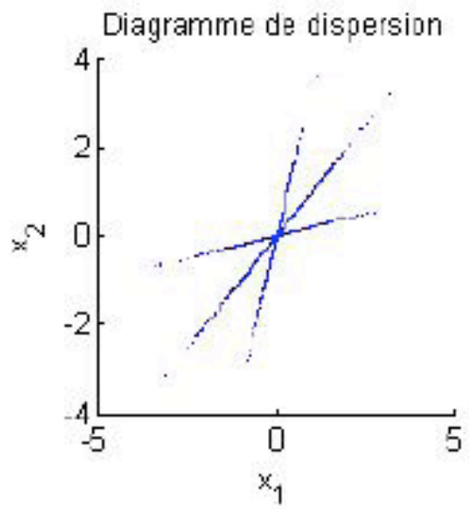



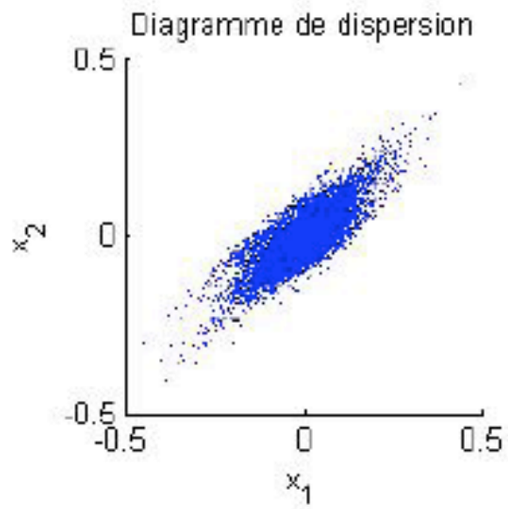



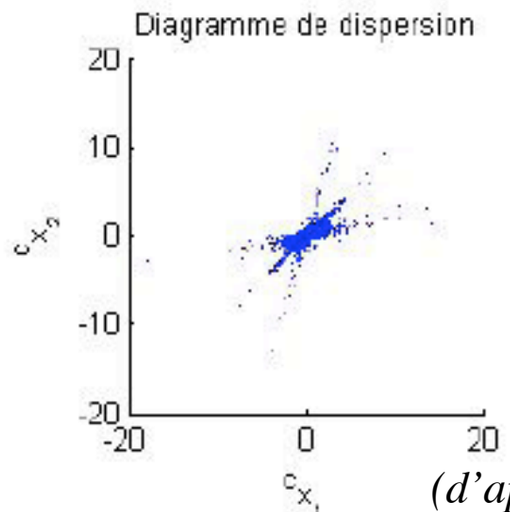



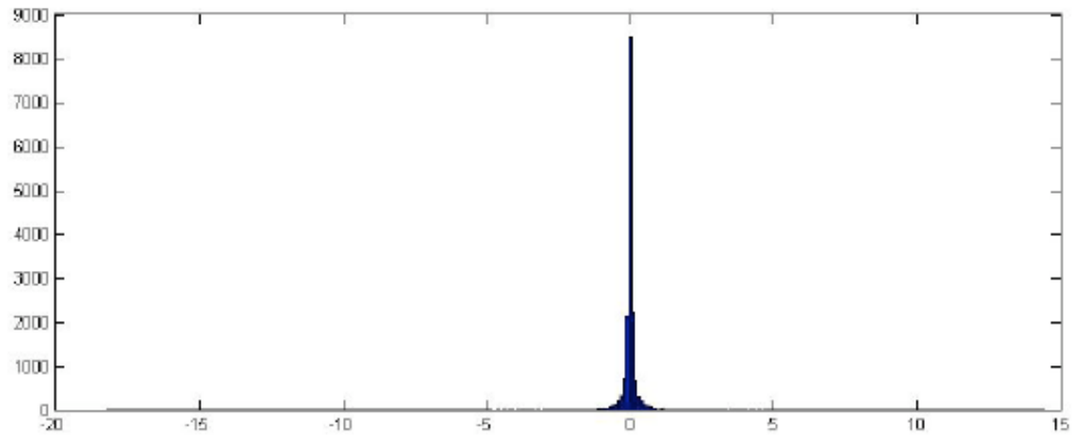




*(d'après R. Gribonval)*

# Modèles parcimonieux

- observation empirique

Histogramme des coeffs
MDCT source 1




*(d'après R. Gribonval)*

- modélisation par Gaussiennes généralisées $p_c(c) \propto \exp(-\alpha |c|^\tau)$
- dictionnaires

Orthogonal : la MDCT (DCT avec fenetre recouvrante)

Redondant : atomes de Gabor

$$\varphi_{n,f}(t) = w(t - nT) \exp(2j\pi ft)$$

Redondance en temps/fréquence/échelle

Dictionnaires appris

- Evidence biologique (Smith/Lewicki) Gammatones, ...

# Modèles parcimonieux

Estimation :

- soit par une transformée linéaire
  - mise en oeuvre simple
  - parcimonie souvent faible (surtout dans le cas redondant)
- soit par une transformée non linéaire minimisant la parcimonie **jointe** $L\tau$ $(0 \leq \tau \leq 1)$

exacte $$C_{\mathbf{x}}^{\tau} := \arg \min_{C_{\mathbf{x}} | C_{\mathbf{x}} \mathbf{\Phi} = \mathbf{x}} \sum_{k=1}^{K} \left( \sum_{p=1}^{P} \left| c_{x_p}(k) \right|^2 \right)^{\tau/2}$$

approchée $$C_{\mathbf{x}}^{\tau,\lambda} := \arg \min_{C_{\mathbf{x}}} \left\{ \| \mathbf{x} - C_{\mathbf{x}} \mathbf{\Phi} \|_F^2 + \lambda \sum_{k=1}^{K} \left( \sum_{p=1}^{P} \left| c_{x_p}(k) \right|^2 \right)^{\tau/2} \right\}$$

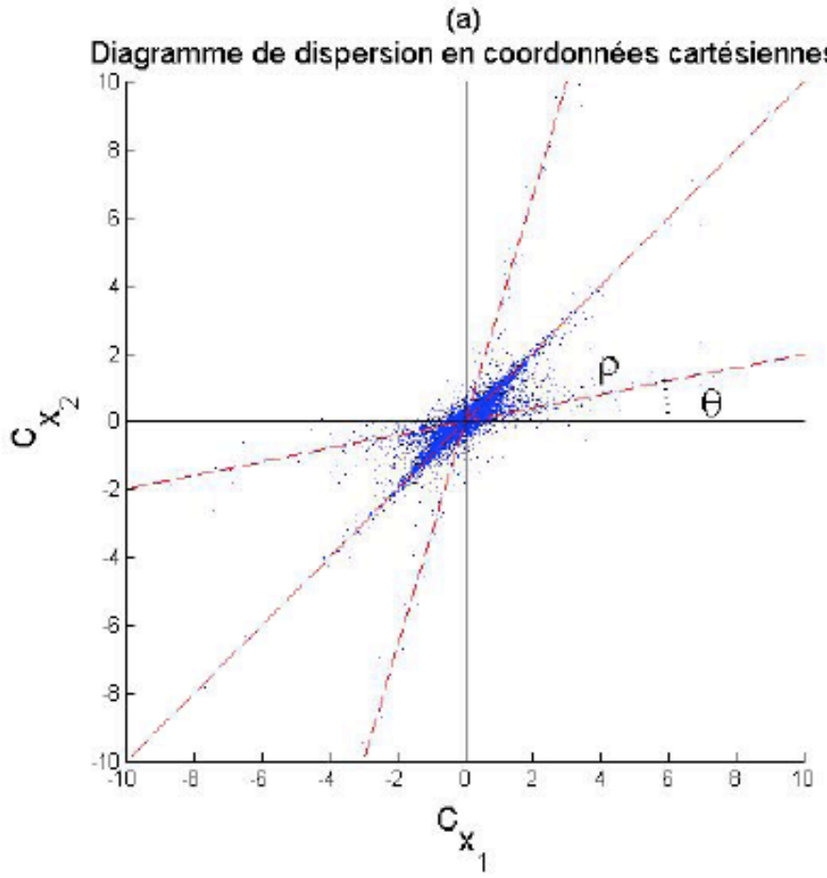
# Modèles parcimonieux

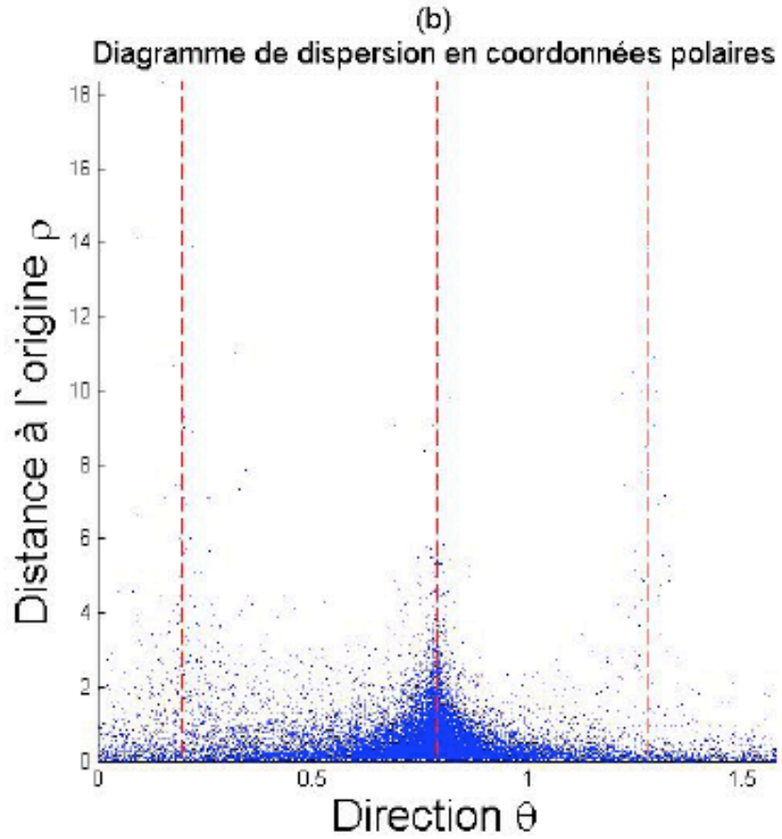
Algorithmes

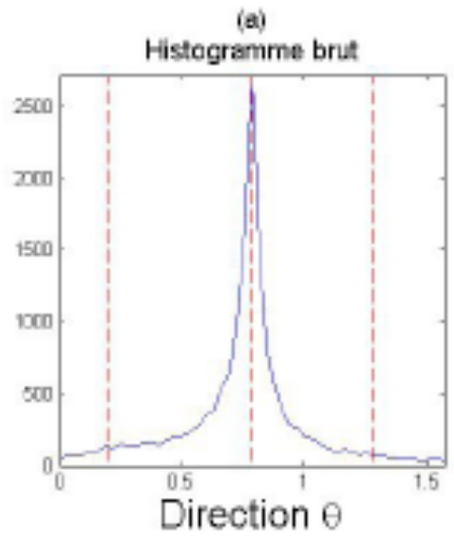
- moindres carrés pondérés (M-FOCUSS)
- $\tau = 1$ Basis Pursuit, BPDN
- Iterated (soft/hard) thresholding
- greedy (MP - aka CLEAN -, OMP, GP, CMP, OOMP, LOMP, ... )
- ...

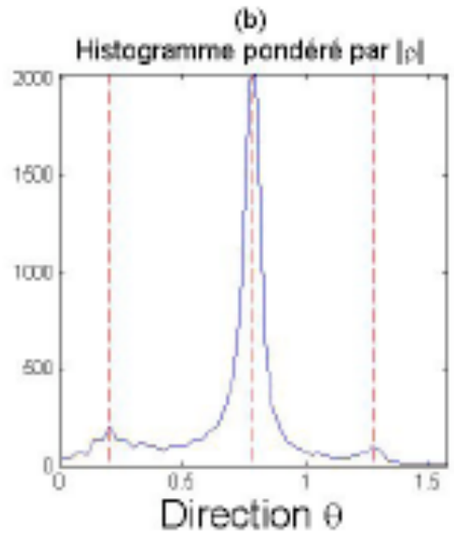
# Modèles parcimonieux

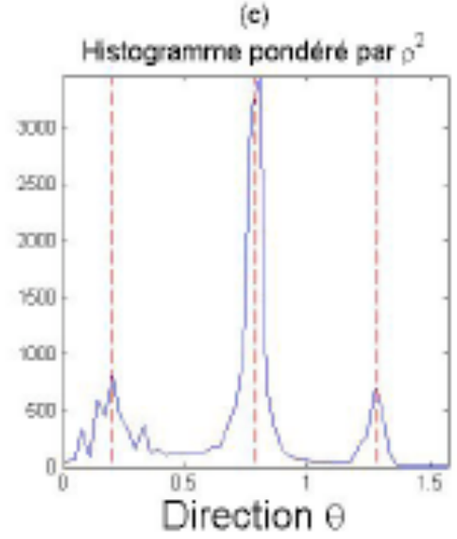
## Estimation de la matrice de mélange




*(d'après R. Gribonval)*

# Modèles parcimonieux

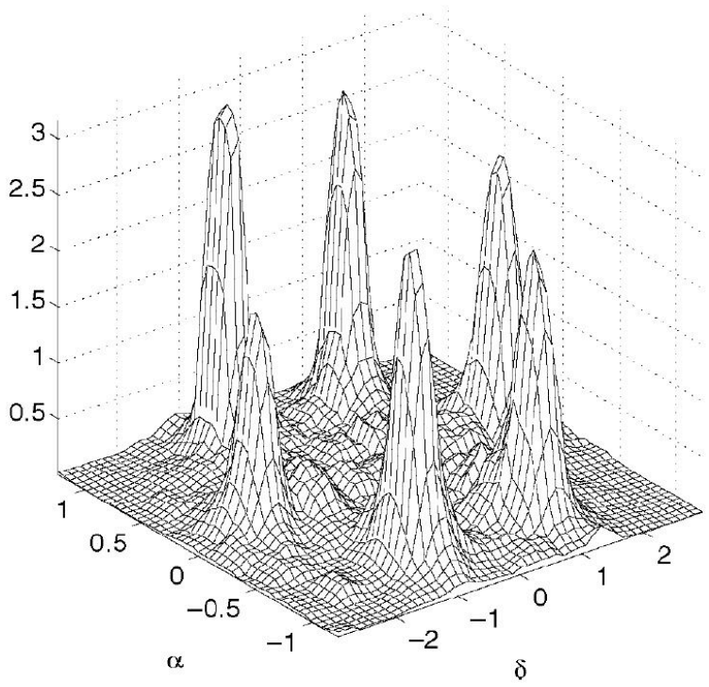
## Séparation par masquage binaire

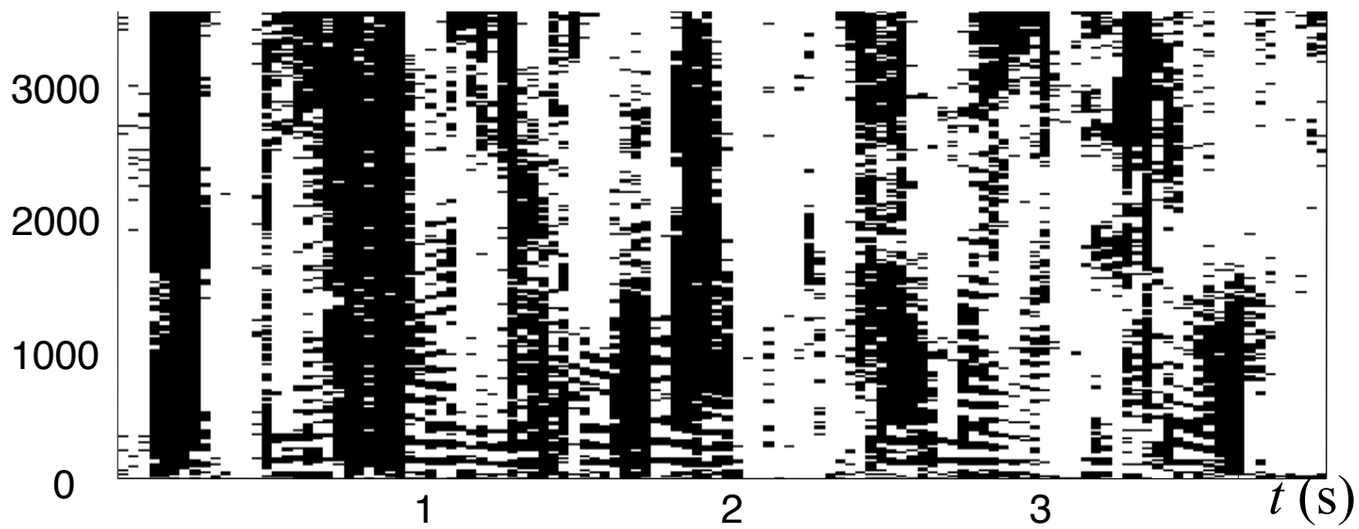
## Séparation par masquage binaire

On suppose qu'il n'existe qu'un seule source active en chaque point (t,f)

$$\chi_n(k) := \begin{cases} 1 & \text{si } n = n(k); \\ 0 & \text{sinon.} \end{cases} \qquad \text{avec } n(k) := \arg\max_n \frac{|\widehat{\mathbf{A}}_n{}^{\mathrm{H}} C_{\mathbf{x}}(k)|}{\|\widehat{\mathbf{A}}_n\|_2^2}$$

- DUET (Yilmaz & Rickard, 2004) : une seule source présente à chaque $(t, f)$

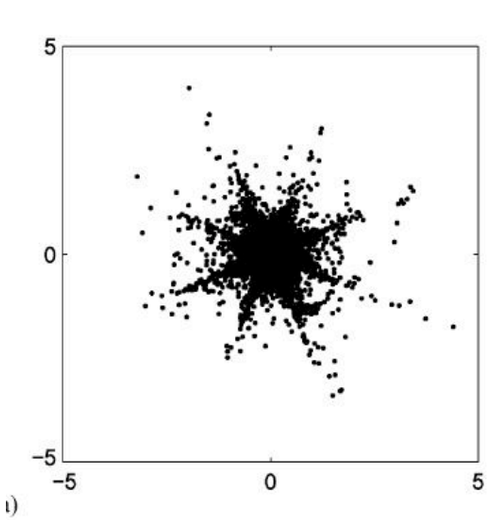


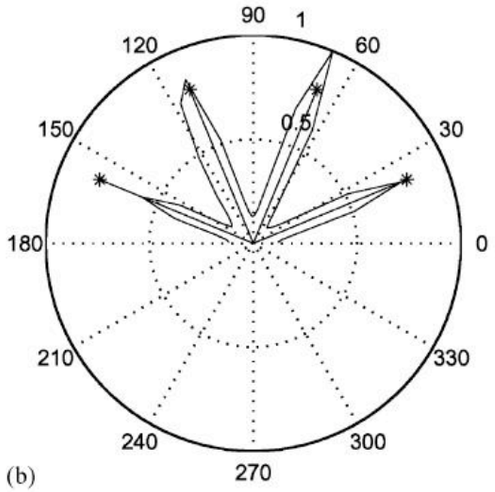



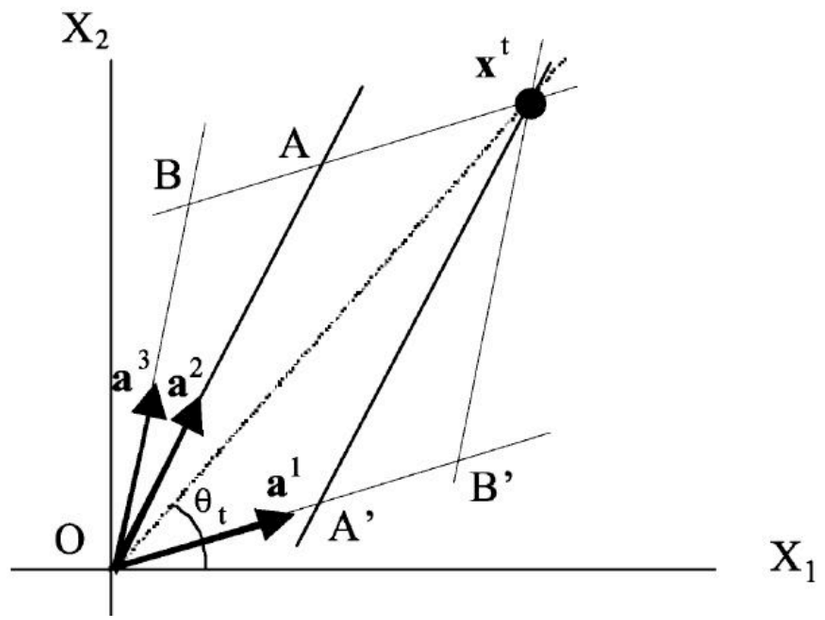

*(d'après R. Gribonval)*

# Modèles parcimonieux

## Bofill / Zibulevsky (2001) :
## 2 canaux - 2 sources présentes à chaque point t-f



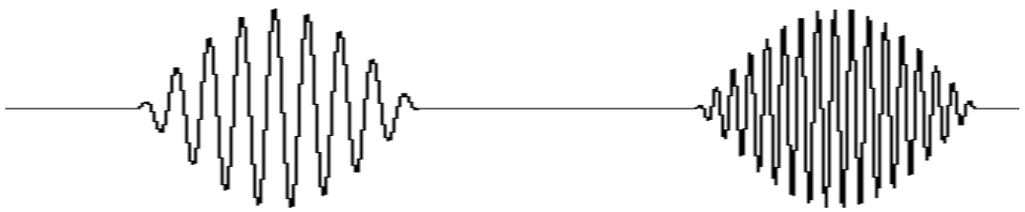



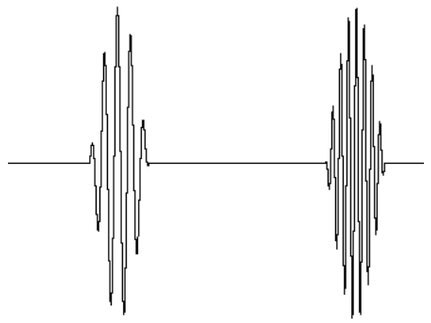

sources originales mélange sources estimées

# Séparation parties sinusoïdale / transitoires

- On cherche à séparer les parties tonales et transitoires d'un enregistrement
- On suppose que l'on a identifié 2 bases, pour chacun d'elles l'une des composantes est sparse (et l'autre non)
- Idée similaire en image : Morphological Component Analysis (Starck et al.)
- Atomes MDCT à 2 résolutions



