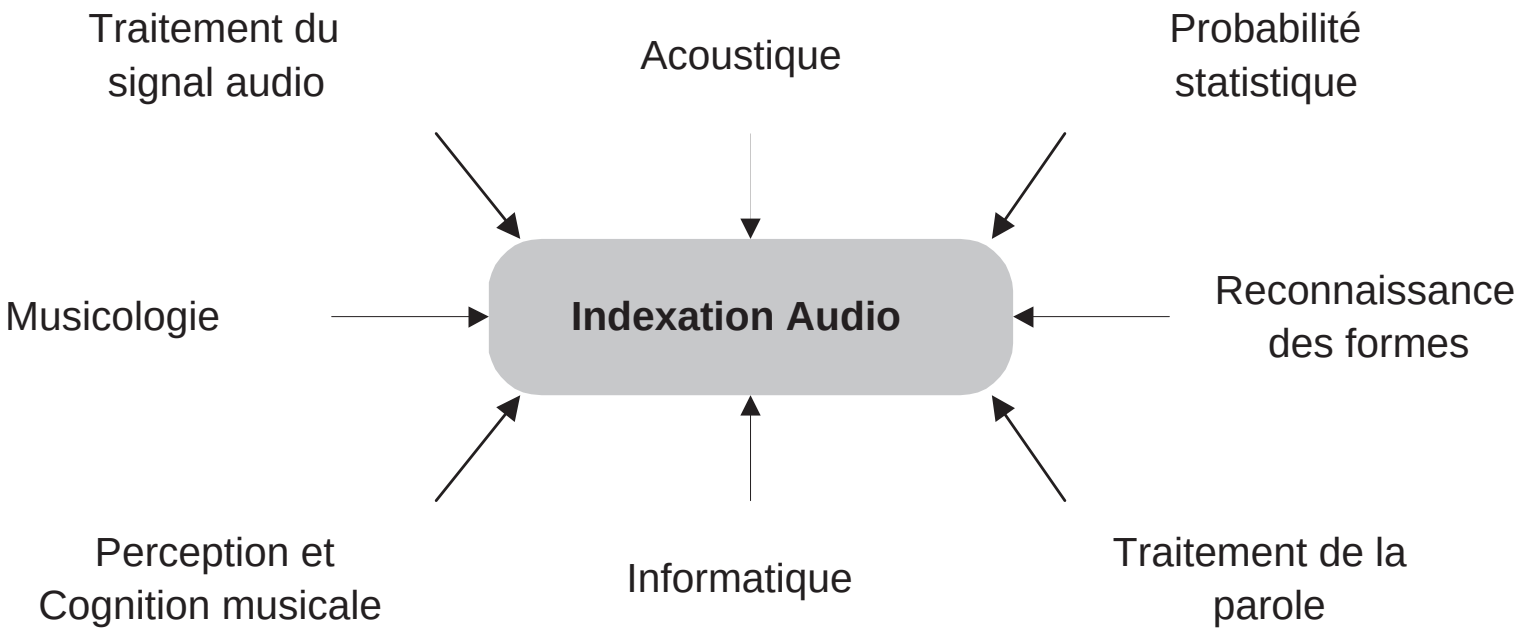
NSY122: Extraction d'information audio (partie 1) - peeters@ircam.fr

Geoffroy.Peeters@ircam.fr

Introduction

Indexation audio= domaine-multi-disciplinaire



Nombreuses applications

d'après Richard

Nombreuses applications

- Recherche d'un contenu audio dans une base de données en ligne autrement que par "artistes", "titres" (Google musical)
- Nouveaux modes de recherche: par chantonnement/ sifflement
- Recherche par similarité de contenu
- Dé-linéarisation d'un flux audio: segmentation de flux radio, télé et étiquetage des parties
- Analyse musicale automatique (audio vers partition)
- Création automatique de play-list sur base de critères de contenu (morceaux à tempo de plus en plus rapide)
- Organisation de données locales (base d'échantillons audio pour la création)
- Pré écoute musicale
- Adaptation de la musique à l'humeur (musique joyeuse, triste), à l'activité (jogging, wake-up)
- Identification audio (recherche de doublons, gestion de copyright, attacher des méta données à une instance d'un morceau)

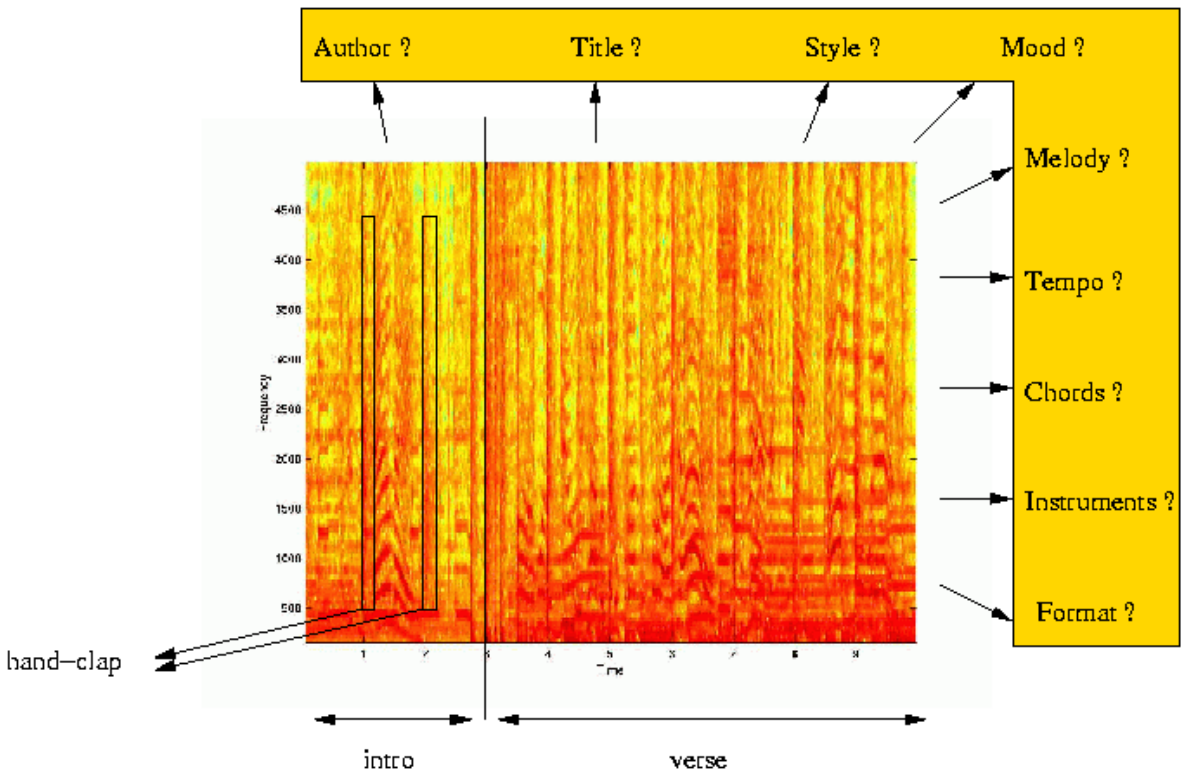
Indexation manuelle ou automatique

Indexation manuelle:

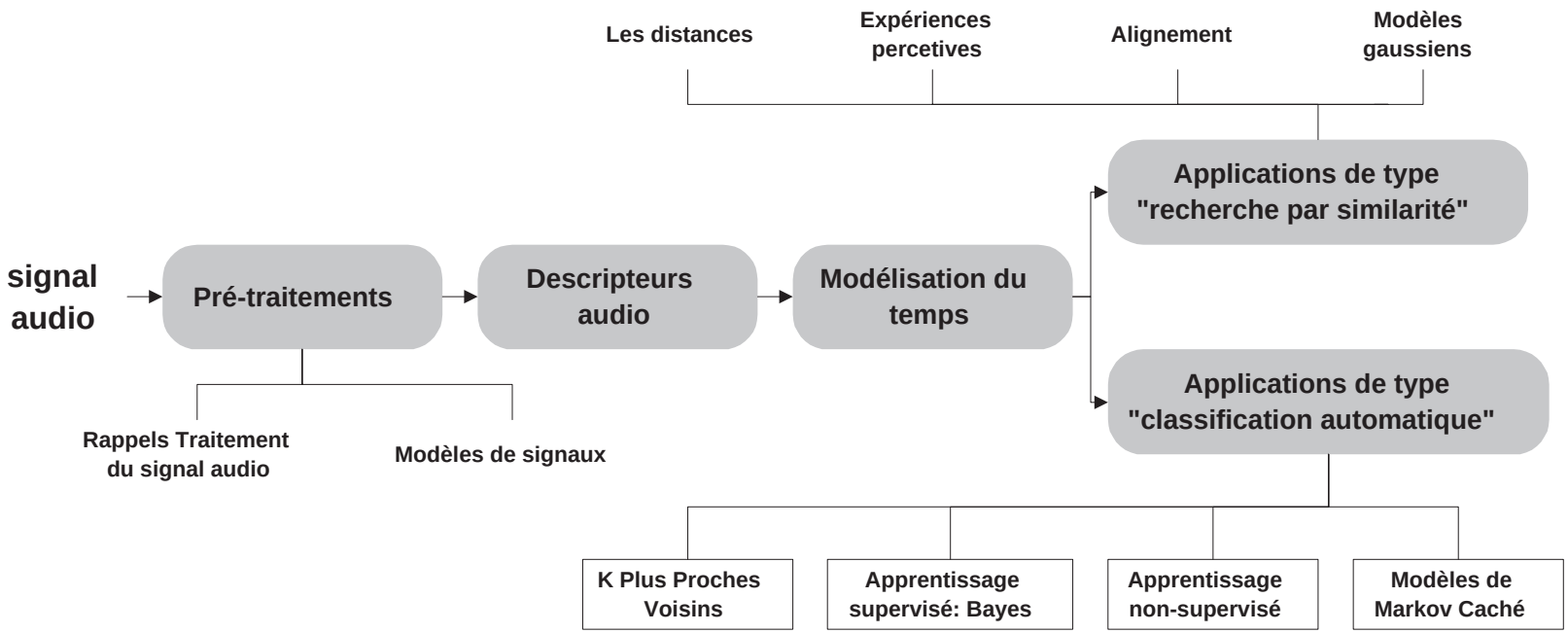
- tout ne peut pas être extrait manuellement (audio ID, timbre musical)

Indexation automatique

- description de contenu
- gain de temps
- tout ne peut pas être extrait automatiquement (exemple: lieu d'enregistrement)

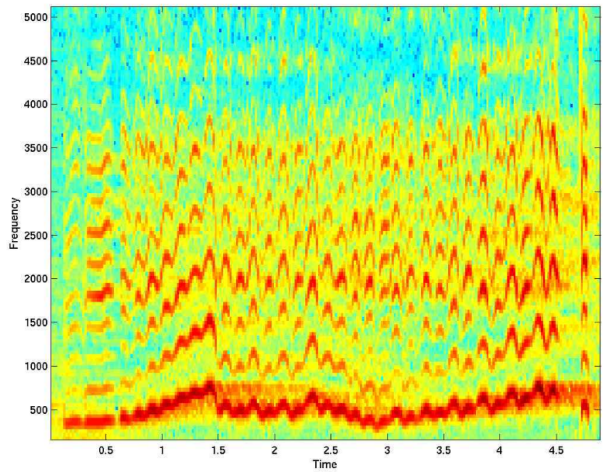


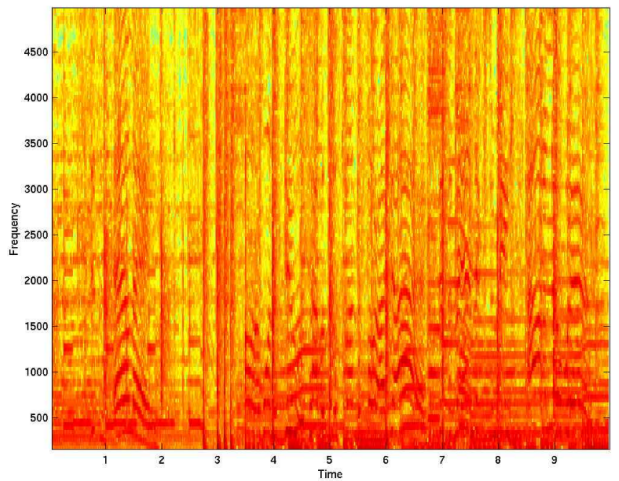
Plans


Les distances
Expériences percetives
Alignement
Modèles gaussiens
Applications de type "recherche par similarité"
signal audio
Pré-traitements
Descripteurs audio
Modélisation du temps
Applications de type "classification automatique"
Rappels Traitement du signal audio
Modèles de signaux
K Plus Proches Voisins
Apprentissage supervisé: Bayes
Apprentissage non-supervisé
Modèles de Markov Caché

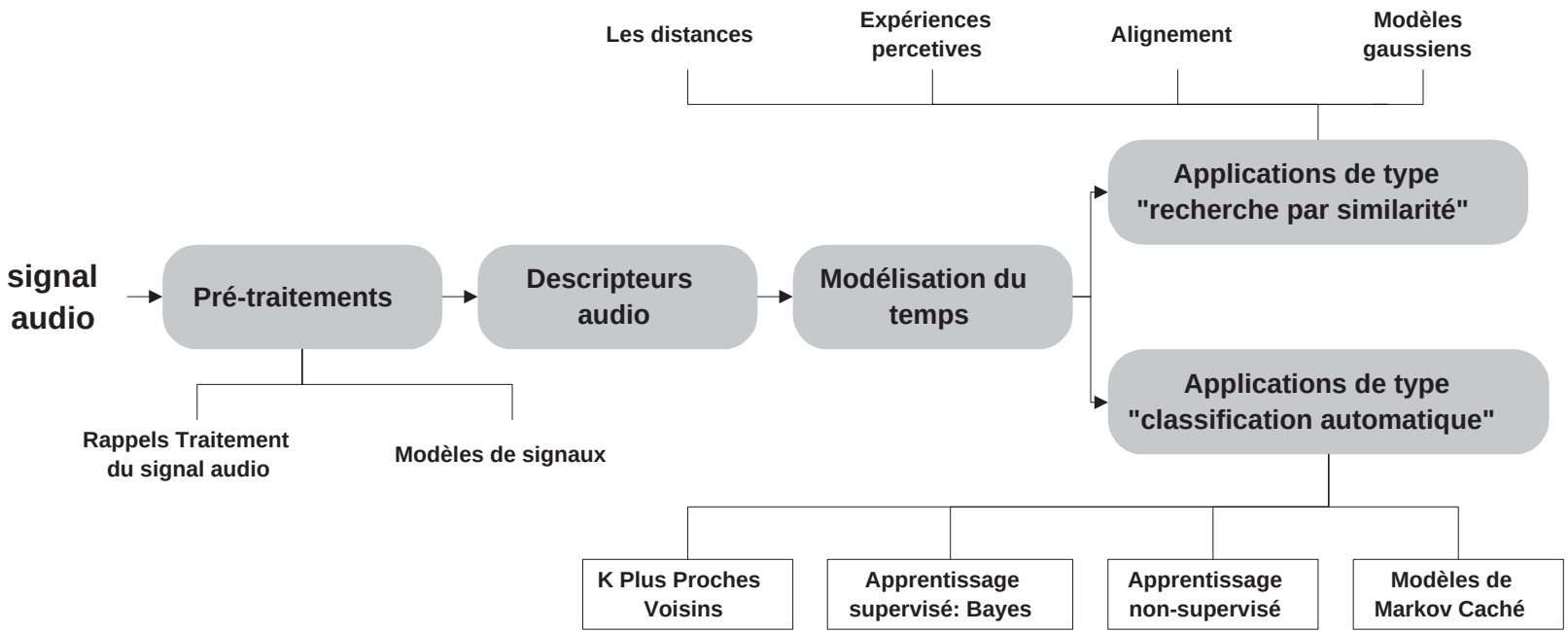
Verrous technologiques actuels

- Toutes les sources sont superposées simultanément (signal audio= une image transparente)
- Les sources varient au cours du temps
- Toutes les hauteurs sont superposées