Grandes fenêtres pour la partie tonale



Petites fenêtres pour la partie transitoire

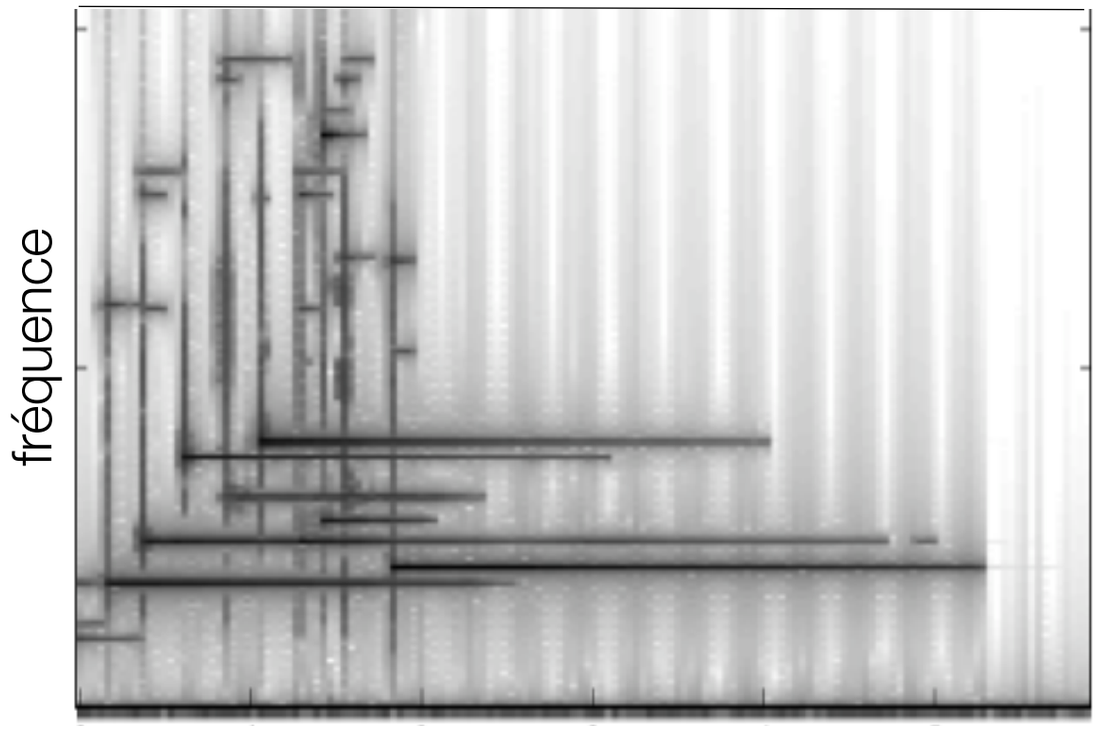
FIRSP: Fast Iteratively Reweighted SParsifier [M. Davies & L.D.,02]

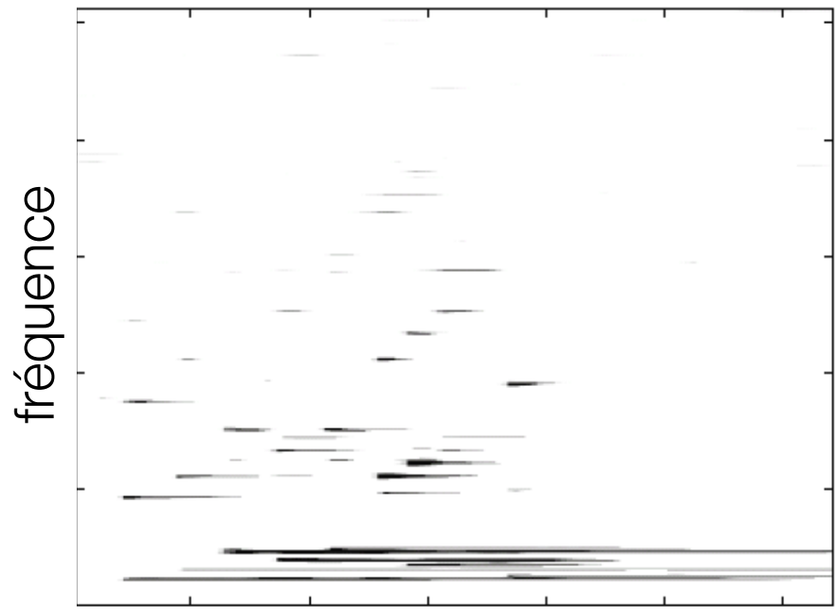
$$u = \arg\min_{u} ||x - \Phi u||_2^2 + \lambda \|u\|_p^p$$

Peut être résolu efficacement avec une variante de l'algorithme EM si $\Phi$ est une union de bases orthogonales

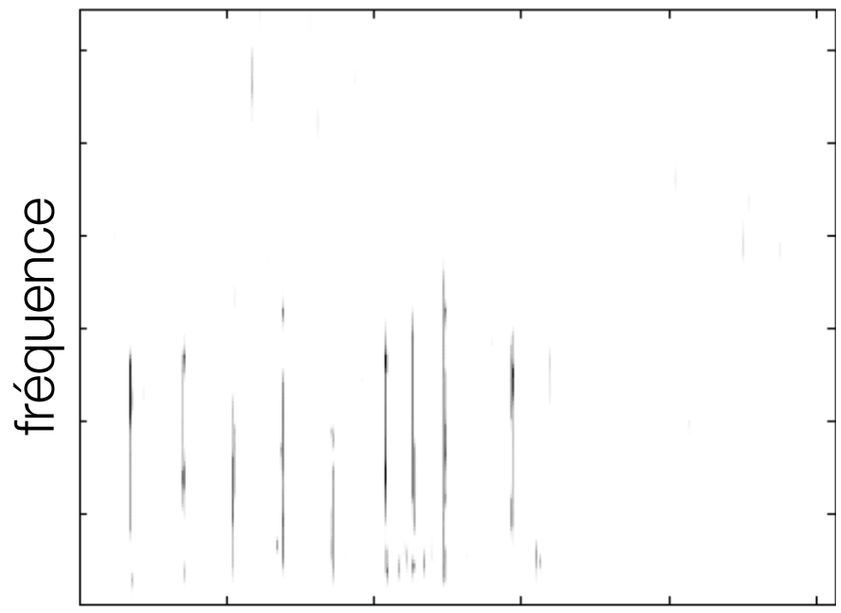
# Estimation simultanée par méthode jointe

Spectrogramme simple résolution



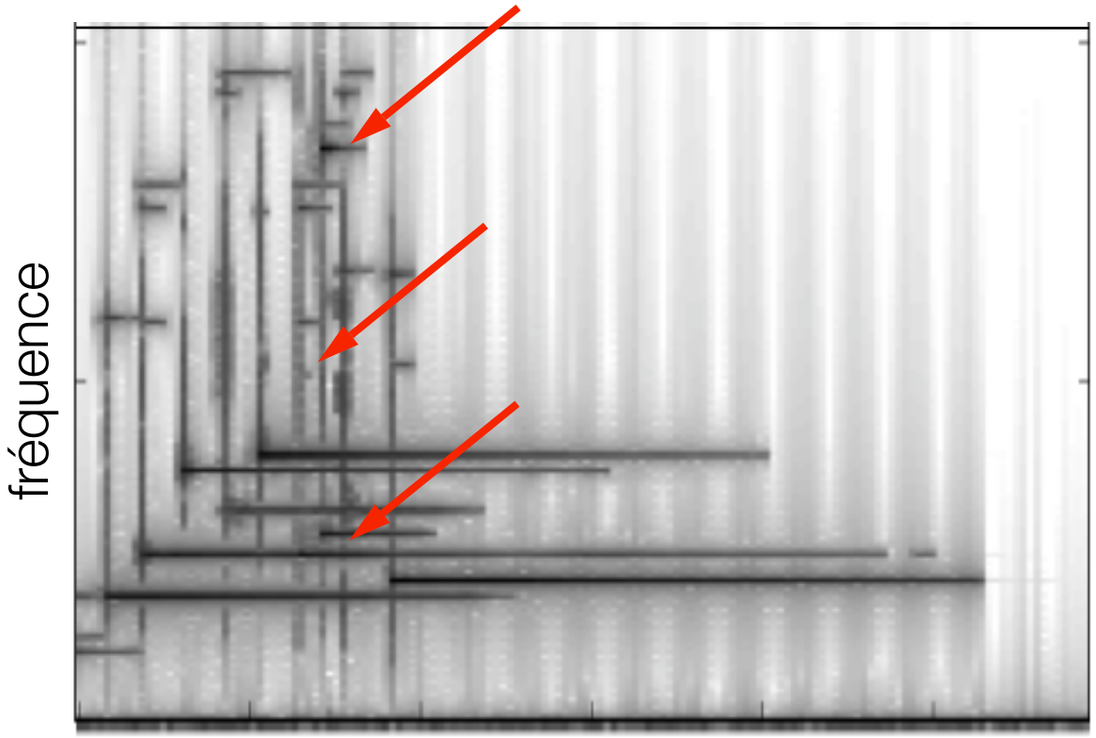




2 “cartes de significance”



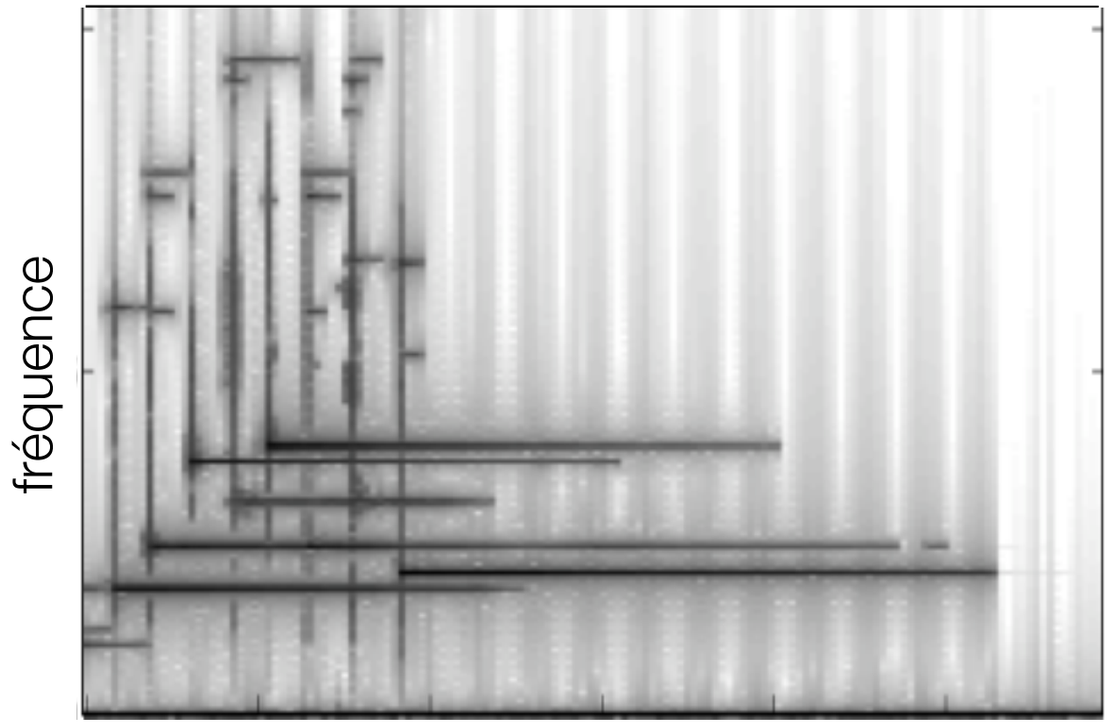
# Estimation simultanée par méthode jointe

## Edition audio : enlever une note




Original

Approximation parcimonieuse




signal retouché

7e note

# Exension avec modèle structuré

## Chaines de Markov horizontales / verticales

(S. Molla /B.Torrésani, C.Févotte/B.Torrésani/LD/S.Godsill IEEE TSALP 2008)

## Parcimonie structurée

(normes mixtes B. Torrésani/M. Kowalski)

# Approches “aveugles”

- Modèle “perceptif” (Computational Auditory Scene Analysis)
- Modèle additif (ICA - SCA)
- Modélisation de l’amplitude du spectrogramme

# Principe

On modélise le module carré de la STFT que l'on va (abusivement) supposer additif

Si les phases des sources sont indépendantes et distribuées de manière uniforme

$$E(|Y(k)|^2) = |X_1(k)|^2 + |X_2(k)|^2$$

- choix de fenêtre
- module donne souvent de meilleurs résultats
- approximation qui ne va pas sans poser de problèmes ultérieurs ...

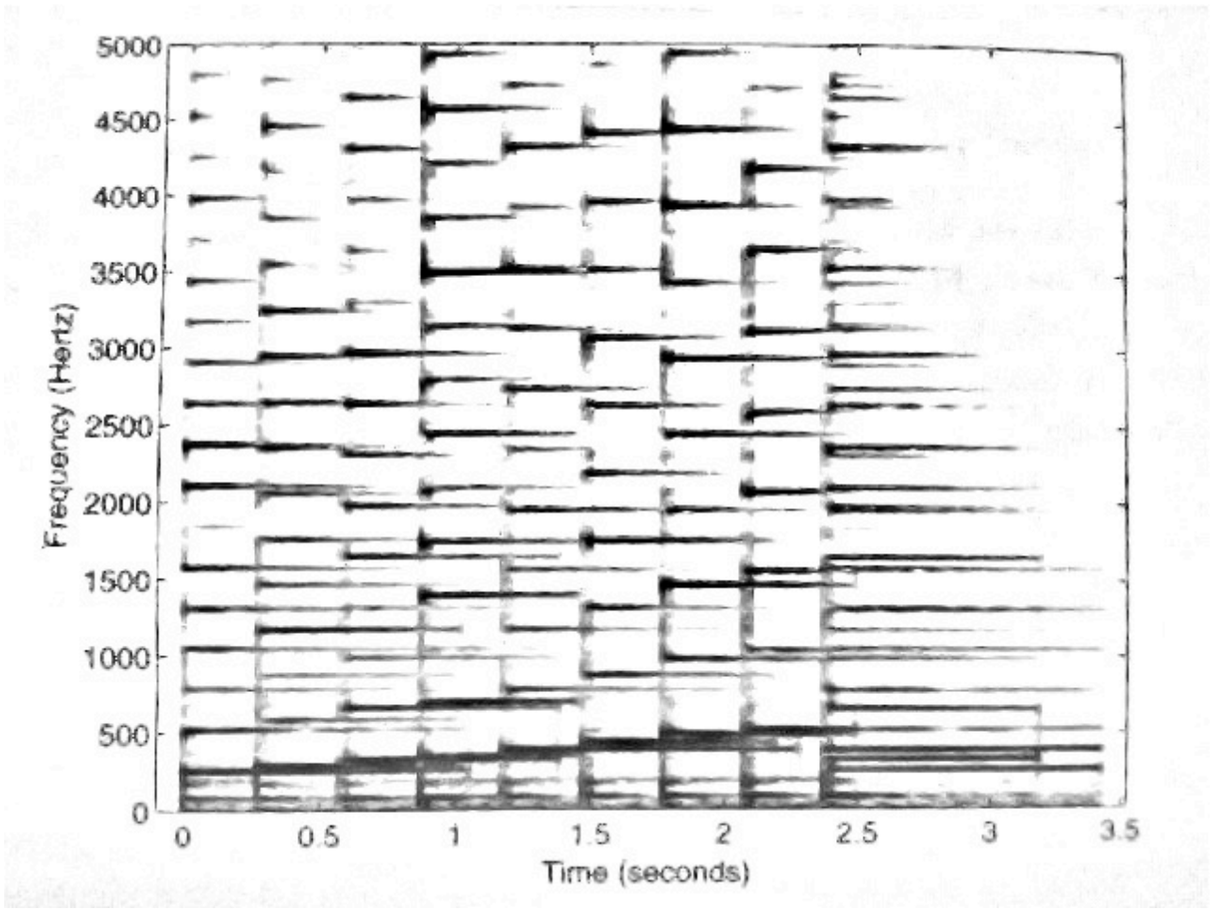
# Principe

modèle X = B G  de réduction de rang

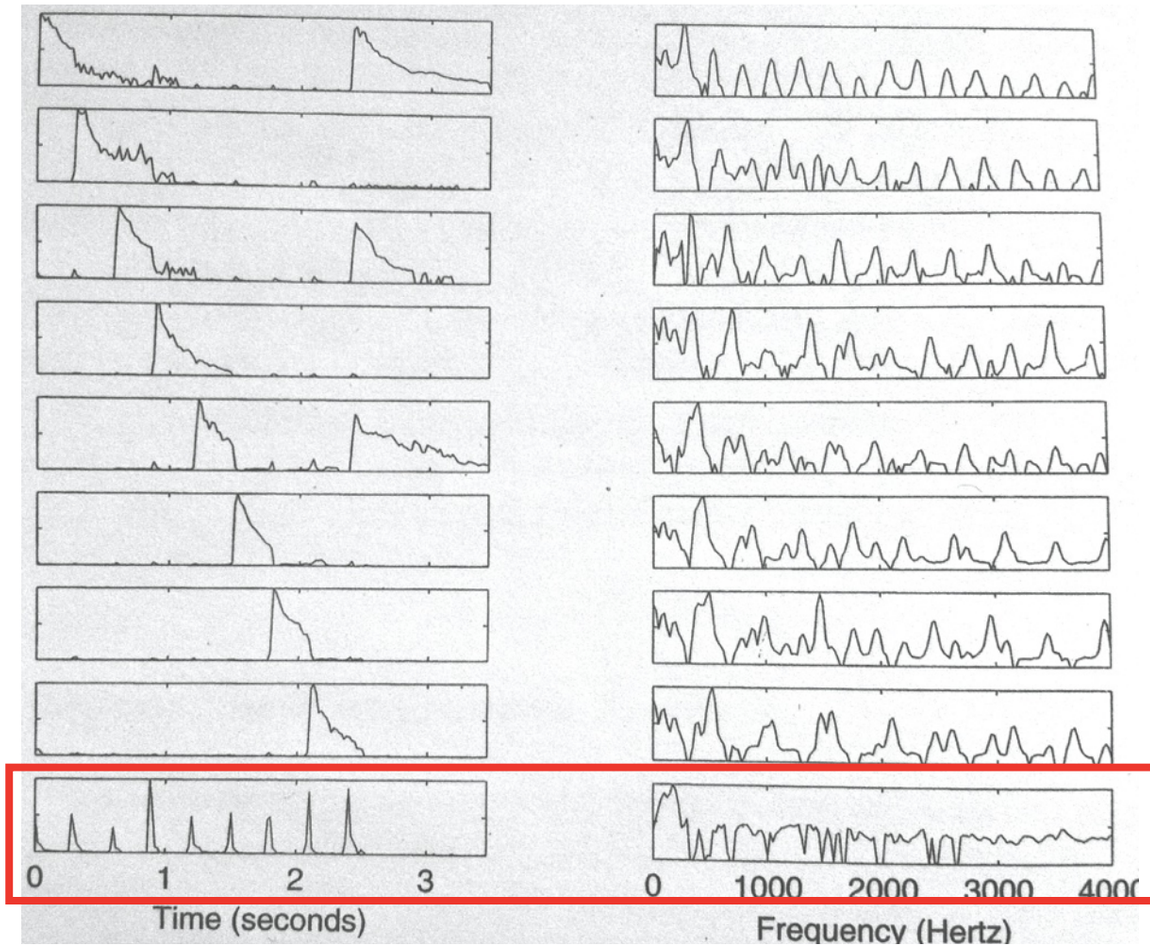
X (observation)

G (gain)
partition

B (mélange)
~notes




=



# Reconstruction des sources

- On suppose la factorisation effectuée X = BG
- Le module de la source (1) est estimé comme
  $X_{(source1)} = B_{(source1)} G_{(source1)}$
- pour reconstruire la source en temporel
  - soit on reconstruit des phases en utilisant la redondance de la STFT (algorithme itératif de Griffin&Lim)
  - soit on utilise une technique type Filtre de Wiener sur le mélange

$$S_i(n,k) = \frac{\sigma_i^2(k)}{\sum_{j=1}^{M} \sigma_j^2(k)} X_1(n,k)$$

utilise la phase du mélange : OK si peu de recouvrement spectral

  - techniques hybrides (travaux LeRoux, N. Sturmel / LD)

# Comment factoriser ?

La variance de chacune des sources peut être estimée par NMF

Article fondateur Lee & Seung

$$d_{euc}(B,G) = \|[X]_{k,t} - [BG]_{k,t}\|_F^2$$

$$d_{div}(B,G) = \sum_{k,t} D([X]_{k,t}, [BG]_{k,t}) \qquad D(p,q) = p\log\frac{p}{q} - p + q$$

(Kullback-Leibler pour des valeurs normalisées)

autres divergences possibles: Itakura-Saito, ou + généralement beta-divergences (Févotte)

$$D_{IS}(p,q) = \frac{p}{q} - 1 + \log\frac{p}{q}$$

(Gamma distrib sur le bruit)

# NMF

règles d'updates multiplicatifs

$d_{euc}$

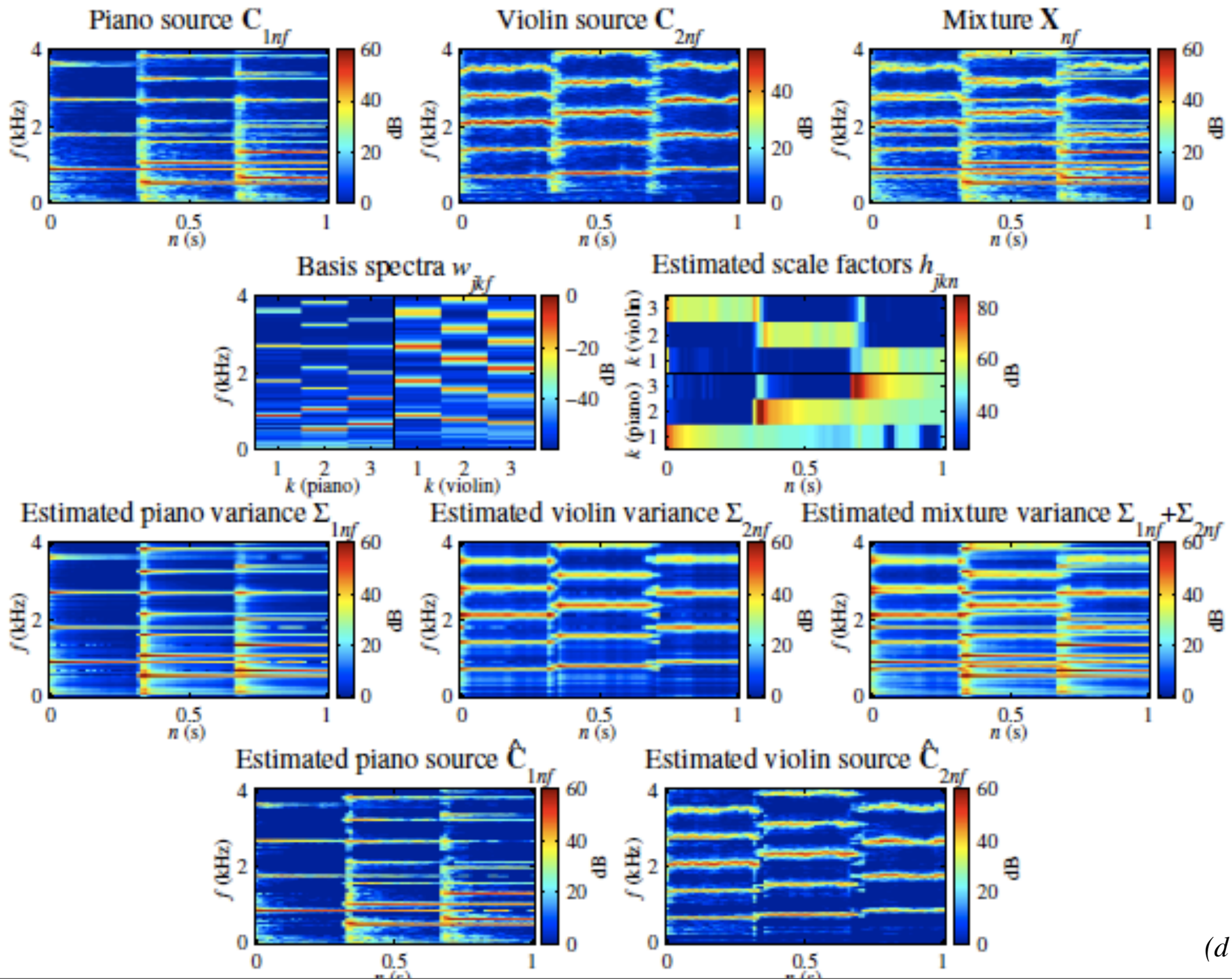
$$B \leftarrow B. \times (XG^T)./(BGG^T)$$
$$G \leftarrow G. \times (B^T X)./(B^T BG)$$

$d_{KL}$

$$B \leftarrow B. \times ((X./BG)G^T)./(1G^T)$$
$$G \leftarrow G. \times (B^T(X./BG))./(B^T 1)$$

Bon résultats perceptifs avec KL div (invariant par chgt d'échelle et donc plus sensible aux petites valeurs)

# NMF




*(d'après Vincent / Ono)*

# Approches “semi-aveugles”

# Addition d'information supplémentaire sur les sources

- Dans la plupart des cas "réels" les hypothèses habituelles (indépendance, non-négativité, parcimonie ...) sont insuffisantes pour obtenir des résultats de qualité suffisante pour les applications citées.
- Risque de sur-apprentissage
- De plus nous avons souvent de l'information a priori sur les sources
- Ex: analyse d'instruments à sons entretenu : sources harmoniques, spectre ≈ stationnaire
- Plus souvent, cette information est spécifique à une source et peut être obtenue par apprentissage.
- Parfois l'info a priori est juste utilisée comme initialisation.

# Modèles de sources dans le cas X = BG

En général l'estimation est faite en 2 étapes :

- Apprentissage de $B_{sources}$ à partir de matériau d'apprentissage (sources isolées). Parfois les caractéristiques des gains peuvent être stockées, par exemple en vue d'une modélisation de leur distribution.

- Représentation du signal polyphonique comme combinaison linéaire des $B_{sources}$ appris, uniquement estimation des gains

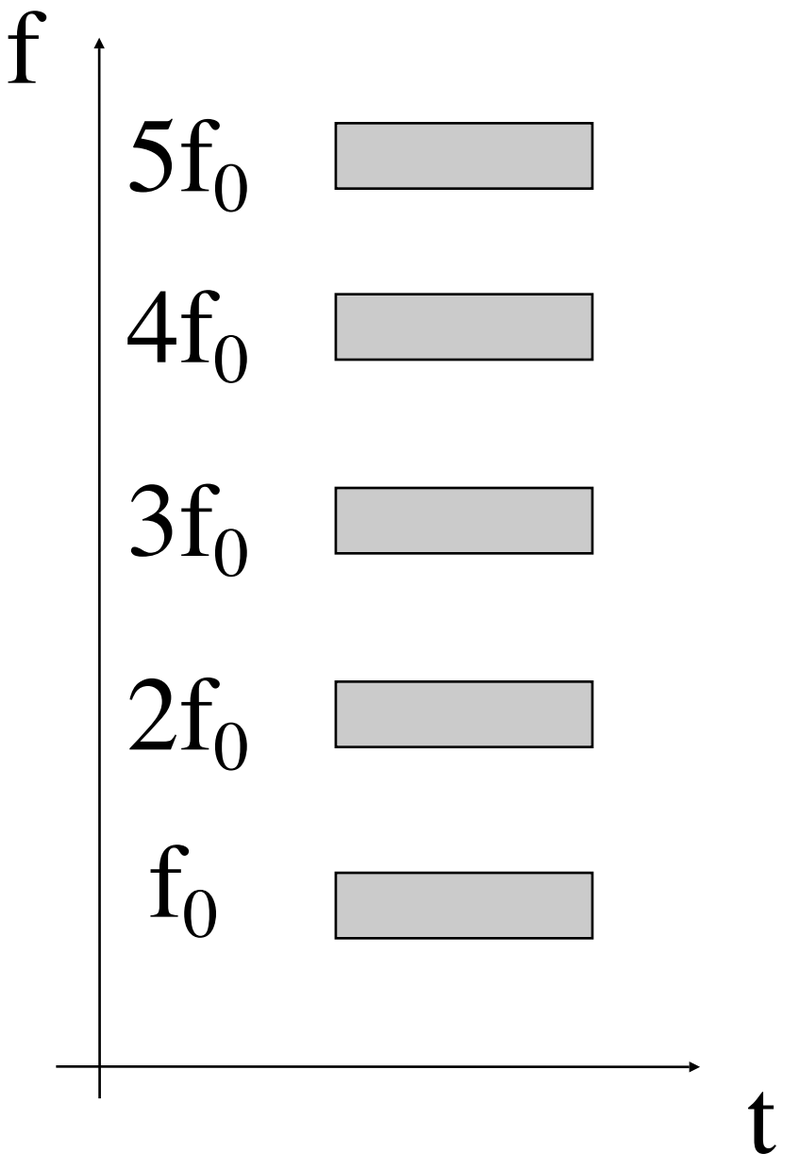
# Exemple dans le cas linéaire parcimonieux

Représentation parcimonieuse par atomes harmoniques appris.

P. Leveau, E. Vincent, G. Richard and LD, Instrument-specific harmonic atoms for mid-level music representation, IEEE Trans Audio Speech and Language Processing,16(1), 2008

# Harmonic Gabor atoms

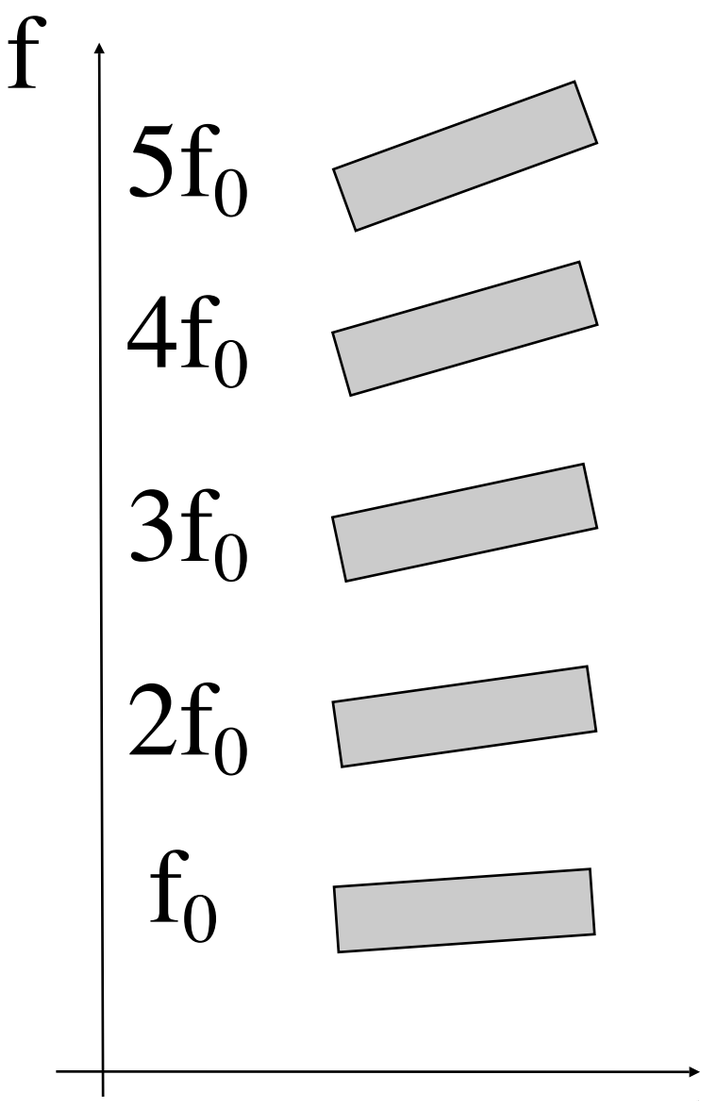
[Gribonval TSP 2003]

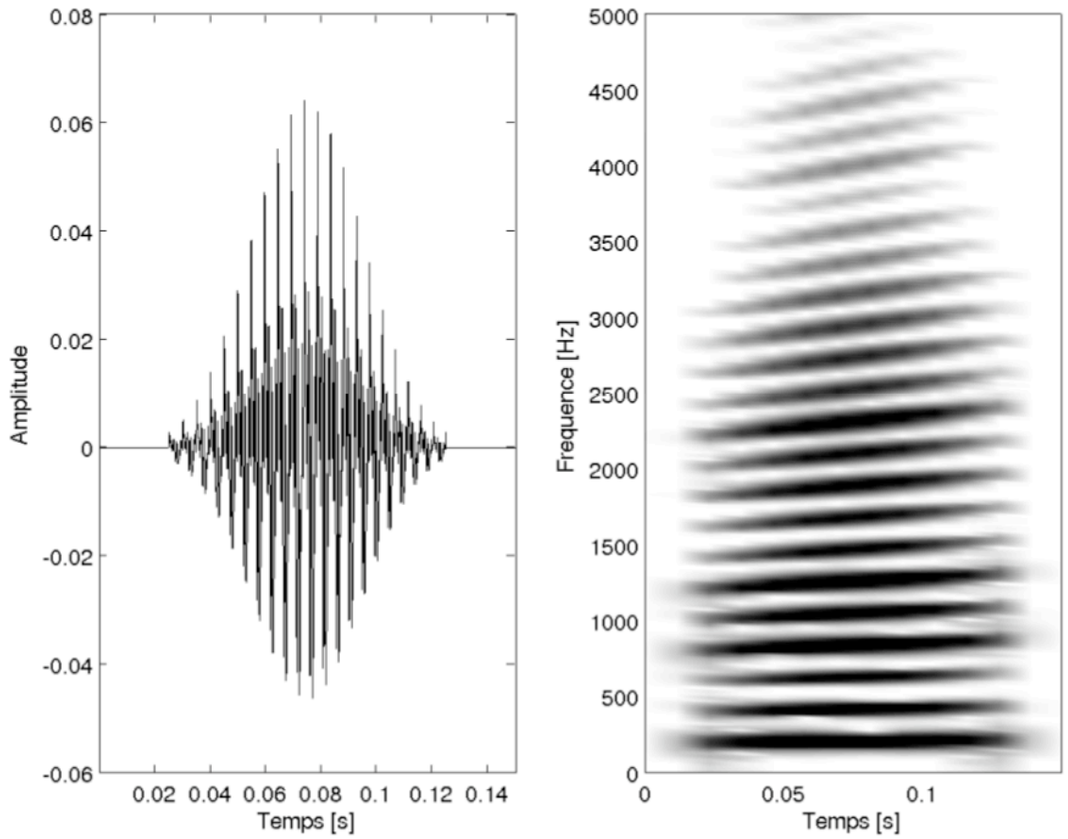



$$g_{s,u,f} = w(\frac{t-u}{s})e^{2\pi jf.t}$$

$$h_{s,u,f_0,A,\Phi} = \sum_{m=1}^{M} a_m e^{j\phi_m} g_{s,u,m.f_0}$$

# Atomes de Gabor harmoniques avec facteur de “chirp”



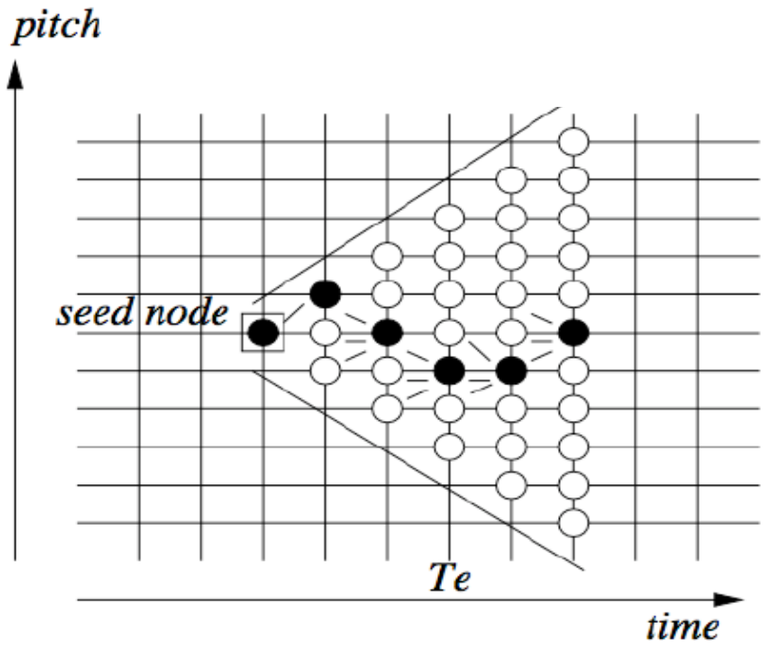



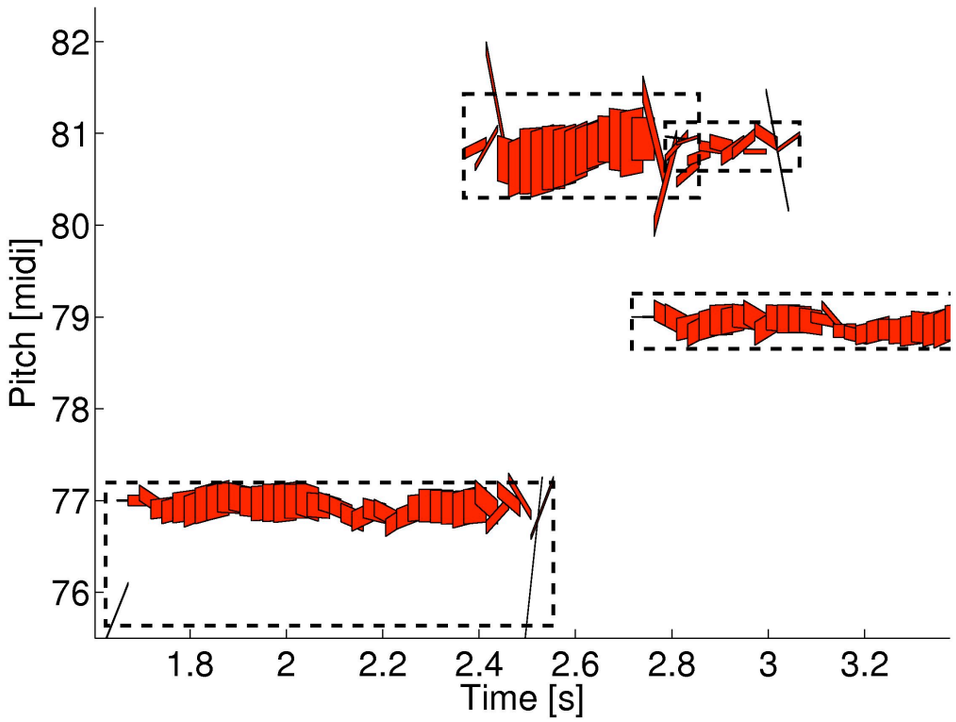

Amplitudes relatives des partiels apprises sur des notes isolées.