Plans


Les distances
Expériences percetives
Alignement
Modèles gaussiens
Applications de type "recherche par similarité"
signal audio
Pré-traitements
Descripteurs audio
Modélisation du temps
Applications de type "classification automatique"
Rappels Traitement du signal audio
Modèles de signaux
K Plus Proches Voisins
Apprentissage supervisé: Bayes
Apprentissage non-supervisé
Modèles de Markov Caché

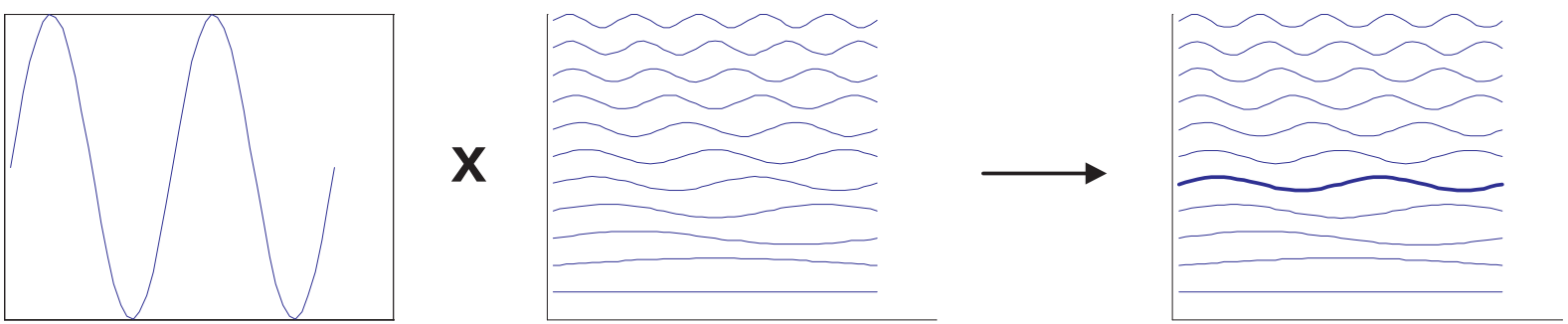
Transformée de Fourier

- Transformée de Fourier du signal $x(t)$

$$X(\omega) = \int_t x(t) exp(j\omega t) dt \quad (1)$$

$\omega = 2\pi f$ les fréquences exprimées en radian,
$exp(j\omega t) = cos(2\pi ft) + jsin(2\pi ft)$.

- Pourquoi la Transformée de Fourier ?
  - Il est difficile d'extraire des observations directement à partir de la forme d'onde $x(t)$ d'un signal audio
  - On cherche à reproduire la décomposition en fréquences de l'oreille humaine



Transformée de Fourier Discrète

- Echantillonnage temporel:
  - $x(n)$ est le signal $x(t)$ échantillonné à la fréquence sr (sampling rate)
  - Exemple: $sr = 44100Hz$ on prend 44100 valeurs de $x(t)$ par seconde
- Transformée de Fourier à temps n et fréquences k discrètes sur N points= TFD

$$X(k) = \sum_{n=0}^{N-1} x(n) exp\left(j2\pi \frac{k}{N} n\right) \forall k \in [0, N] \quad (2)$$

  - $\int \leftrightarrow \sum$
  - $t \leftrightarrow n$
  - $\omega = 2\pi f \leftrightarrow 2\pi \frac{k}{N}$. Les fréquences correspondantes en Hz sont $f = \frac{k}{N} sr$

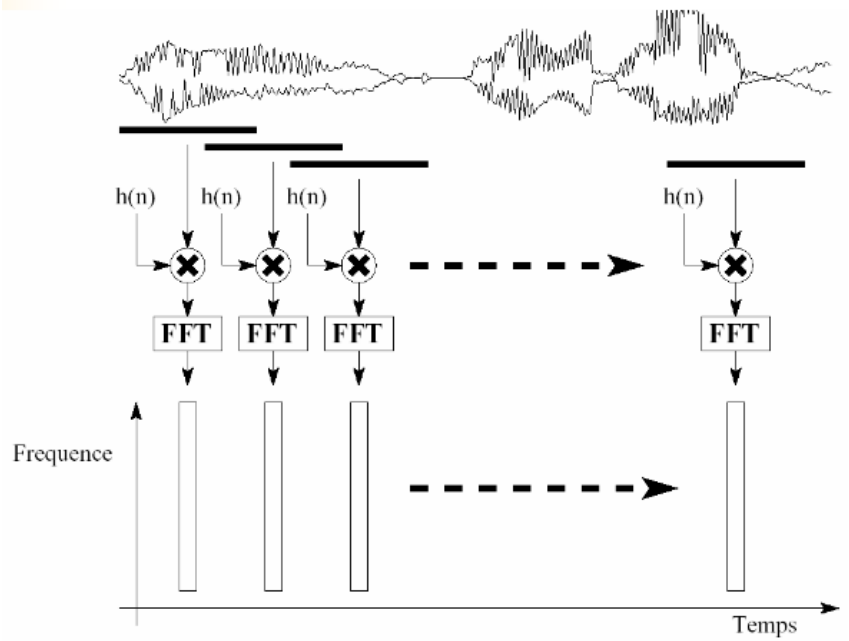
Transformée de Fourier à Court Terme

- Transformée de Fourier à Court Terme: application de la TFD à une portion du signal centrée autour de l'échantillon m

$$X(k,m) = \sum_{n=0}^{N-1} x(n)h(m-n)exp\left(j2\pi\frac{k}{N}(m-n)\right) \forall k \in [0,N] \quad (3)$$

Pourquoi la Transformée de Fourier à Court Terme?

- Le signal audio est non-stationnaire: ses propriétés varient au cours du temps
- On considère cependant que ses propriétés sont "localement" (en temps) stationnaires: sur une durée de $\pm$ 40ms
- On effectue une suite d'analyse de Fourier sur des durées de $\pm$ 40ms= analyse à Court Terme ("trames/frames" en vidéo)



d'après Laroche

Fenêtre de pondération

- $h(n)$ est appelé "fenêtre de pondération". $h(n)$ est défini sur un horizon fini $[0, N]$.
- Différents choix de fonction de pondération; détermine les caractéristiques de l'analyse spectrale.
  - rectangulaire: $h(n) = 1$
  - hanning: $h(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1}))$
  - hamming: $h(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$
  - blackman: $h(n) = a_0 - a_1\cos(\frac{2\pi n}{N-1}) + a_2\cos(\frac{2\pi n}{N-1})$ avec $a_0 = ...$
- Différents choix de longueur de l'horizon de $h(n)$ détermine également les caractéristiques de l'analyse spectrale.
  - Au plus la fenêtre est longue, au plus on observe précisément les fréquences.
  - Au plus la fenêtre est courte, au plus on observe précisément les temps.