clarinette violoncelle

# « molécule »: chaîne d'atomes

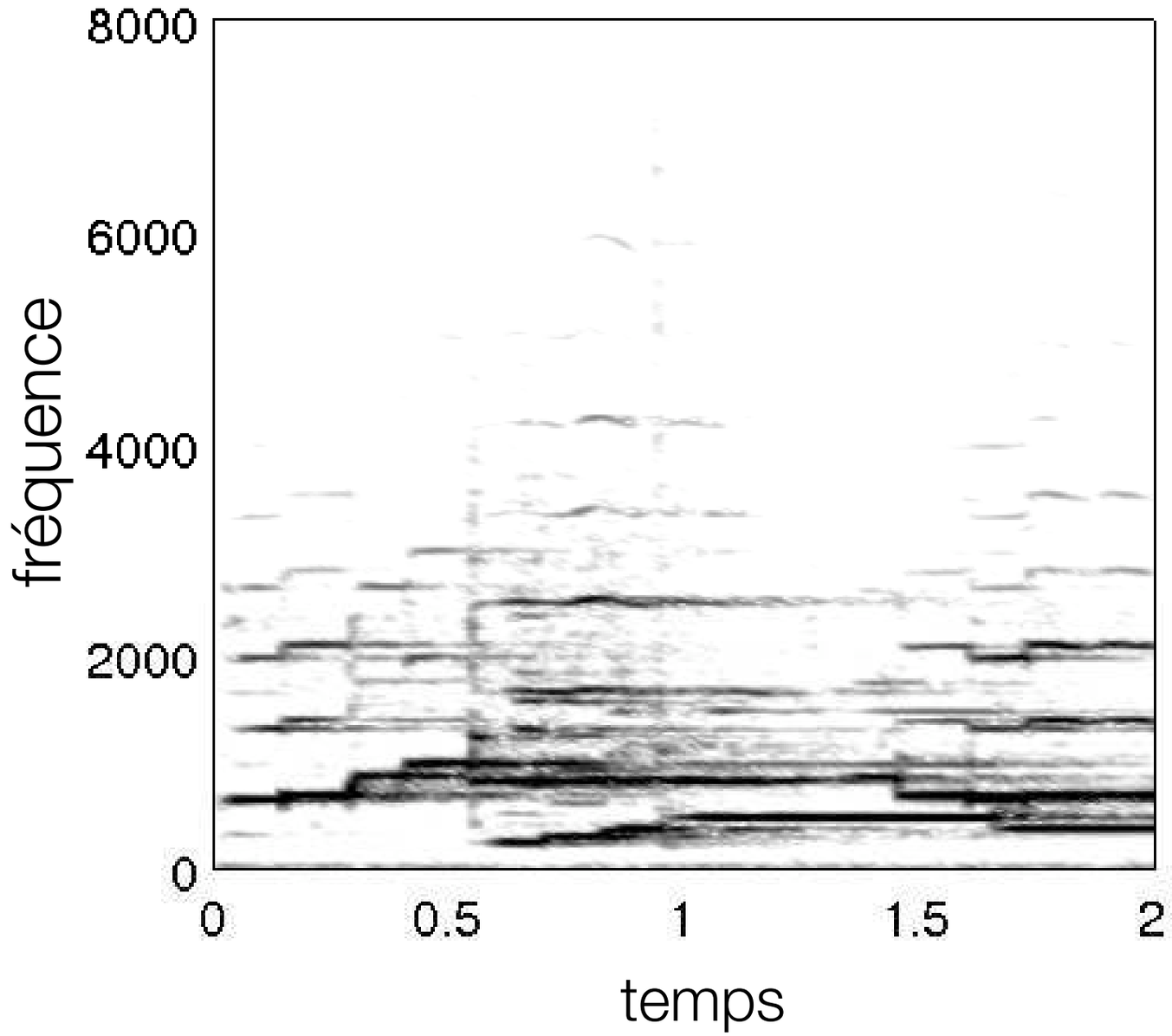
- Dans une représentation temps-fréquence, les atomes d'énergie importante se répartissent le long de chaînes.
- Chaîne d'atomes = molécule
- "Matching Pursuit Moléculaire"



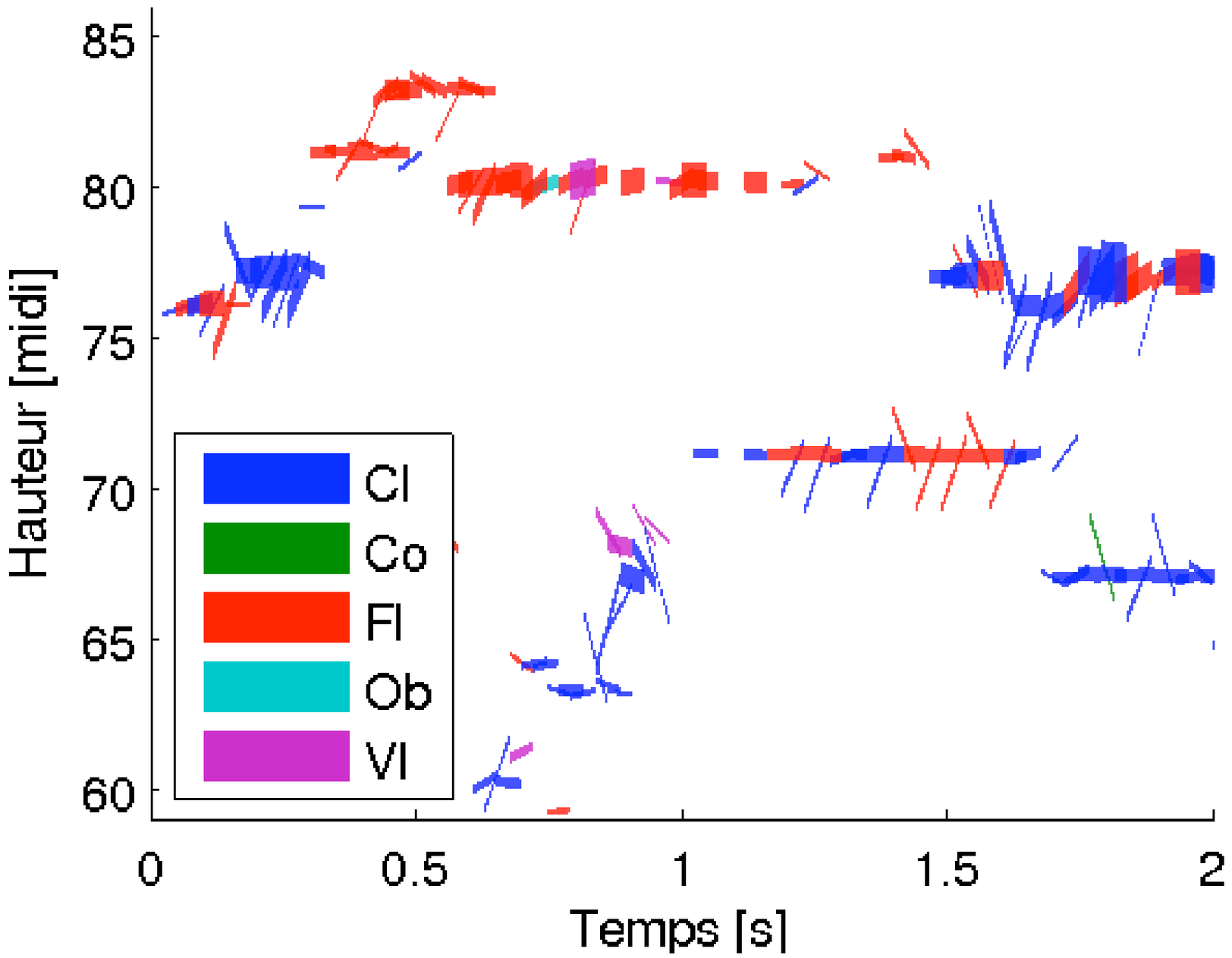
# Visualisation

## Spectrogramme duo Flûte-Clarinette



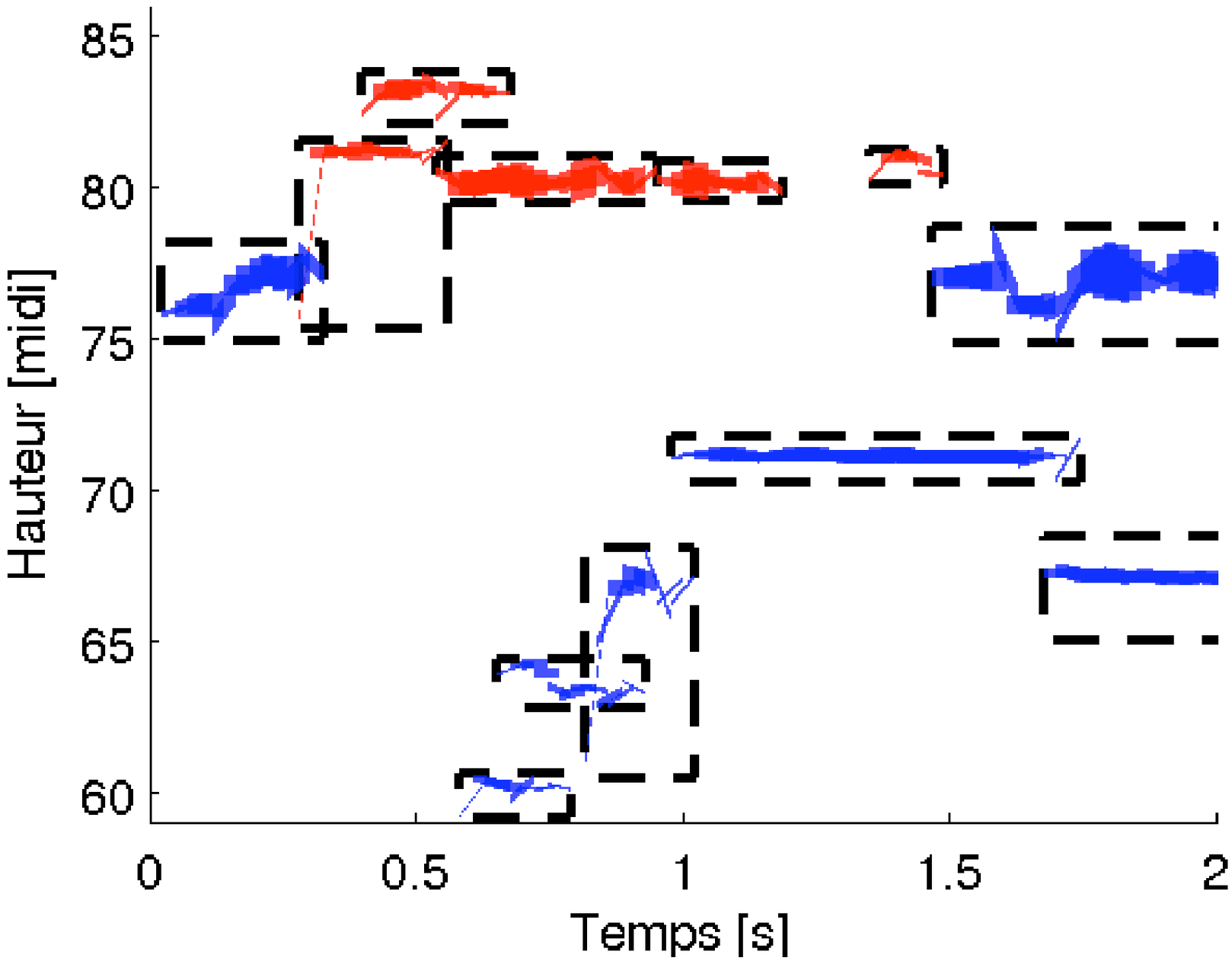
# Visualisation

## Représentation atomique

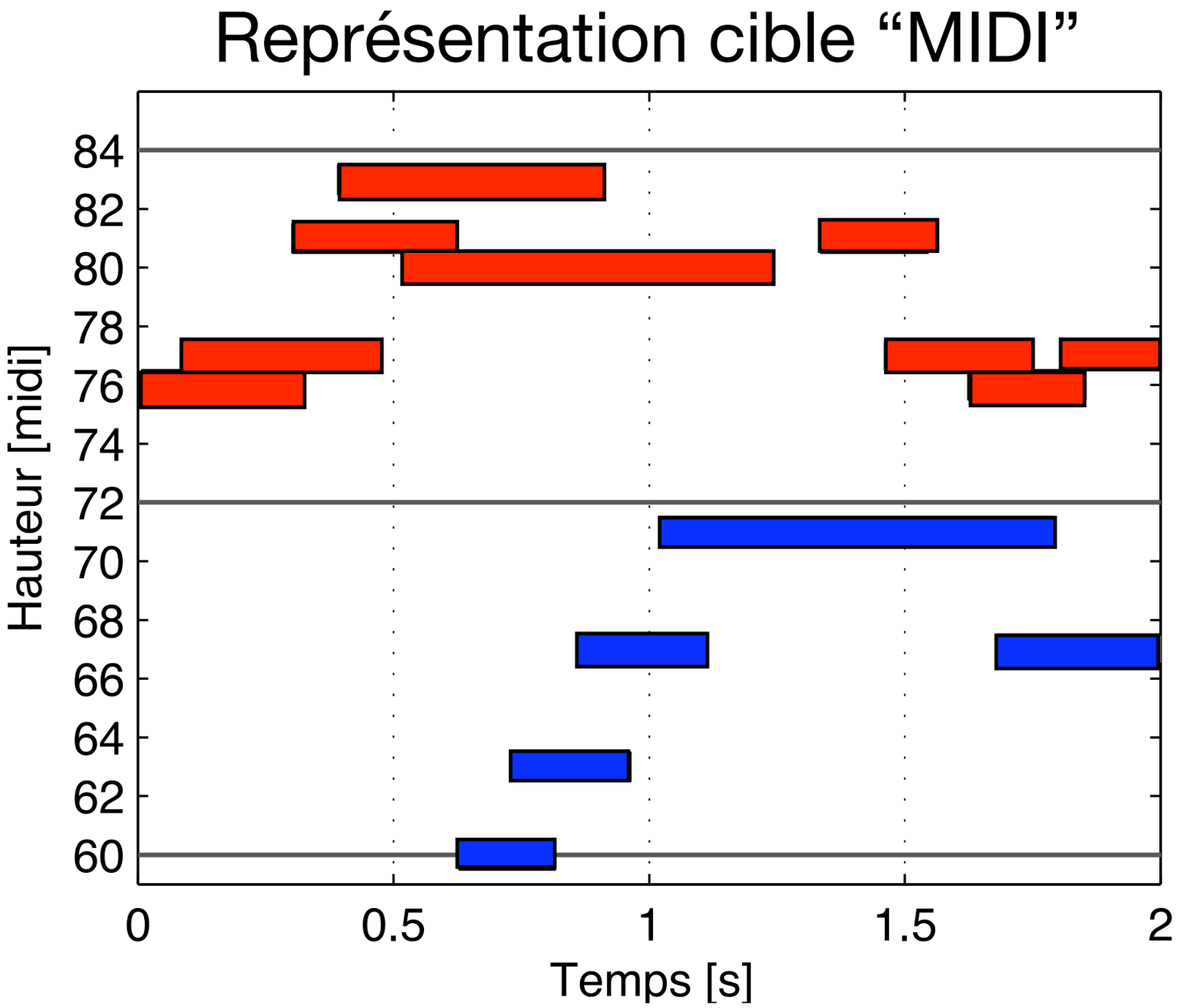


# Visualisation

## Représentation moléculaire



# Visualisation



# Application à la reconnaissance automatique des instruments

**Reconnaissance d'instrument solo**
*(parmi 5 candidats)*

| | |
|---|---|
| décomposition atomique | 76 % |
| décomposition moléculaire | 71% |
| référence (SVM / MFCC) | 78 % |

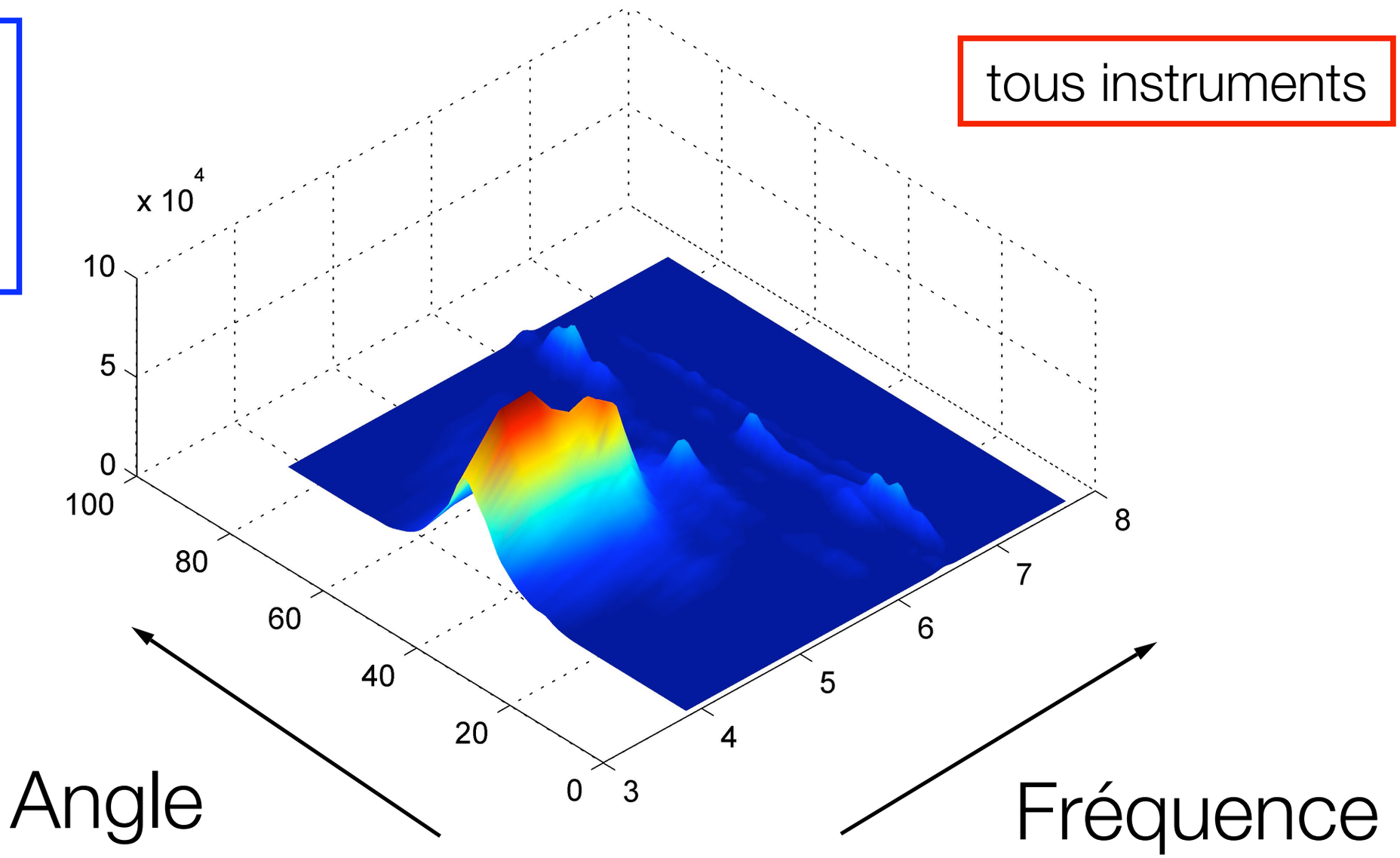
**Reconnaissance de duo**
*(parmi 25+5 candidats)*

| | |
|---|---|
| décomposition atomique | 35 % |
| décomposition moléculaire | 41% |
| référence (SVM / MFCC) | n/a |

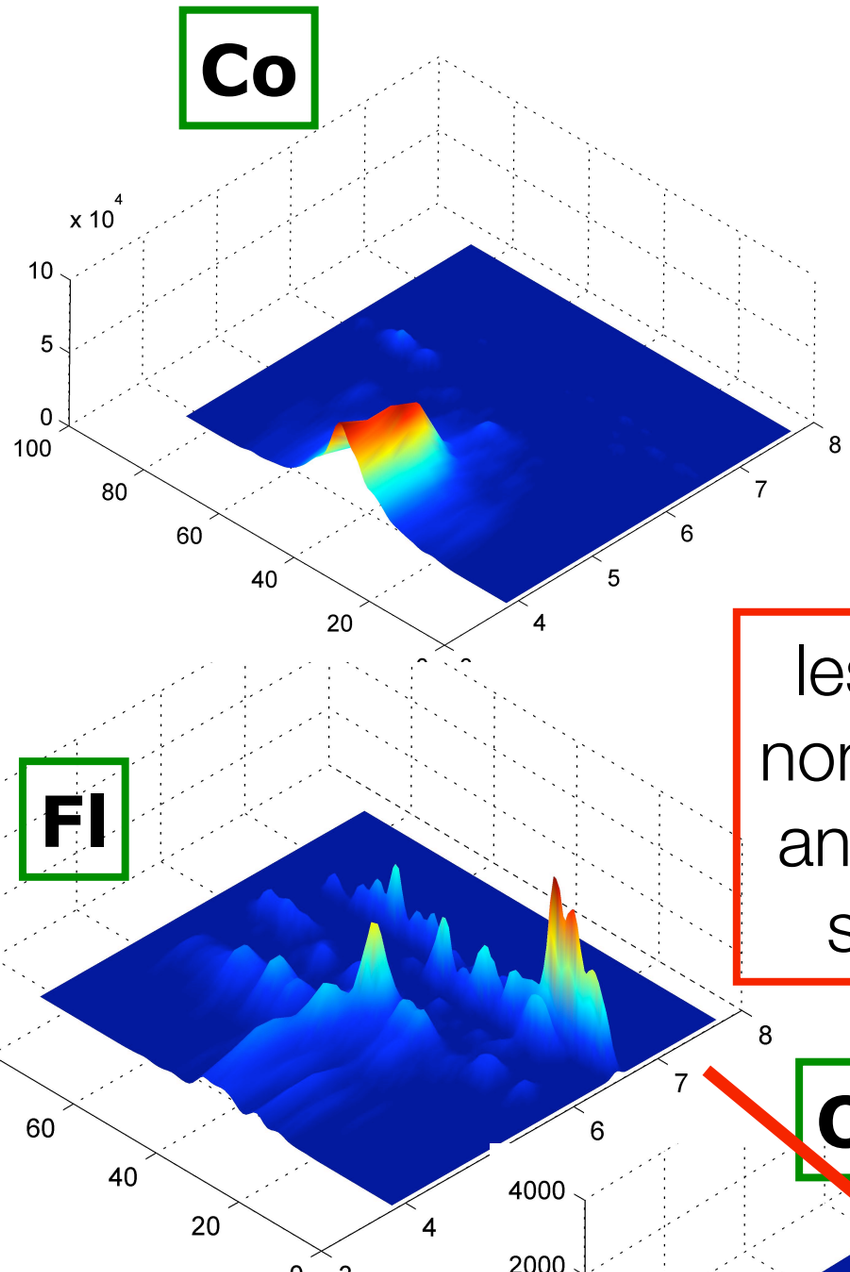
# Prise en compte de la stéréo

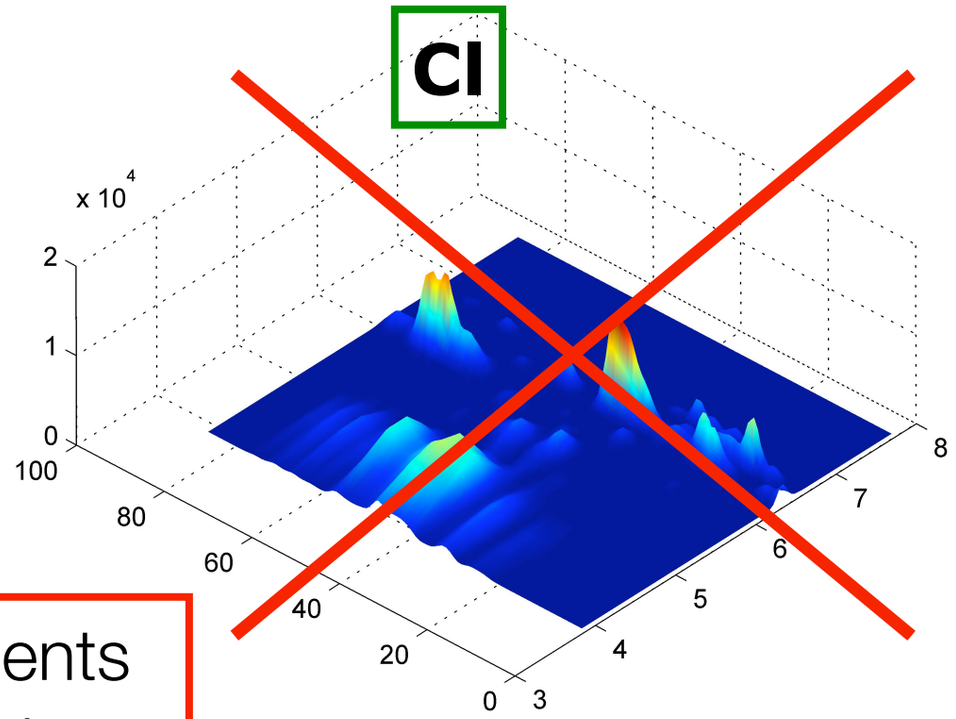
diagramme fréquence-angle

tous instruments



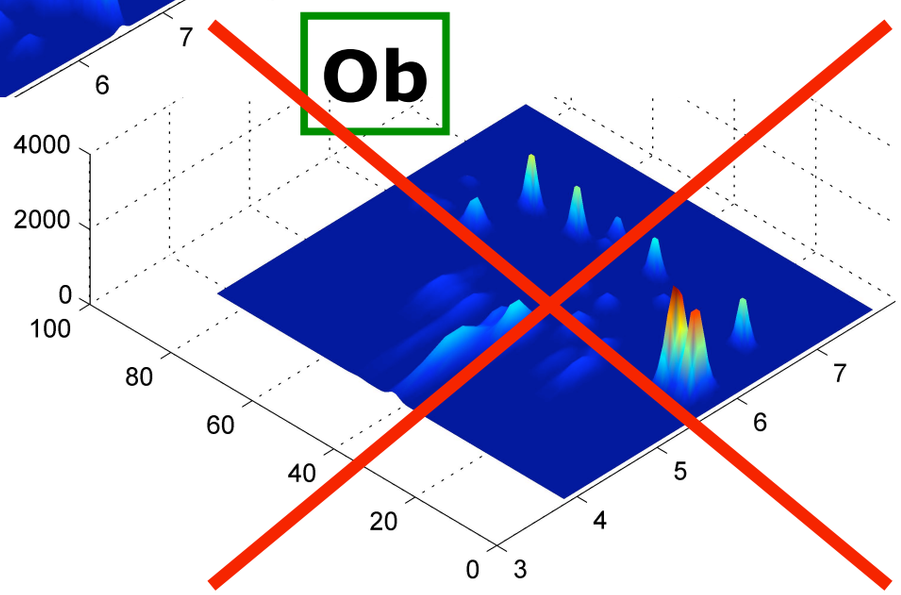

D. Sodoyer, P. Leveau, and LD, Using stereo information for instrument identification in polyphonic mixtures, Proc. WASPAA, 2007

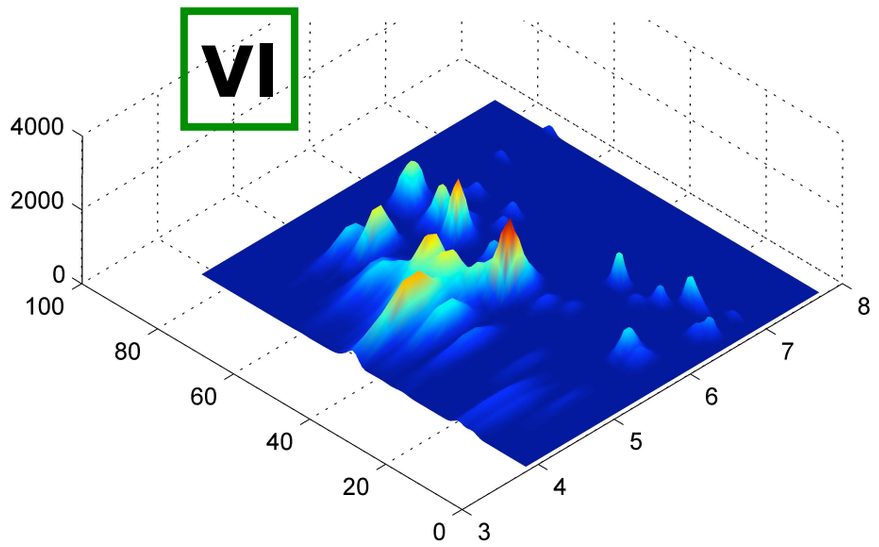



les instruments non cohérents en angle-fréquence sont éliminés

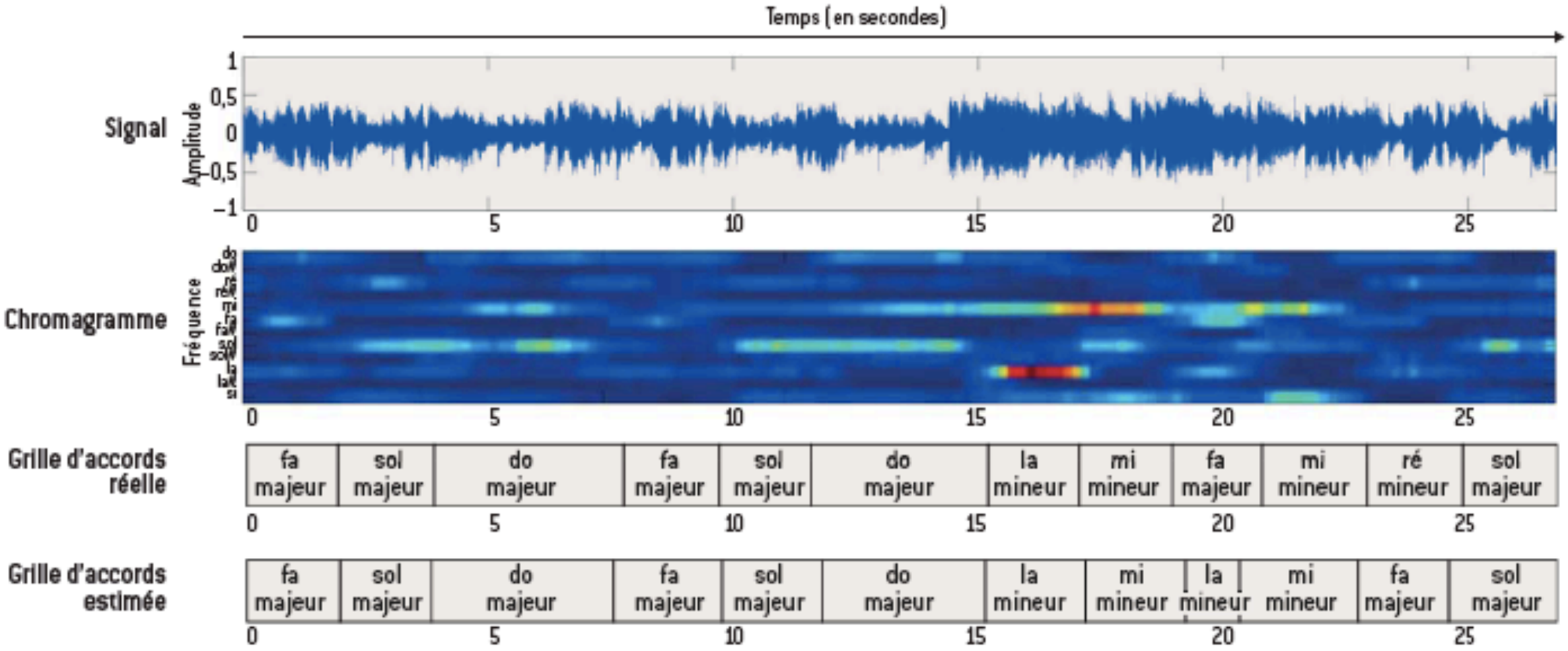



D. Sodoyer, P. Leveau, and LD, Using stereo information for instrument identification in polyphonic mixtures, Proc. WASPAA, 2007

# Example : chord recognition



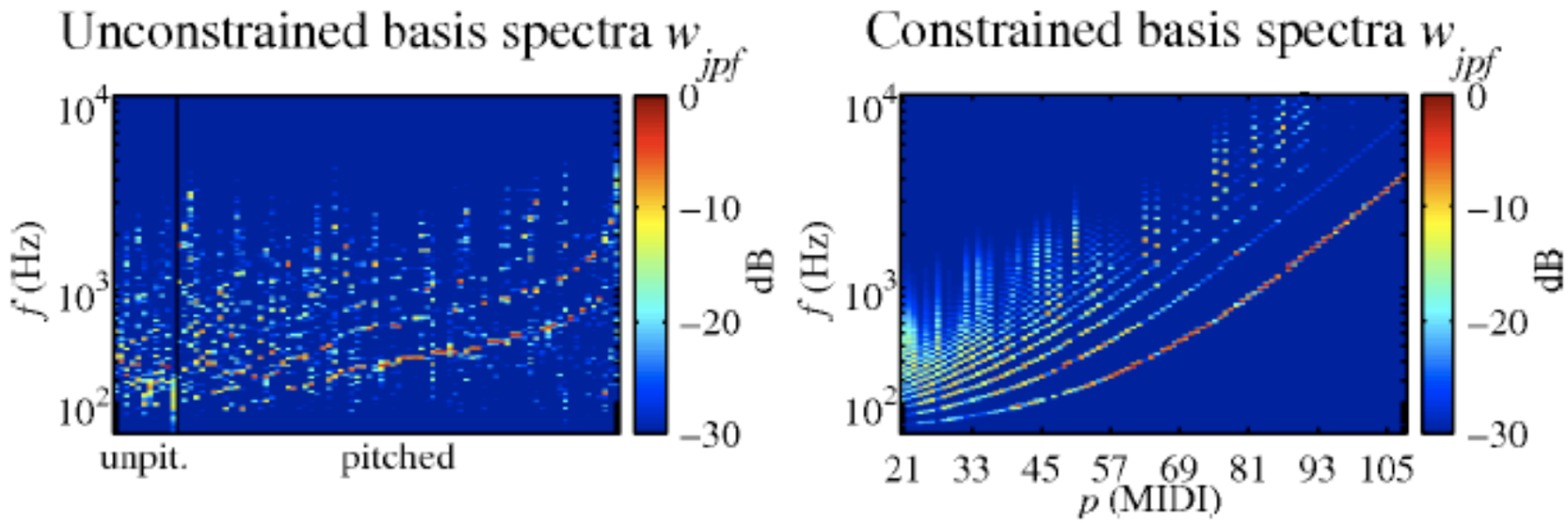

E. Ravelli, G. Richard, LD, *IEEE TSALP 2010*.

LD, *Représenter les sons musicaux*, Pour La Science, Nov 2008.

# Contraintes sur la base B

Abdallah and Plumbley : initialisation des bases avec des infos a priori, puis relaxation de la contrainte

Contraintes d’harmonicité dans la base (E. Vincent)




*(d’après E. Vincent et N. Ono)*

# Estimation de G si B est supposé connu

- Sans utiliser l'hypothèse de parcimonie au sens des moindres carrés :

$$G_{est} = ( B^T B )^{-1} B^T x$$

- Moindres carrés avec contrainte de positivité

- On peut aussi utiliser un modèle temporel pour les durées de notes (par ex. HMM comme dans la thèse d'E. Vincent).

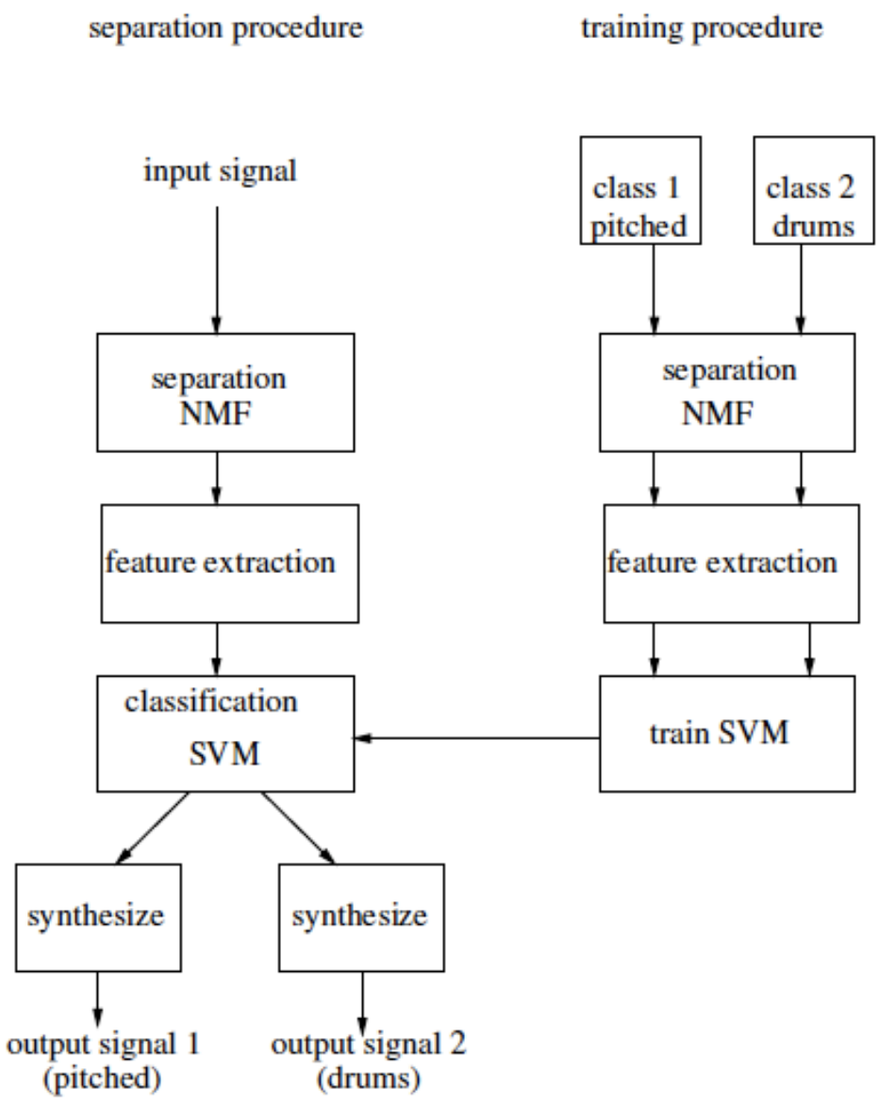
# Utilisation des composants extraits

Si on estime B à partir du mélange, il faut associer chaque composant à une source.

Cas non supervisé : clustering
Cas supervisé : si on a un modèle de source, les composants peuvent être associés en fonction d’une distance au modèle.

# Utilisation des composants extraits

## Ex. étude de Helén/Virtanen sur la séparation de percussions




Original

Drums

sons tonaux (modèle)

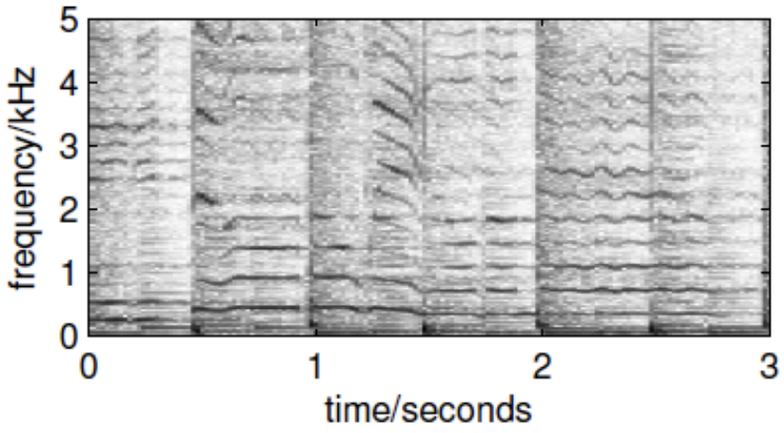
sons tonaux (orig-drums)

# Séparation de la voix / karaoke

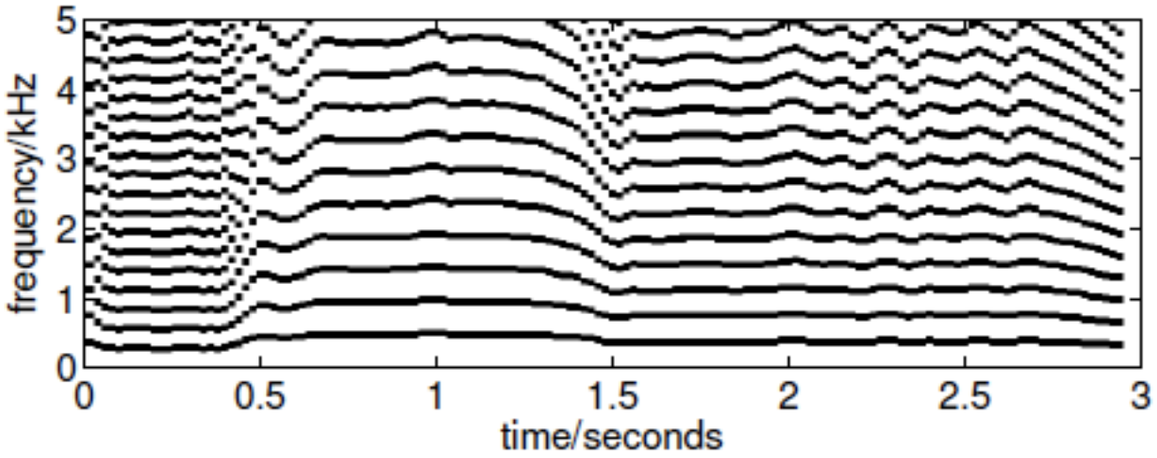
Détection de pitch dominant
templates harmoniques séparant la voix
NMF sur résidu pour mieux connaitre
résidu là où voix.