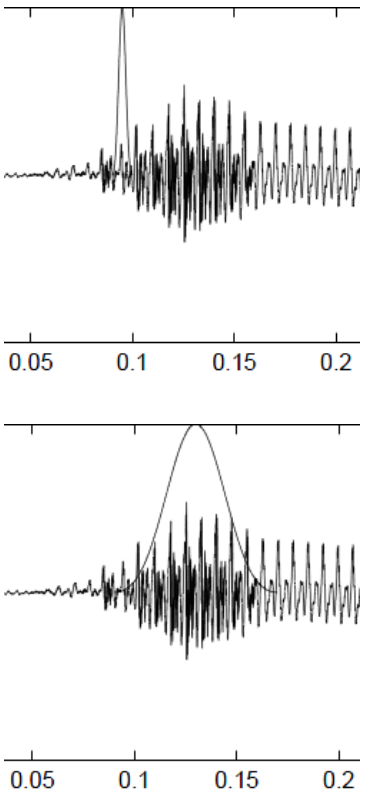
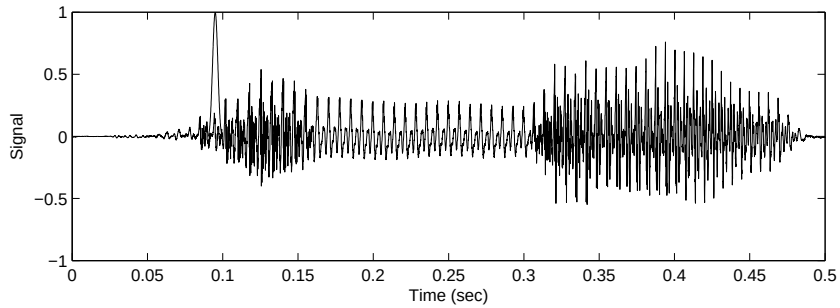
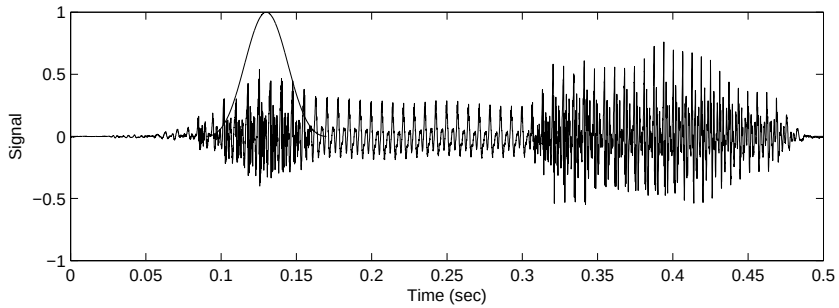
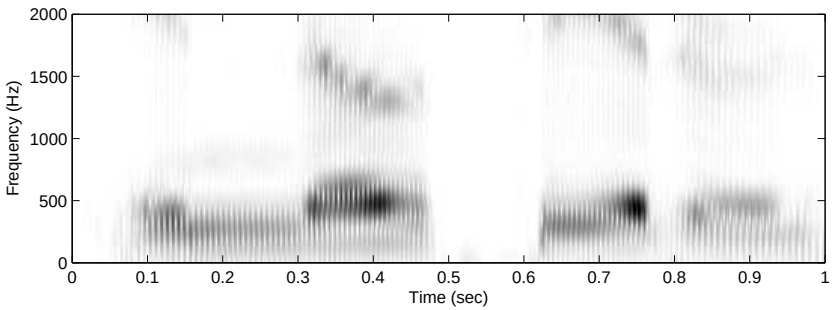
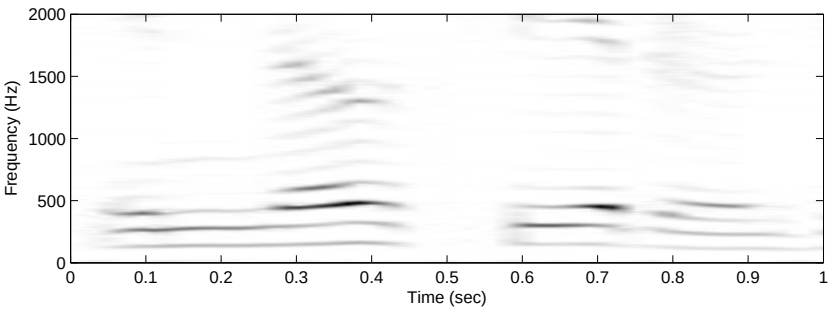


Illustration du paradoxe temps/ fréquence









Comme résoudre ce problème ?

- Utiliser d'autres tranformées que celle de Fourier
- Ondelette, Transformée à Q-Constant, ...

Illustrations logiciel

Transformée de Fourier à Court Terme

- $X(k,m)$ est un nombre "complexe": 0.32+j0.64 (point dans le plan complexe)

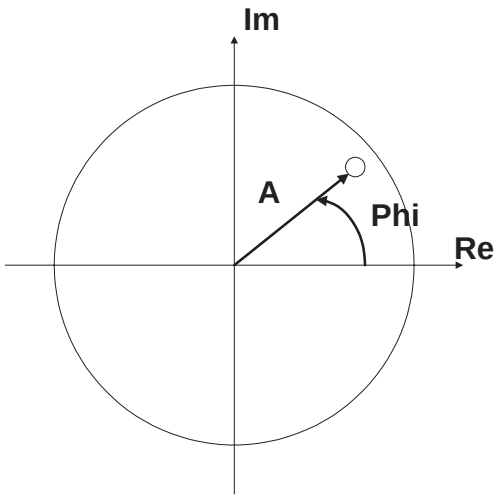
$$exp\left(j2\pi\frac{k}{N}n\right) = \cos\left(j2\pi\frac{k}{N}n\right) + j\sin\left(j2\pi\frac{k}{N}n\right) \quad (4)$$

  - Le spectre d'amplitude $A(k,m)$ est le module de $X(k,m)$ (distance du point à l'origine du plan complexe)

$$A = \sqrt{Re^2 + Im^2} \quad (5)$$

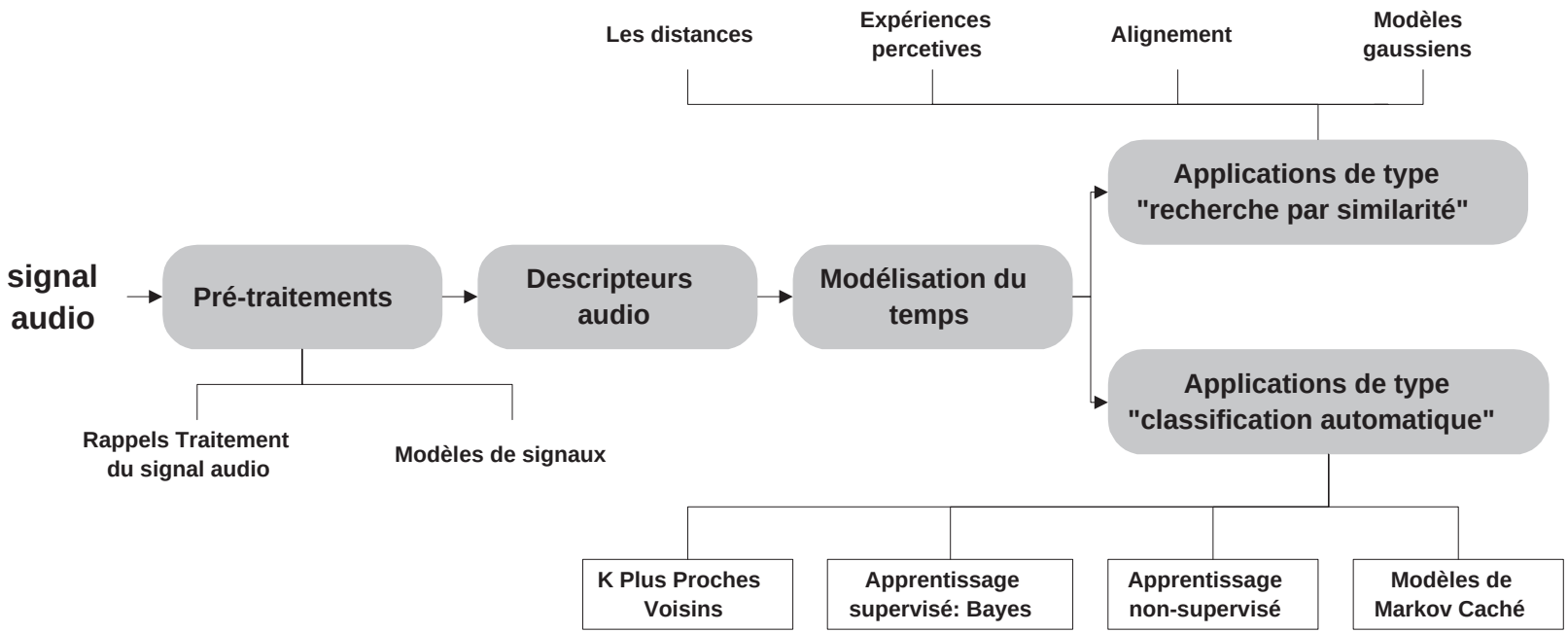
  - Le spectre de phase $\phi(k,m)$ est l'angle de $X(k,m)$