Etudes de Virtanen, Mesaros & Rynnänen (SAPA2008) et Ryynänen, Virtanen, Paulus, and Klapuri (ICME'08)

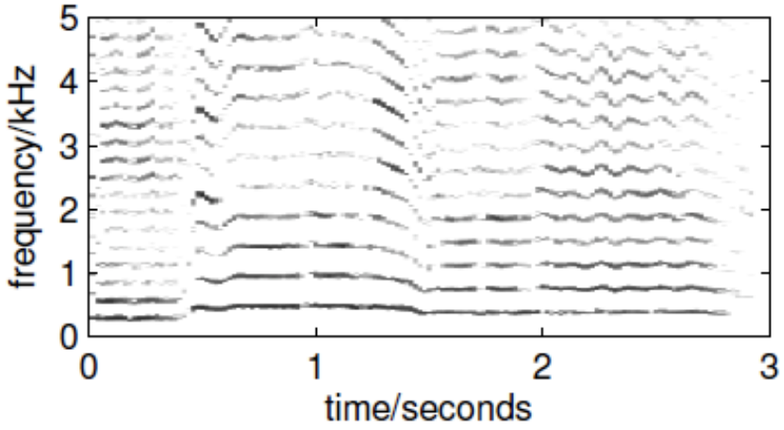
# Séparation de la voix



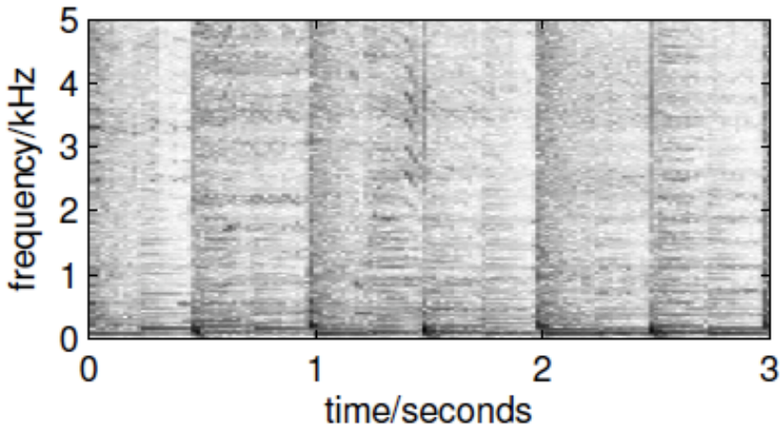

mix




masque pour la voix




voix




autre

# Séparation de la voix

## Etude de Virtanen, Mesaros & Rynnänen

| Mix | Voix (prop. meth) | Voix (modèle sinusoidal) | Voix (masquage binaire) |
|---|---|---|---|
| S1 | V1 | V’1 | V”1 |
| S2 | V2 | V’2 | V”2 |
| S3 | V3 | V’3 | V”3 |

# Karaoke (ICME'08)

| | Mix | Karaoke |
|---|---|---|
| Jamiroquai | S1 | V1 |
| Roxette | S2 | V2 |

# Karaoke (ICME'08)

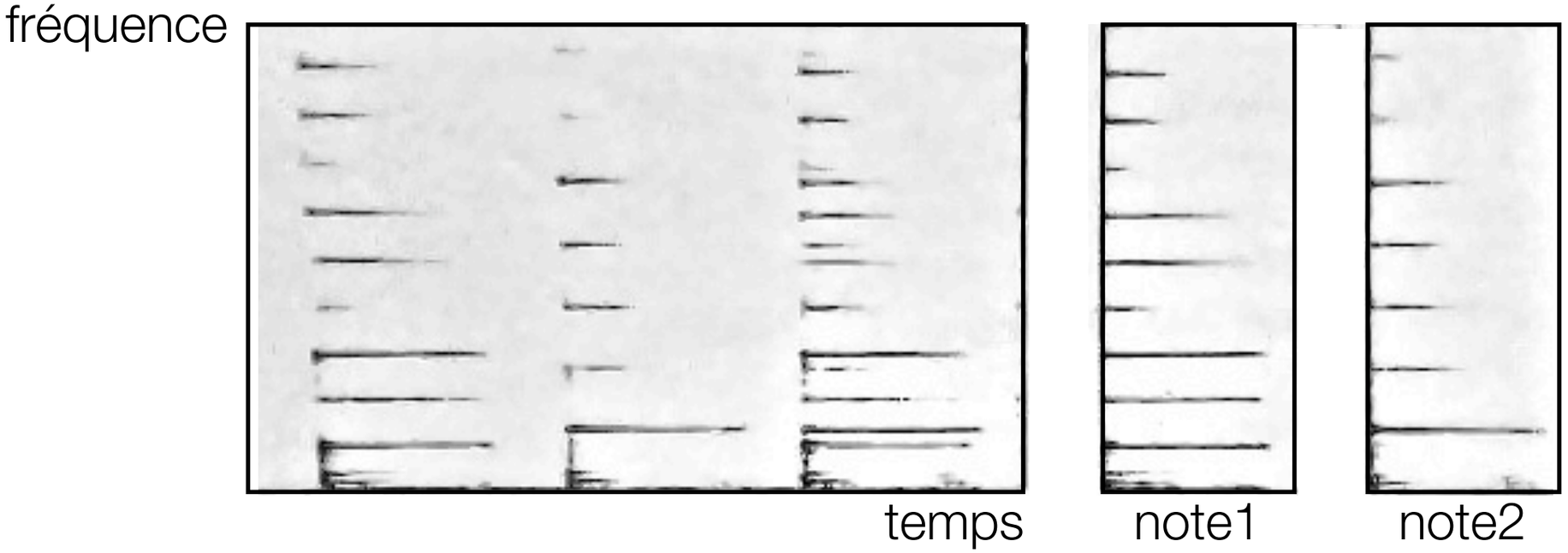
Ré-accorde la voix du chanteur
en temps réel

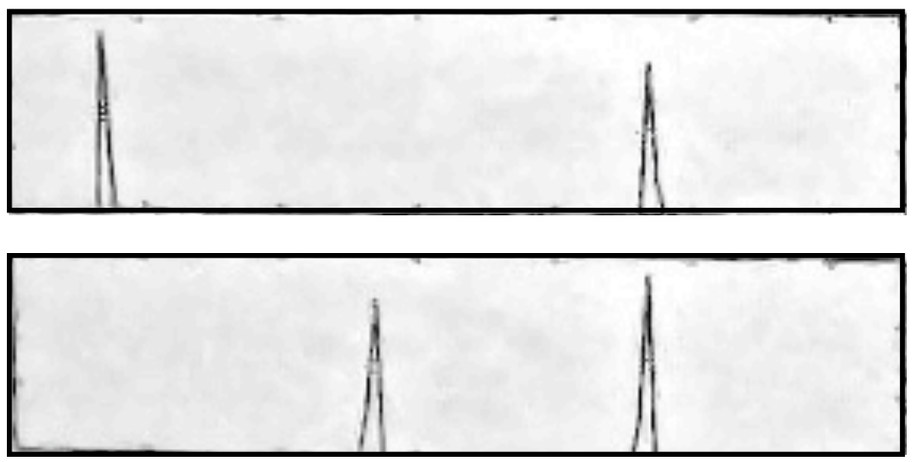
(Demo video sur le site web de l'auteur)

# Composants variant dans le temps

## Cas 1- Spectres variant dans le temps

### travaux de Smaragdis



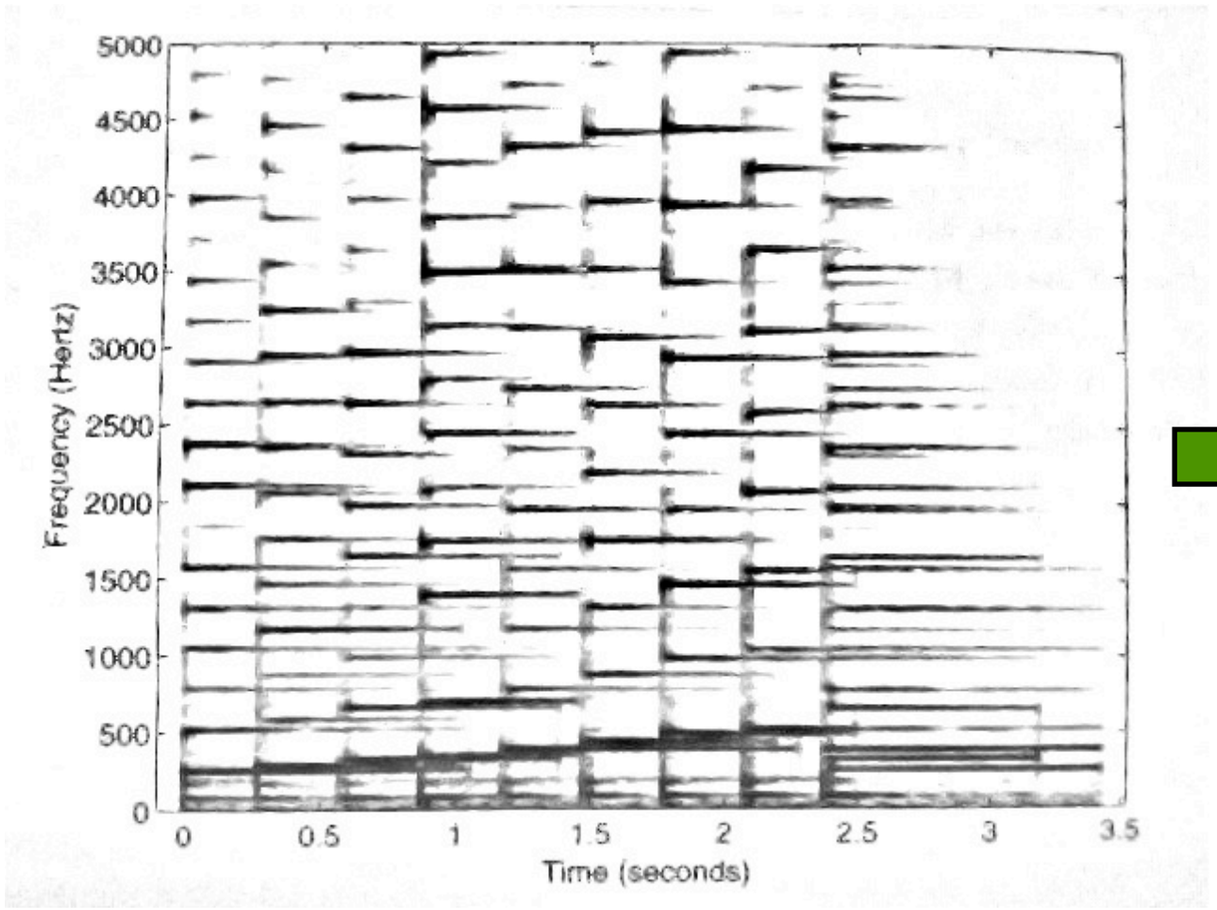




$$\mathtt{x}_t \approx \sum_{n=1}^{N} \sum_{\tau=0}^{L-1} \mathtt{b}_{n,\tau} g_{n,t-\tau}$$

# Composants variant dans le temps

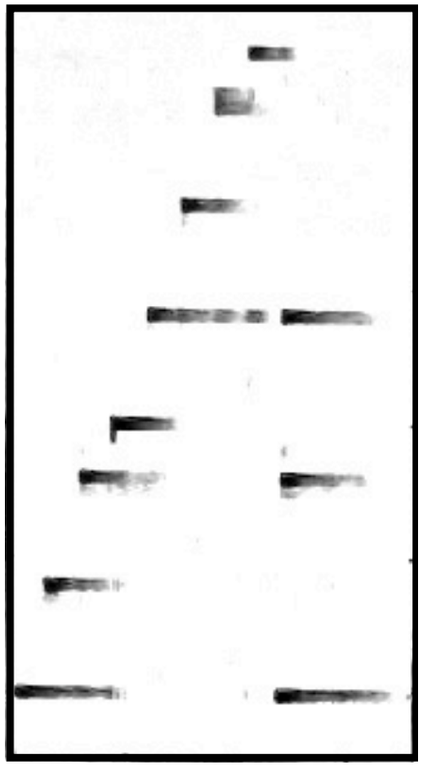
## Cas 2 - Fréquences variant dans le temps

$$\mathbf{x}_t \approx \sum_{n=1}^{N} \sum_{z=0}^{Z} b_{n,k-z} g_{n,t,z}$$

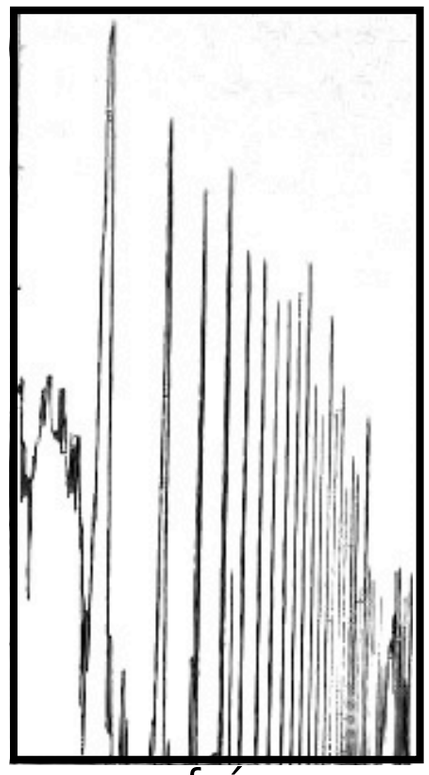



z décalage fréquentiel **(échelle log)**

amplitude



temps



fréquence **(échelle log)**

# Composants variant dans le temps

## Cas 2 - Fréquences variant dans le temps

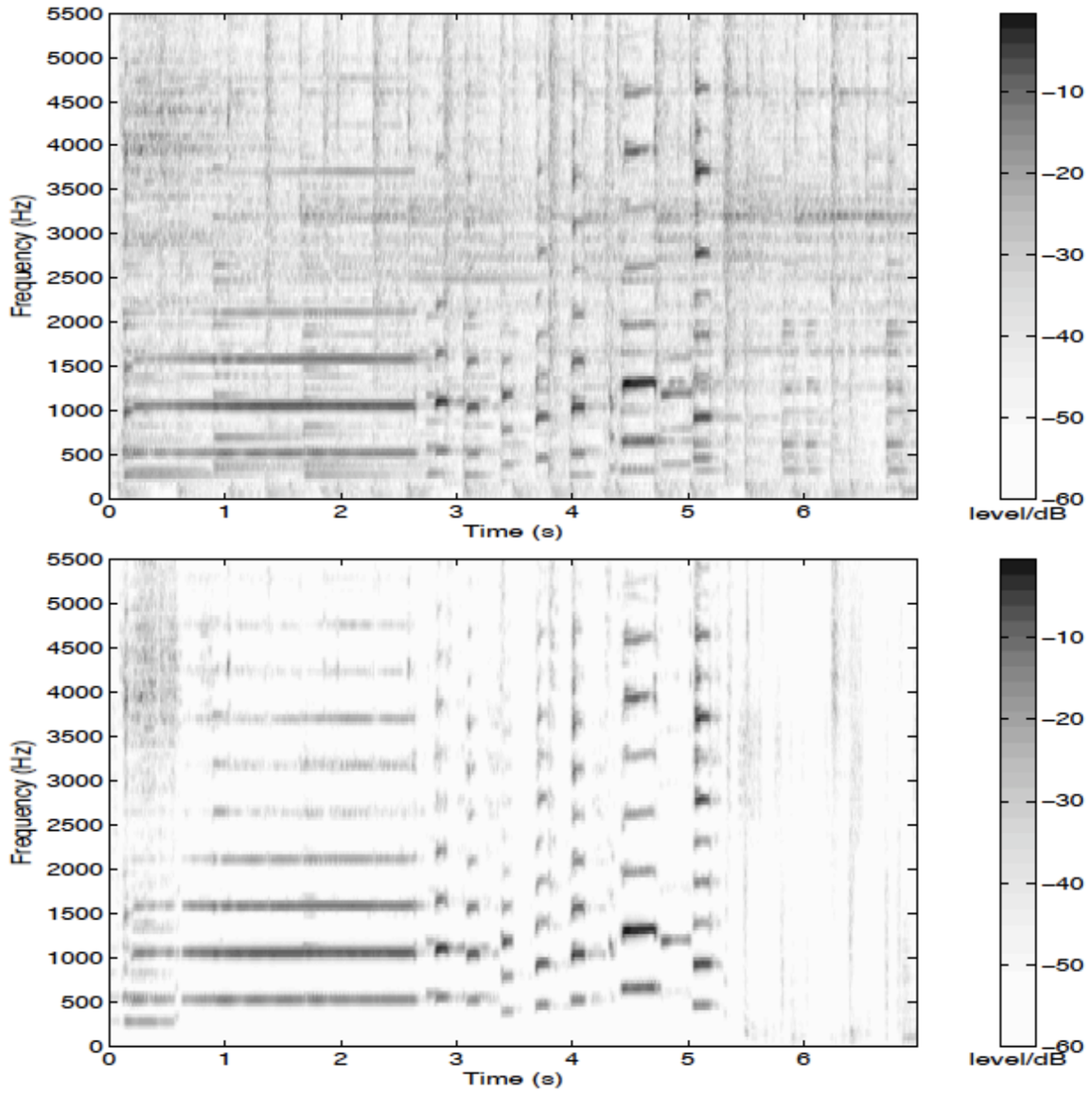
Une variante:

- application : Extraction d'instrument dominant
- hypothèse : modèle source-filtre sur l'instrument dominant dont le filtre est invariant (sans apprentissage préalable).

J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

# Composants variant dans le temps

## Cas 2 - Fréquences variant dans le temps




Mix

Trompette

autre

J.-L. Durrieu, G. Richard, B. David, and C. Févotte, “Source/filter model for unsupervised main melody extraction from polyphonic audio signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

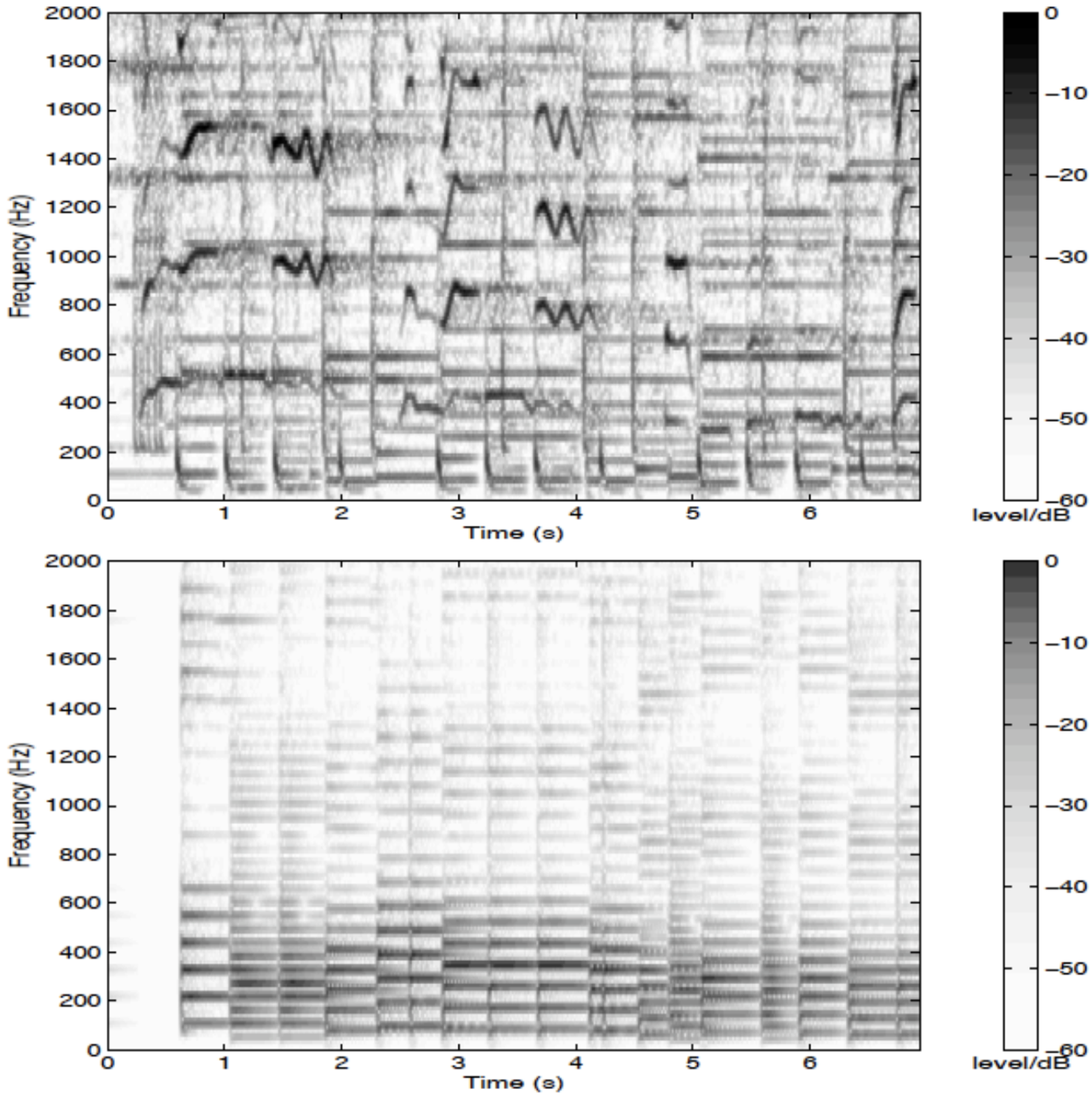
# Avec modèle “musicologique”

M. Ryynänen and A. Klapuri, “Automatic bass line transcription from streaming polyphonic audio,” *ICASSP*, 2007.

## Modèle hiérarchique

- 2 états : note ou repos
  - Modèle de note par HMM a 3 états ASR (Attack Sustain Release)
- Modèle musicologique des transitions entre note et repos dépendant de l’accord et des notes précédentes
- Apprentissage sur données symboliques (MIDI)
- Décodage par Viterbi

# Avec modèle “musicologique”




mix

basse

mix (g) avec basse (d)

M. Ryynänen and A. Klapuri, “Automatic bass line transcription from streaming polyphonic audio,” *ICASSP*, 2007.

# Séparation de sources informée

# Séparation de sources informée

- Pour une séparation de grande qualité, il est pafois nécessaire d'avoir un a priori plus précis que juste des *modèles* de sources

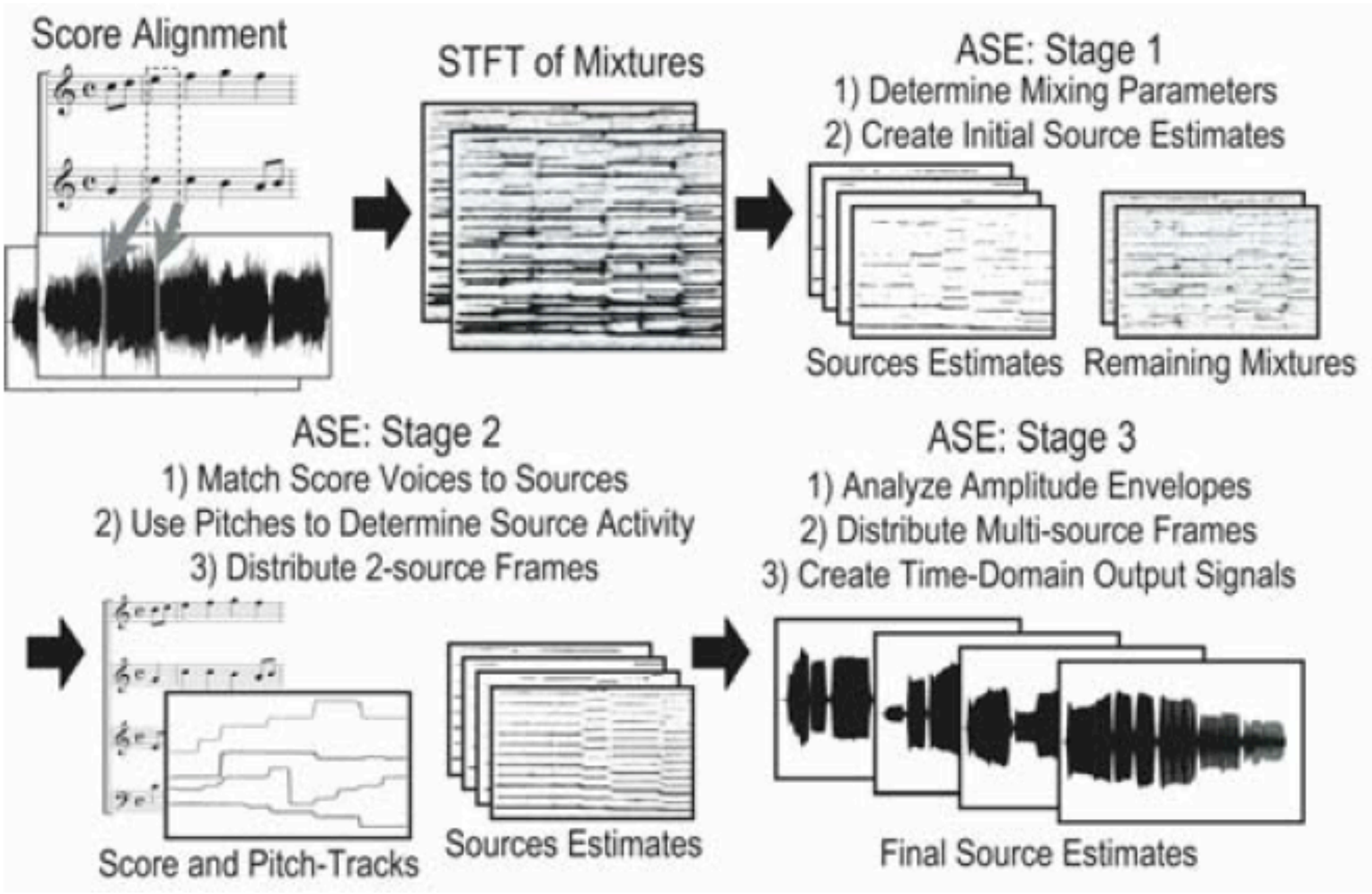
- Séparation avec l'aide d'informations sur la fonction d'activation (les gains)

    - Séparation avec la partition (Woodruff, Pardo, Dannenberg)

    - Séparation guidée par l'utilisateur (Mysore / Smaragdis)

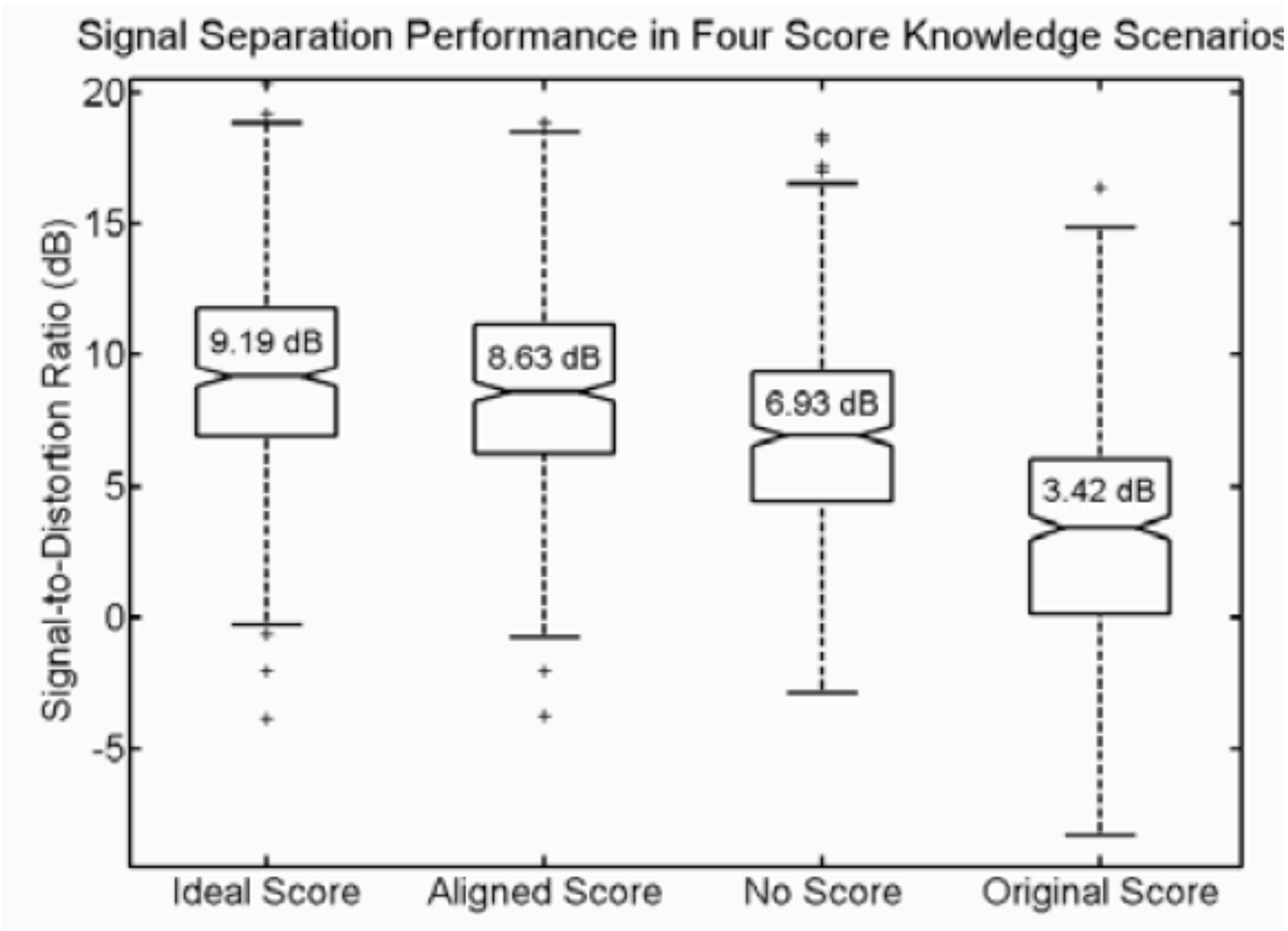
# Séparation de sources informée par la partition

Woodruff, Pardo, Dannenberg, *ISMIR 2006*



# Séparation de sources informée par la partition

Woodruff, Pardo, Dannenberg, *ISMIR 2006*



# Séparation de sources guidée par l'utilisateur

G. Mysore et P. Smaragdis, *Waspaa 2009, LVA/ICA 2010*

---

# Modèle PLCA (Probabilistic Latent Component Analysis)

Amplitude du spectrogramme F vue comme histogramme mesurant l'energie en (t,f), approximé comme somme pondérée de produits de distributions marginales en t et f

$$\mathbf{F} \approx \gamma \sum_{z=1}^{M} P(z)P(f|z)P(t|z)$$

modèle discuté par P. Comon ce matin !

M=1 $P(f|z)$ spectre $P(t|z)$ enveloppe temporelle

On suppose ces distributions multinomiales

# Séparation de sources guidée par l'utilisateur

G. Mysore et P. Smaragdis, *Waspaa 2009, LVA/ICA 2010*

## Estimation par EM

$$\text{E-step: } P(z|f,t) = \frac{P(z)P(f|z)P(t|z)}{\sum_{z'} P(z')P(f|z')P(t|z')}$$

$$\text{M-step: } P(f|z) = \frac{\sum_t \mathbf{F}_{f,t} P(z|f,t)}{\sum_{f'} \sum_t \mathbf{F}_{f',t} P(z|f',t)}$$

$$P(t|z) = \frac{\sum_f \mathbf{F}_{f,t} P(z|f,t)}{\sum_f \sum_{t'} \mathbf{F}_{f,t'} P(z|f,t')}$$

$$P(z) = \frac{\sum_f \sum_t \mathbf{F}_{f,t} P(z|f,t)}{\sum_{z'} \sum_f \sum_t \mathbf{F}_{f,t} P(z'|f,t)}$$

Prise en compte d'a priori sur $P(f|z)$ et $P(t|z)$

On suppose que l'on dispose de "modèles" $\alpha(f|z)$ et $\alpha(t|z)$

$$P(f|z) = \frac{\sum_t \mathbf{F}_{f,t} P(z|f,t) + \kappa_z \alpha(f|z)}{\sum_{f'} \sum_t \mathbf{F}_{f',t} P(z|f',t) + \kappa_z \alpha(f'|z)}$$

$$P(t|z) = \frac{\sum_f \mathbf{F}_{f,t} P(z|f,t) + \mu_z \alpha(t|z)}{\sum_f \sum_{t'} \mathbf{F}_{f,t'} P(z|f,t') + \mu_z \alpha(t'|z)}$$

$\kappa_z$ et $\mu_z$

paramètres scalaires permettant de régler à quel degré on veut que les distrib apprises ressemblent aux modèles
Décroissent au cours des itérations

# Séparation de sources guidée par l'utilisateur

G. Mysore et P. Smaragdis, *Waspaa 2009, LVA/ICA 2010*

(Demo video sur le site web de l'auteur)

Lien avec le principe de "Common Spatial Pattern"

# Cas ultime

- Le cas ultime représente l'information "oracle" des pistes séparées : ISS ou Informed Source Separation

- On "voit" les sources séparées un moment, on peut en extraire de l'info (à débit variable) qui aidera le décodage.

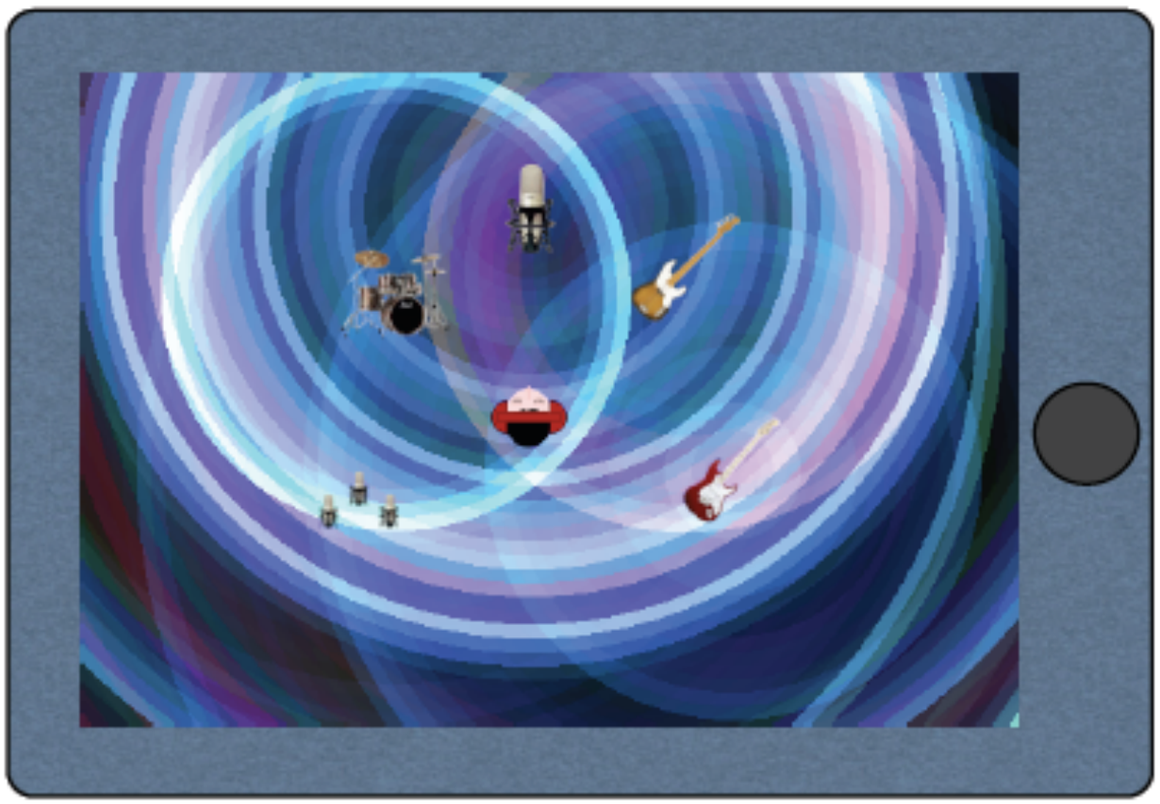
# Le projet ANR DReaM


## le Disque Repensé pour l'écoute active de la Musique

- LaBRI (coordinateur)
- GIPSA-lab
- Télécom ParisTech
- Institut Langevin
- iKlax Media

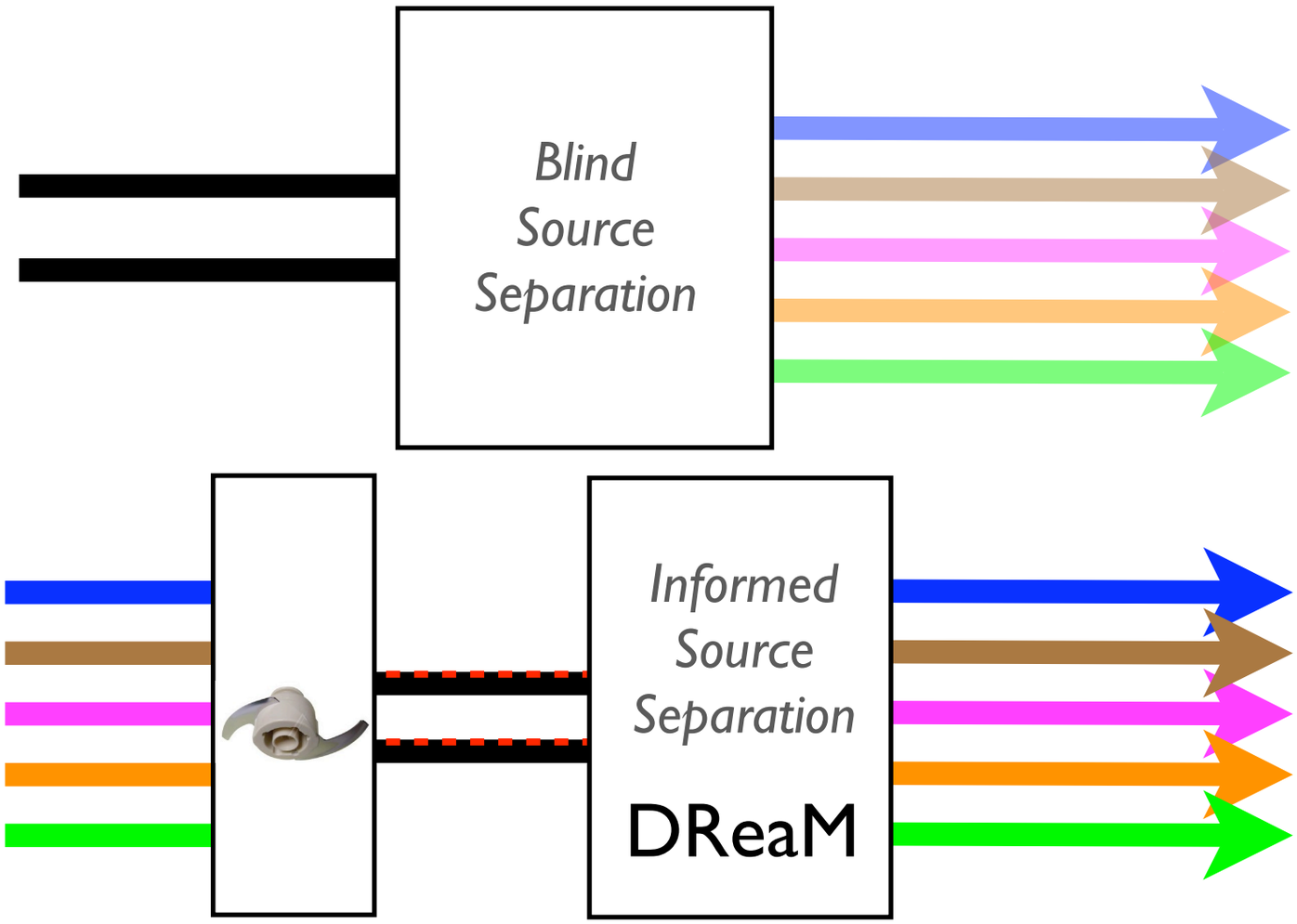
# Le projet DReaM

- L’info de démix est tatouée dans le signal lui-même (débit de tatouage “transparent” dans un canal type piste de CD: env 150 kbps)
- Fichiers audio “actifs”: transformation des paramètres musicaux (intensité, hauteur, timbre, temps, espace) de chaque entité sonore (source), en temps réel.