$$\phi = -arctan\left(\frac{Im}{Re}\right) \quad (6)$$



- Le "spectrogramme" / "sonagramme" représente les spectres d'amplitude $A(k,m)$ au cours des différentes analyses à court terme m
  - Représentation en 2D, en 3D

Plans


Les distances
Expériences percetives
Alignement
Modèles gaussiens
Applications de type "recherche par similarité"
signal audio
Pré-traitements
Descripteurs audio
Modélisation du temps
Applications de type "classification automatique"
Rappels Traitement du signal audio
Modèles de signaux
K Plus Proches Voisins
Apprentissage supervisé: Bayes
Apprentissage non-supervisé
Modèles de Markov Caché

Modèles de signaux

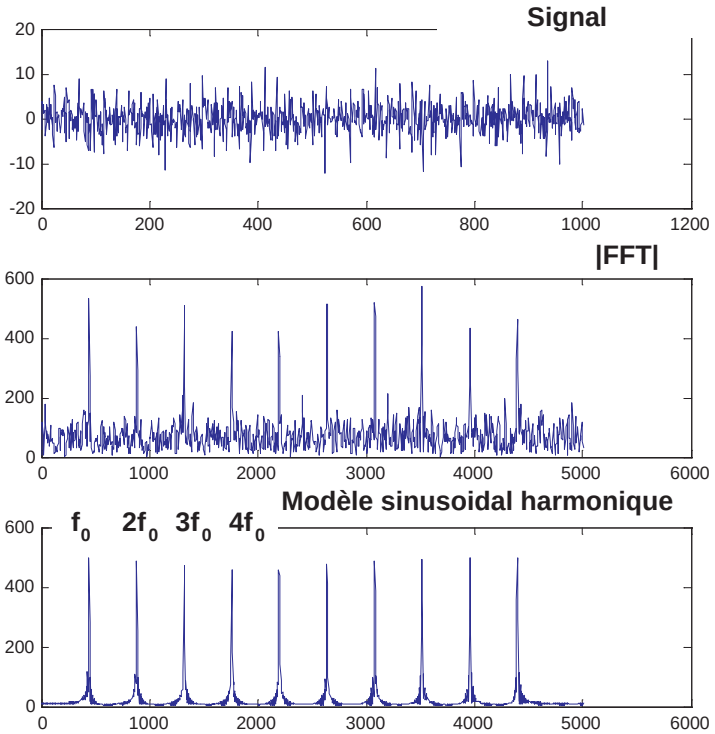
Pourquoi des modèles de signaux ?

- On suppose que le signal a été produit par un certain modèle
- Permet de réduire le nombre de paramètres observés du signal (la TFCT contient beaucoup trop d'information)
- Permet d'obtenir des paramètres plus facilement interprétables (indexation) et manipulables (transformation, synthèse) (la TFCT fournit des paramètres non directement exploitables)
- Quels modèles ?
  1. Modèle sinusoidal harmonique
  2. Modèle source/ filtre
  3. Modèle autorégressif

Modèle sinusoidal harmonique

- Hypothèse: le signal $x(t)$ est un signal harmonique
- Exemple: une note de musique (les instruments de musique produisent généralement des sons harmoniques), les parties voisées de la voix
- Modélisation: le signal est représenté comme une somme de sinusoides dont les fréquences sont des multiples entiers de la fréquence fondamentale (la hauteur de la note) + un résiduel (du bruit)
- Exemple: Pour un instrument de musique jouant un *la*3($A4$), sa fréquence fondamentale (f_0) est $440Hz$, il peut être modéliser comme la somme de ses harmoniques:
  $f_0 = 440Hz, 2f_0 = 880Hz, 3f_0 = 1320Hz, ...$

$$X(f) = \sum_{h=1}^{H} A_h sin(2\pi h f_0 + \phi_h) + \epsilon(f) \qquad (7)$$

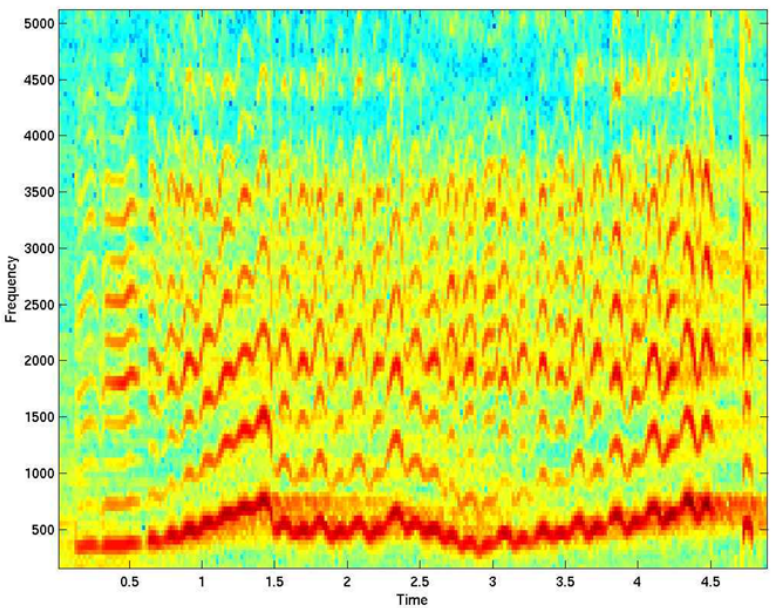


- A_h est l'amplitude de l'harmonique h, hf_0 la

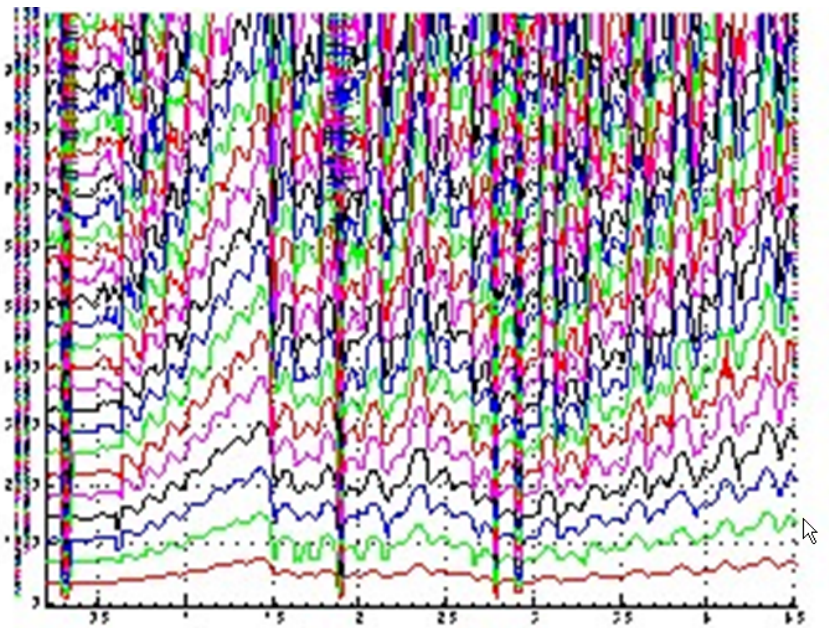
Modèle sinusoidal harmonique

Les sinusoides harmoniques varient au cours du temps

TFCT



Modèle sinusoidal



Modèle harmonique

Utilisation du modèle sinusoidal ?

- Synthèse, modification du signal
- Codage (Paramétrisation du signal comme f_0 + enveloppe spectrale $[A_1, A_2, ..., A_H]$)
- Extraction de paramètres pour l'indexation

Différentes estimations de la fréquence fondamentale

- Temporel: auto-corrélation, Average Mean Difference Function, Yin, Cepstre
- Filtre en peigne harmonique, Maximum de vraisemblance, ...

Différentes paramétrisation de l'enveloppe spectrale

- LPC (Linear Predictive Coding)
- Cepstre
- MFCC

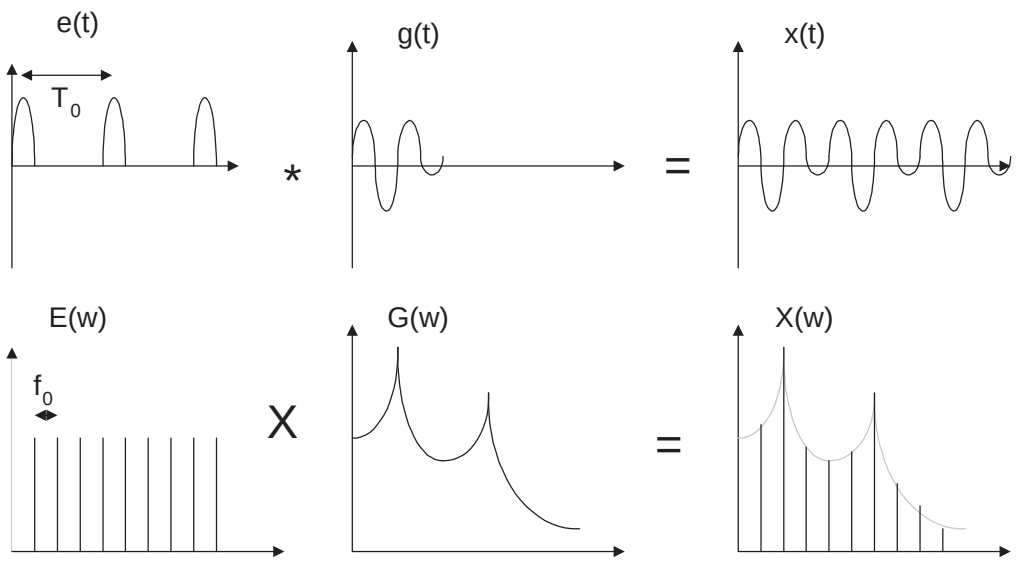
Modèle source/ filtre

- Hypothèse: le signal $x(t)$ est le résultat du passage d'une excitation (un pulse, une série de pulse) dans un filtre (résonnant)
- Exemples: le signal de parole, certains instruments de musique (trompette)
- Représentation mathématique: un signal d'excitation $e(t)$ passe (convolution) à travers un filtre $g(t)$

$$x(t) = e(t) * g(t) \qquad (8)$$

- Représentation dans le domaine de la TF

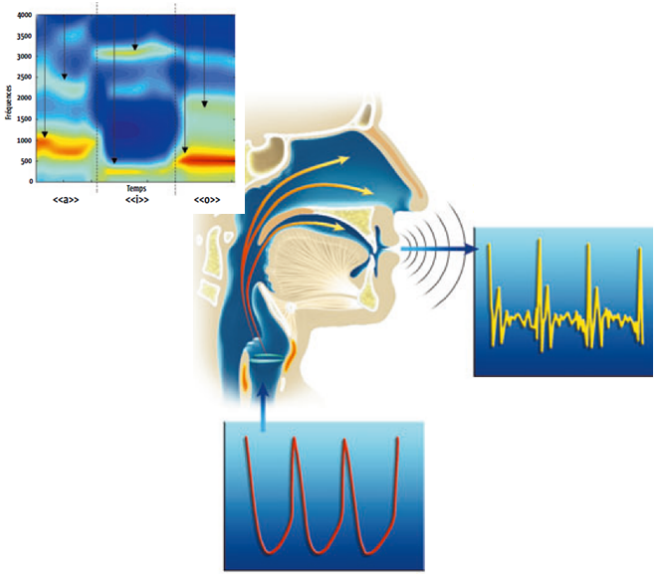
$$X(\omega) = E(\omega)G(\omega) \qquad (9)$$



Modèle source/ filtre

Utilisation pour la modélisation du signal de parole:

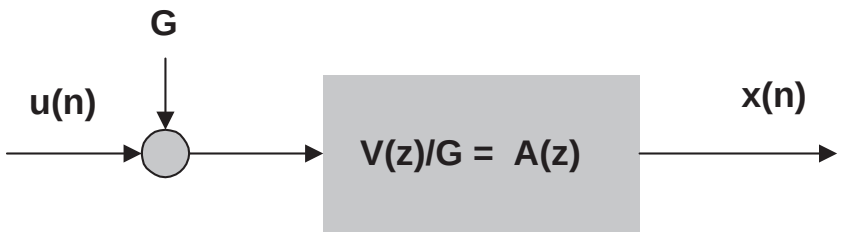
- Le signal de parole (pour sa partie voisée) est créé par une excitation régulière (les cordes vocales) passant ensuite dans le conduit bucco-nasal créant des résonances (bouches) et anti-résonances (nez).
- Périodicité des cordes vocales (ouverture/fermeture) détermine la hauteur
  - Hauteur de 100Hz ? pulses d'air sont espacés de $T_0 = 1/f_0 = 1/100 = 10ms$.
  - Appelé signal d'excitation (ou signal source), $e(t)$.
- Conduit bucco-nasal permet de créer les différentes voyelles pour une hauteur donnée en renforcant (résonance) et retirant (anti-résonances) certains fréquences.
  - Filtre résonant (AR: Auto-Regressif) et anti-résonant (MA: Moving Average): un filtre dit "ARMA".



d'après Peeters, La Recherche

La prédiction linéaire: le modèle auto-régressif

- Modèle auto-régressif: le signal à un instant n peut être prédit à partir du (est une combinaison linéaire de) du signal aux instants précédents
- $x(n) = a_1 x(n-1) + a_2 x(n-2) + a_3 x(n-3) ... + a_P x(n-P)$
- Modèle plus général: $x(n) = G \cdot u(n) + \sum_{j=1}^{p} a_j x(n-j)$
- équivalent à passé le signal dans un filtre FIR tout -pôle:



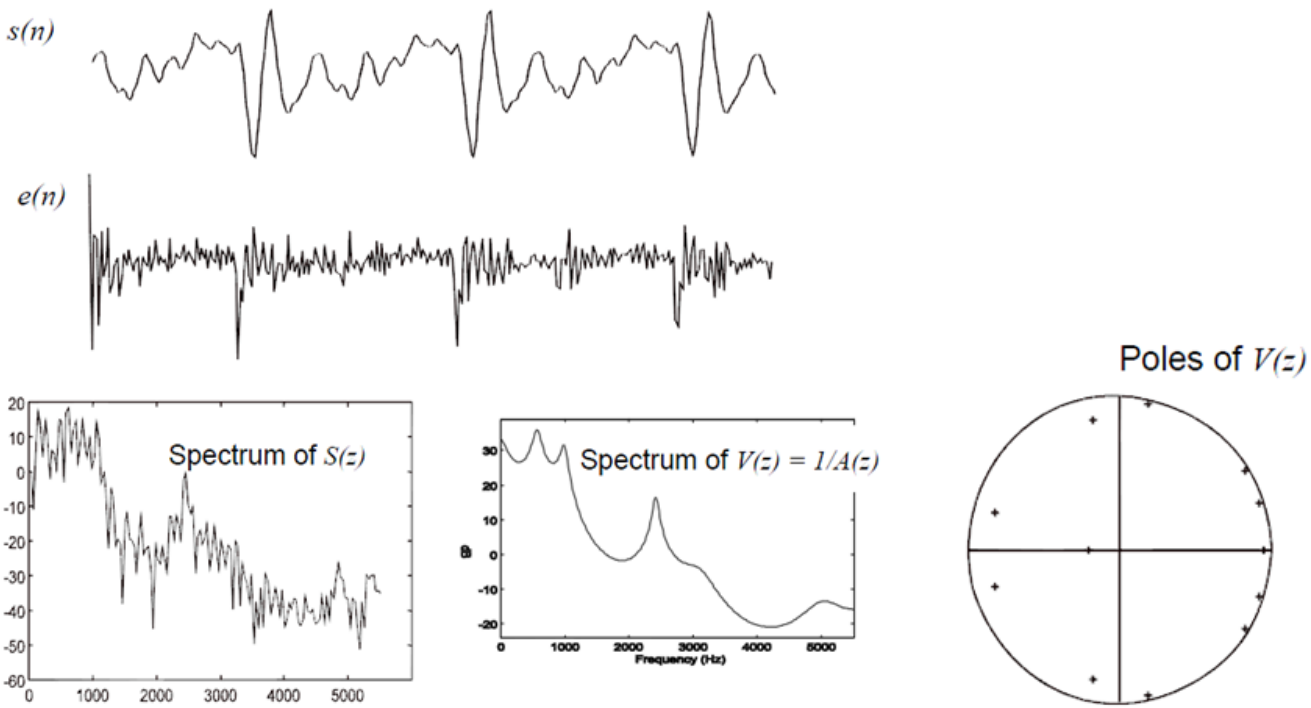
$$V(z) = \frac{G}{1 + \sum_{j=1}^{P} a_j z^{-j}} \tag{10}$$

- Objectif de la prédiction linéaire:
  - déterminer le filtre $V(z)$ (donc les résonances ou les formants dans le cas de la voix) à partir du signal $x(n)$ (signal de pression micro)
- Comment estimer $V(z)$?
  - minimisation de l'erreur quadratique entre signal observé $x(n)$ et sa modélisation $\hat{x}(n)$: $e(n) = x(n) - \sum_{j=1}^{p} a_j x(n-j)$
  - Résolution par inversion de matrice; méthode de l'auto-corrélation, méthode de la co-variance

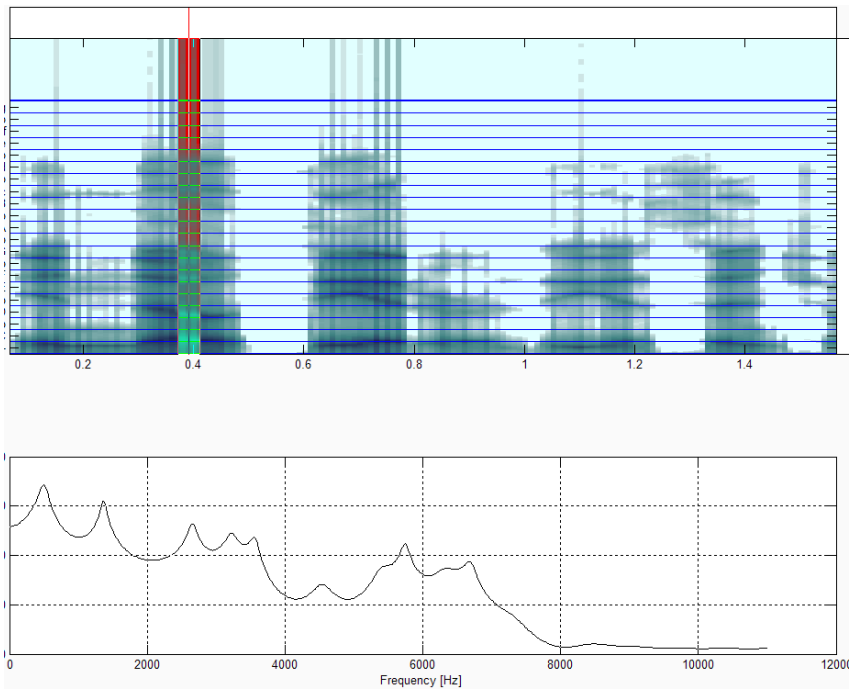
La prédiction linéaire

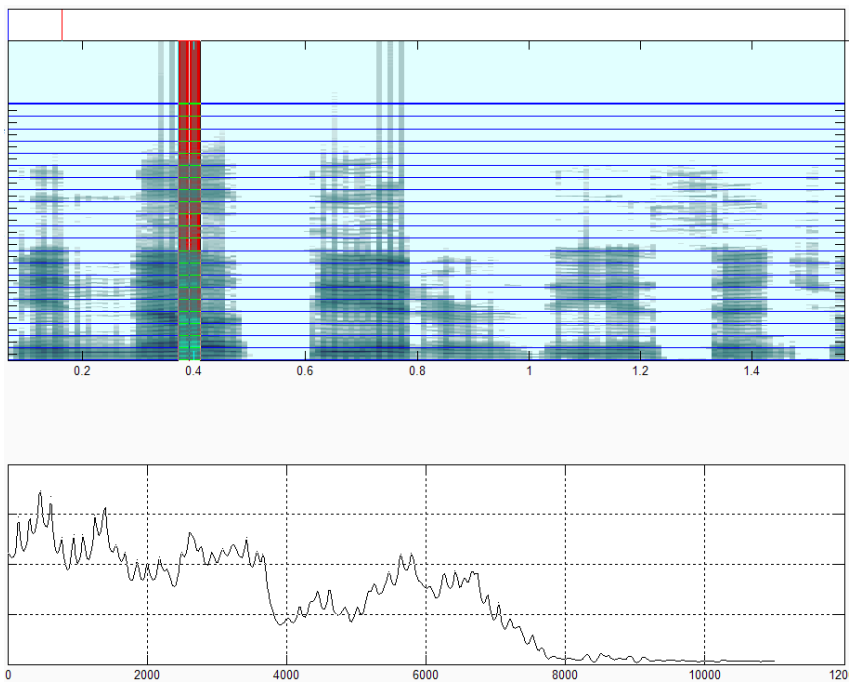
Représentation des résonances:

- Paramètres du filtre AR: a_k,
- Coefficients de réflexion, log area ratios (LAR),
- Coefficients cepstraux
- Line Spectrum Frequencies (LSF)