*(d’après S.Marchand)*

# Le projet DReaM




+

-

Entrée universelle

Qualité sonore limitée pour les pistes isolées (OK pour une re-spatialisation limitée)

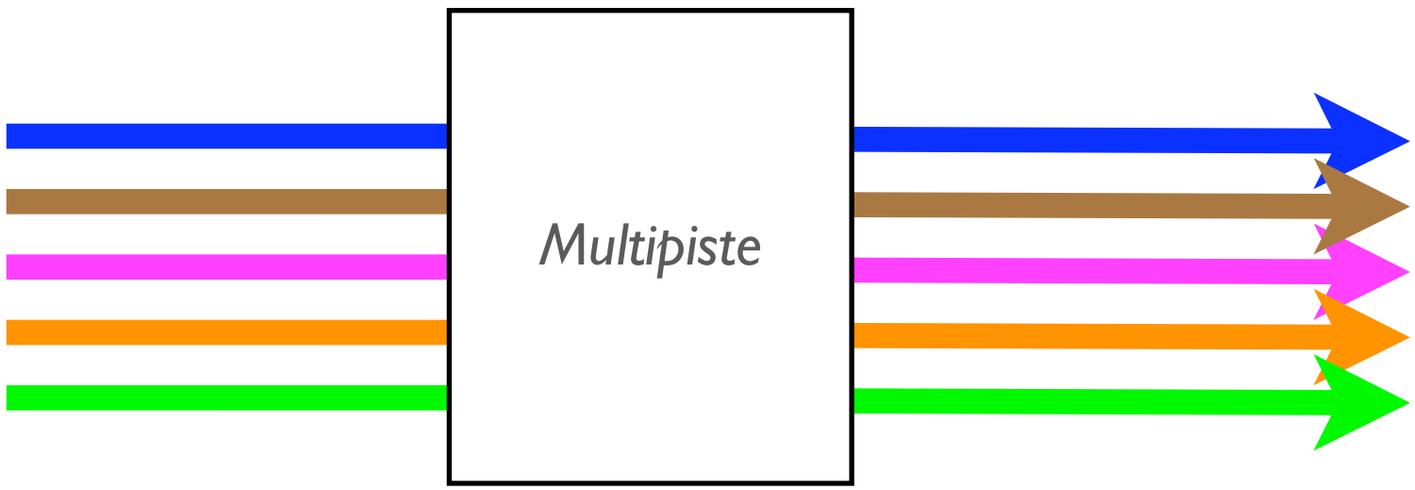



+

-

Très bonne qualité sonore
Compatible tous formats stéréo

Nécessite la «préparation» des fichiers



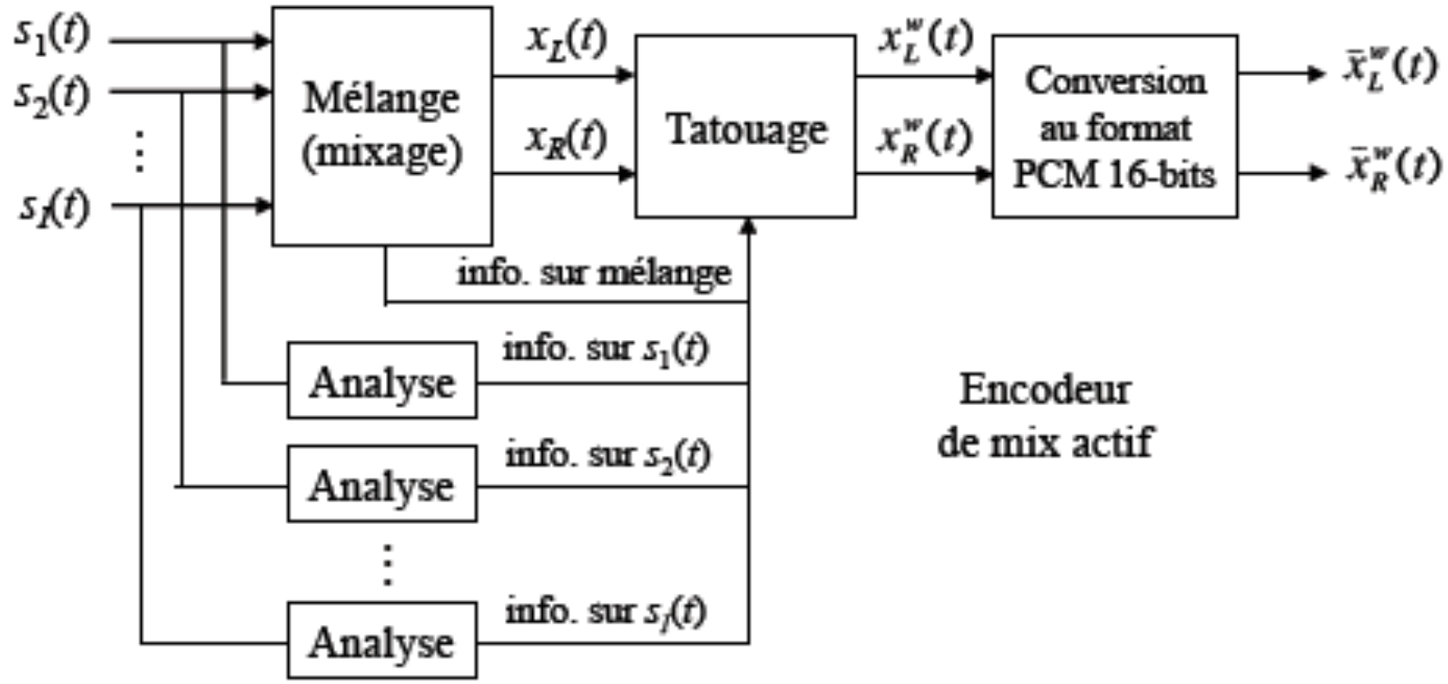

+

-

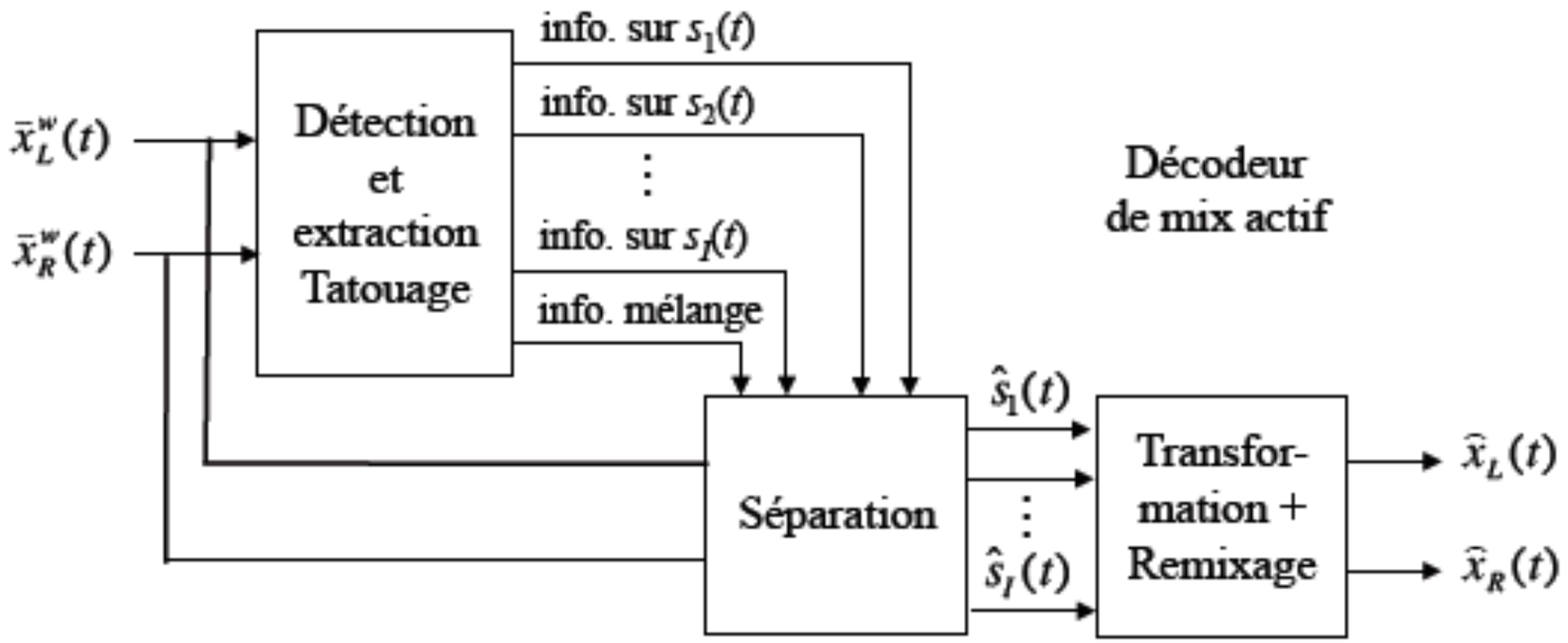
Qualité sonore parfaite, sans technologie

Nouveau *format* de diffusion multipiste

- codeur (au niveau du studio de production)




- décodeur (contrôlé par l'auditeur)




*(d'après S.Marchand)*

# Une implémentation

On peut ainsi “informer” toutes les méthodes précédentes.

Cf Bofill / Zibulevsky (2001) : à chaque bin t-f on fait correspondre 2 sources
Erreurs de choix des 2 sources (en particulier quand plus que 2 sources présentes)
Erreurs sur la matrice de mélange

- Idée (L. Girin / S. Marchand) : on informe la séparation en transmettant
    - La matrice de mélange “vraie”
    - En chaque point t-f (ou sur des groupes localement) les indices des 2 sources dominantes

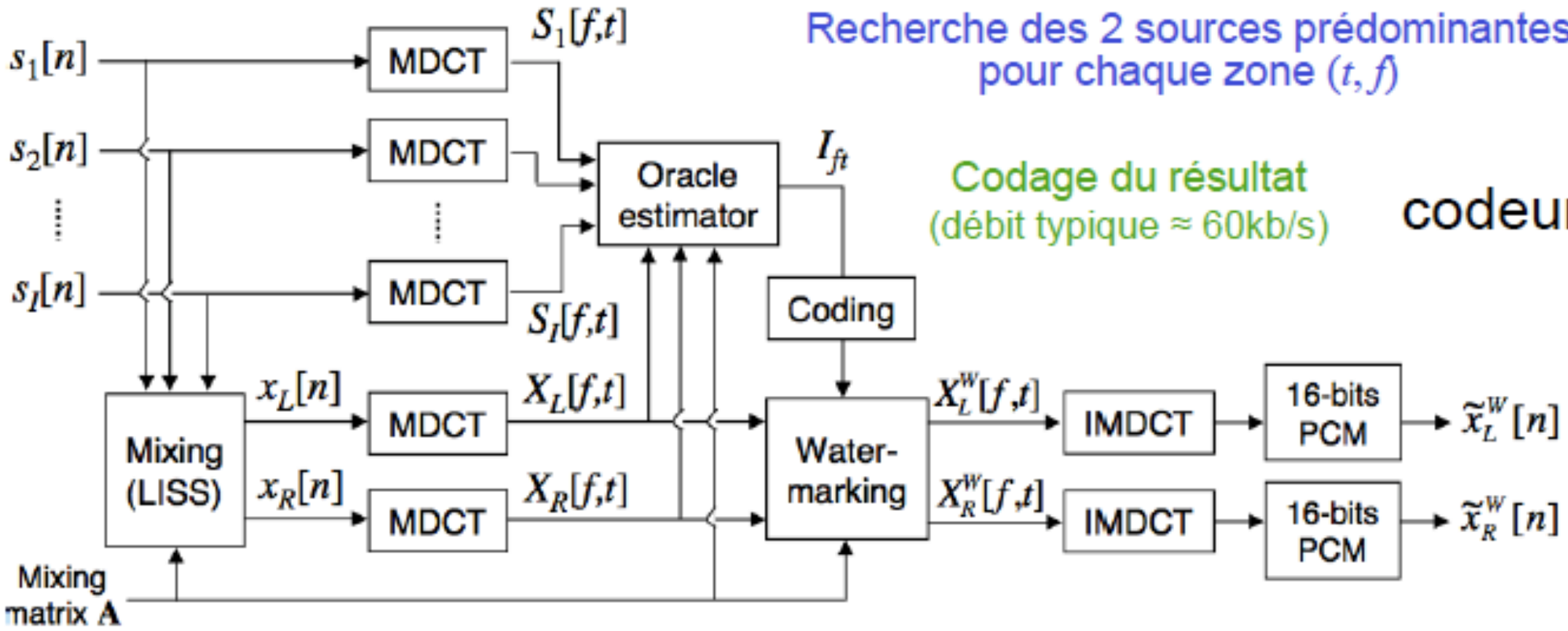
# Séparation par indexation des sources prédominantes

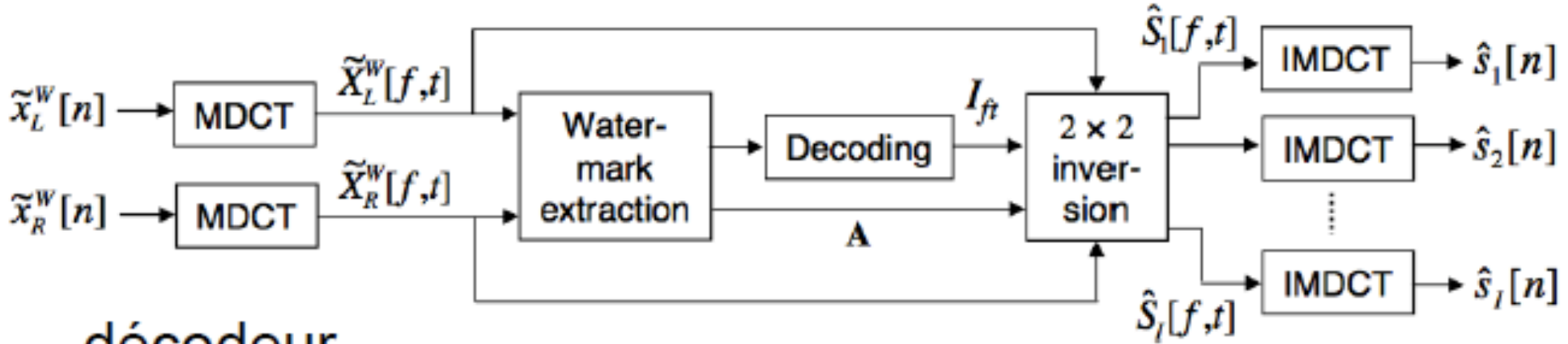
Transformée TF pour la parcimonie

Recherche des 2 sources prédominantes pour chaque zone $(t, f)$

Codage du résultat
(débit typique ≈ 60kb/s)

codeur



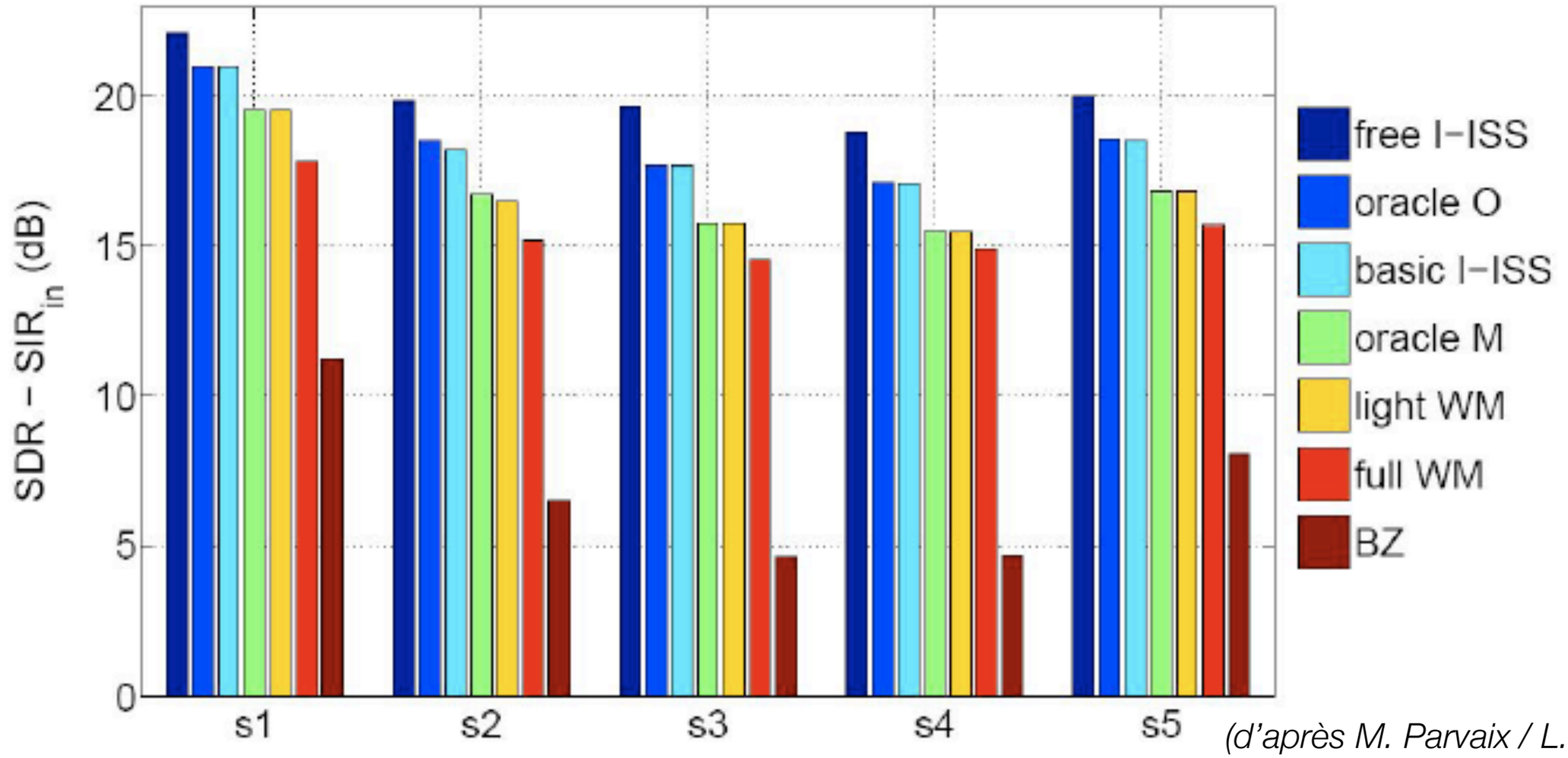




décodeur

Inversion locale du système avec 2 sources prédominantes

*(d'après M. Parvaix / L. Girin)*

# Séparation par indexation des sources : résultats

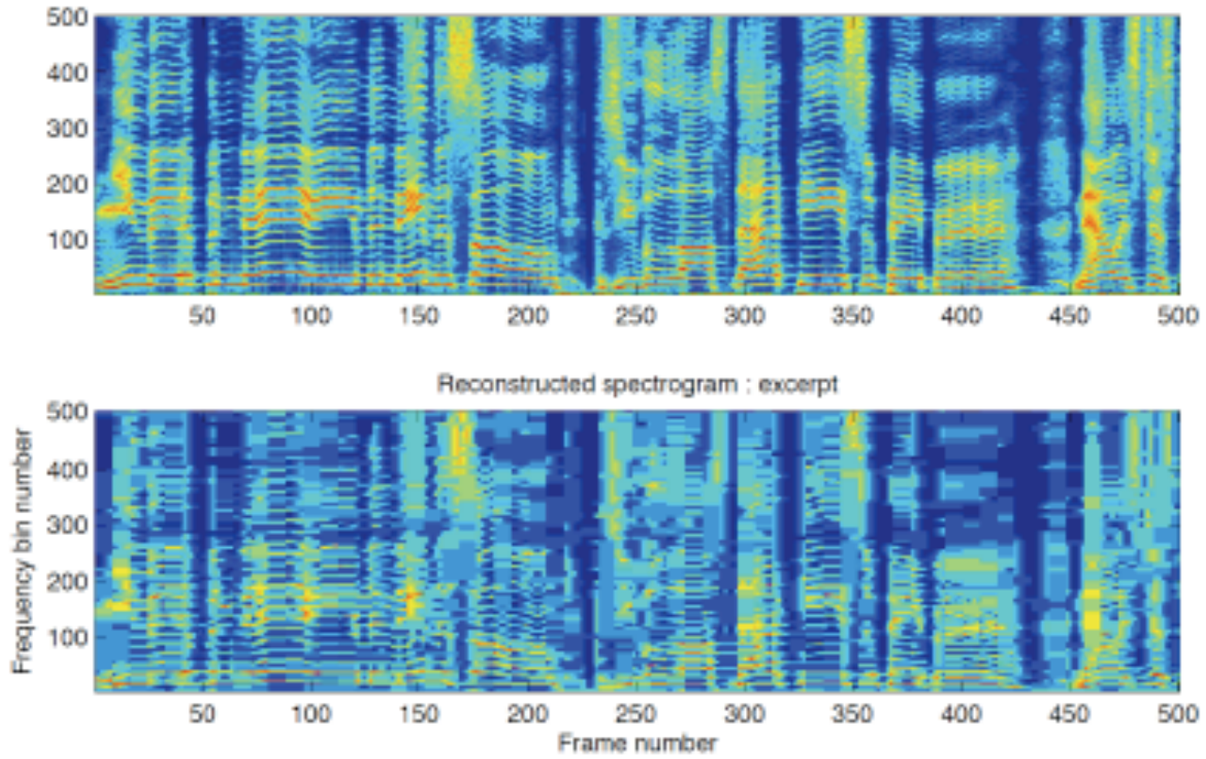



*(d'après M. Parvaix / L. Girin)*

- gains en rapport signal à bruit de l'ordre de 17 à 22 dB pour des mélanges à 5 sources
- qualité suffisante pour application de remix / karaoké
- simplicité du décodeur (faisabilité en temps réel)

# Séparation par filtrage de Wiener

- transmission des spectrogrammes des sources originales
- réduction du débit par compression (codage d'images)
- filtrage de Wiener (conservation de la phase du mélange)




*(d'après A. Liutkus et al.)*

# Autres approches ...

## Equipe de S. Marchand

- cf Dominique Fourer: modèle sinusoidal des sources
- cf Stanislaw Gorlow: clustering des sources prédominantes

## Equipe Telecom ParisTech - G. Richard / R. Badeau

- Antoine Liutkus: modèle gaussien des sources
- Alexey Ozerov: vue unifiée codage - ISS
  - Séparation de sources informée par codage

    ou
  - Codage informé par séparation

Challenge : être compatible avec tous les effets d'un 'vrai' mix

# Conclusion

# Conclusion

- La séparation de sources aveugle dans un contexte musical est un problème en général très (trop ?) difficile
  - nb sources >> nb canaux, grande diversité des sources, réverbération, mixage, perception très exigeante ...

- Les tendances actuelles sont à injecter de plus en plus d'information dans les algorithmes :
  - modèle de sources, modèles de "partitions" (gain(t))
  - séparation informée par l'utilisateur
  - cas extrême : séparation de source informée

- Domaine très actif

# Est-ce juste un problème académique ?

Depuis quelques années on commence à voir les premières applications industrielles en post-production cinéma

- - demo video Audionamix -

# That’s all !

Merci à Rémi Gribonval, Pierre Leveau, Sylvain Marchand, Gaël Richard et Emmanuel Vincent.