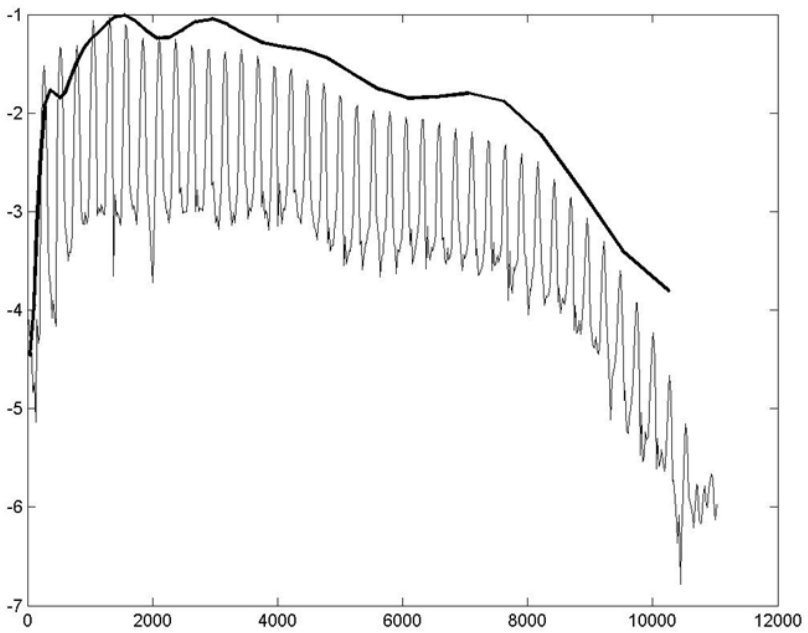
Choix du nombre de pôle P



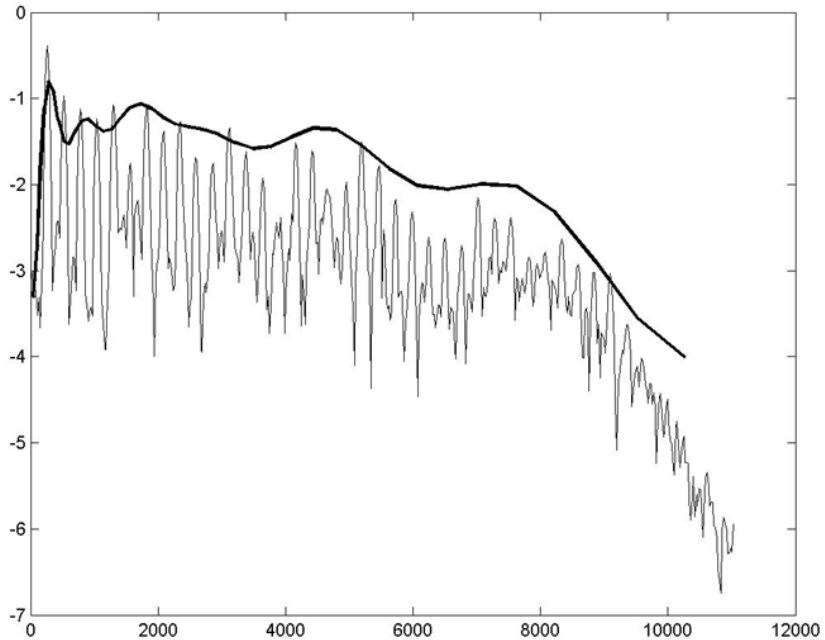


Importance de la fréquence fondamentale et de l'enveloppe spectrale

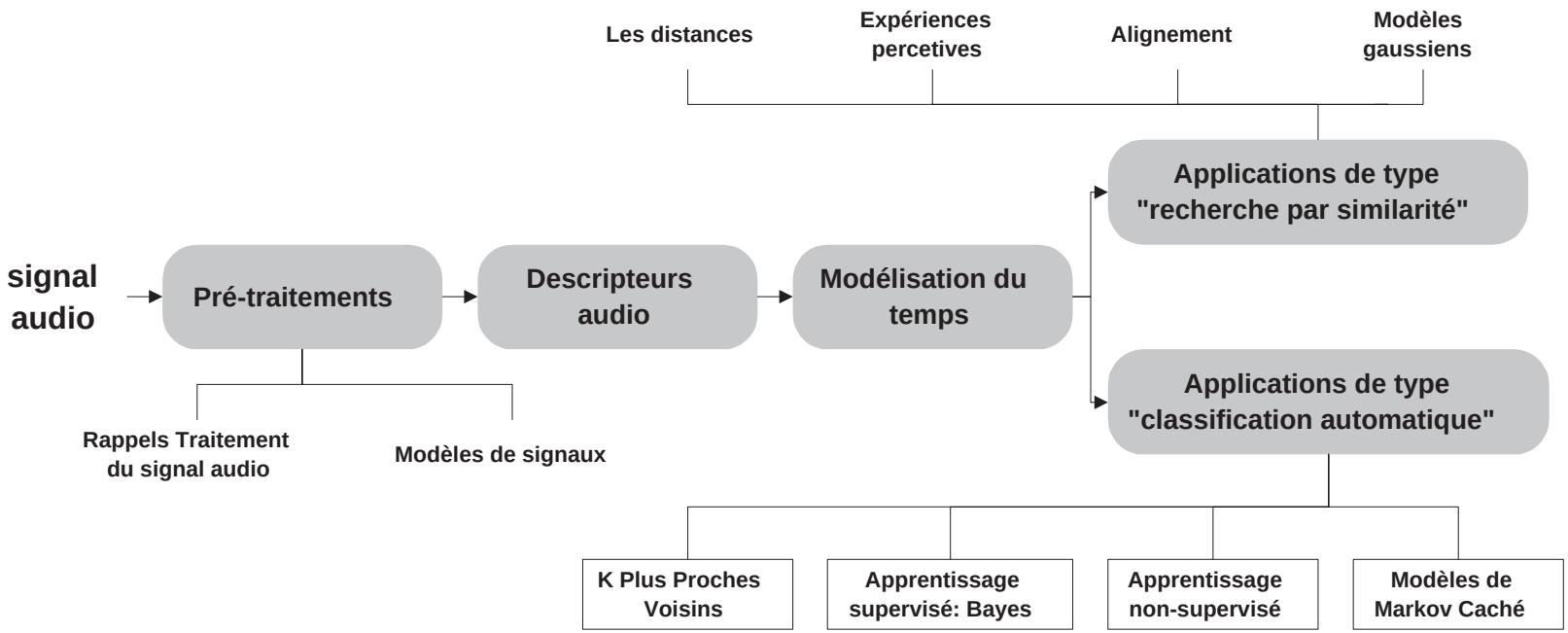
Enveloppe spectrale son de trompette



Enveloppe spectrale son de violon



Plans


Les distances
Expériences percetives
Alignement
Modèles gaussiens
Applications de type "recherche par similarité"
signal audio
Pré-traitements
Descripteurs audio
Modélisation du temps
Applications de type "classification automatique"
Rappels Traitement du signal audio
Modèles de signaux
K Plus Proches Voisins
Apprentissage supervisé: Bayes
Apprentissage non-supervisé
Modèles de Markov Caché

Taxinomie des descripteurs audio

On peut distinguer les descripteurs audio selon plusieurs points de vue

- l'aspect statique/ dynamique d'un descripteur:
  - est-ce que le descripteur représente le signal à un instant donné, où son évolution au cours du temps
- l'étendue temporelle que décrit le descripteur:
  - est-ce que le descripteur décrit un instant du signal (FFT) ou la globalité d'un morceau de musique (artiste)
- le concept que décrit le descripteur:
  - Exemple: l'enveloppe spectrale, le contenu harmonique
- le mode d'extraction du descripteur
  - à partir de l'auto-corrélation, à partir du spectre
- sa forme:
  - scalaire, vecteur, modèle statistique
- à quel type de contenu le descripteur est appliqué (certains descripteurs reposent sur un modèle applicable uniquement sur certains types de signaux):
  - échantillon audio (son unitaire, mono-source, mono-phonique),
  - phrase musicale (son évoluant dans le temps, mono-source, mono-phonique)
  - accords (mono-source, poly-phonique)
  - musique (multi-source, poly-phonique)

Un concept plusieurs méthodes

1. Pour certains concepts (descripteurs audio), il existe plusieurs méthodes d'estimation
2. Pour d'autres concepts, le descripteur audio est intrinsèquement lié à une formulation mathématique

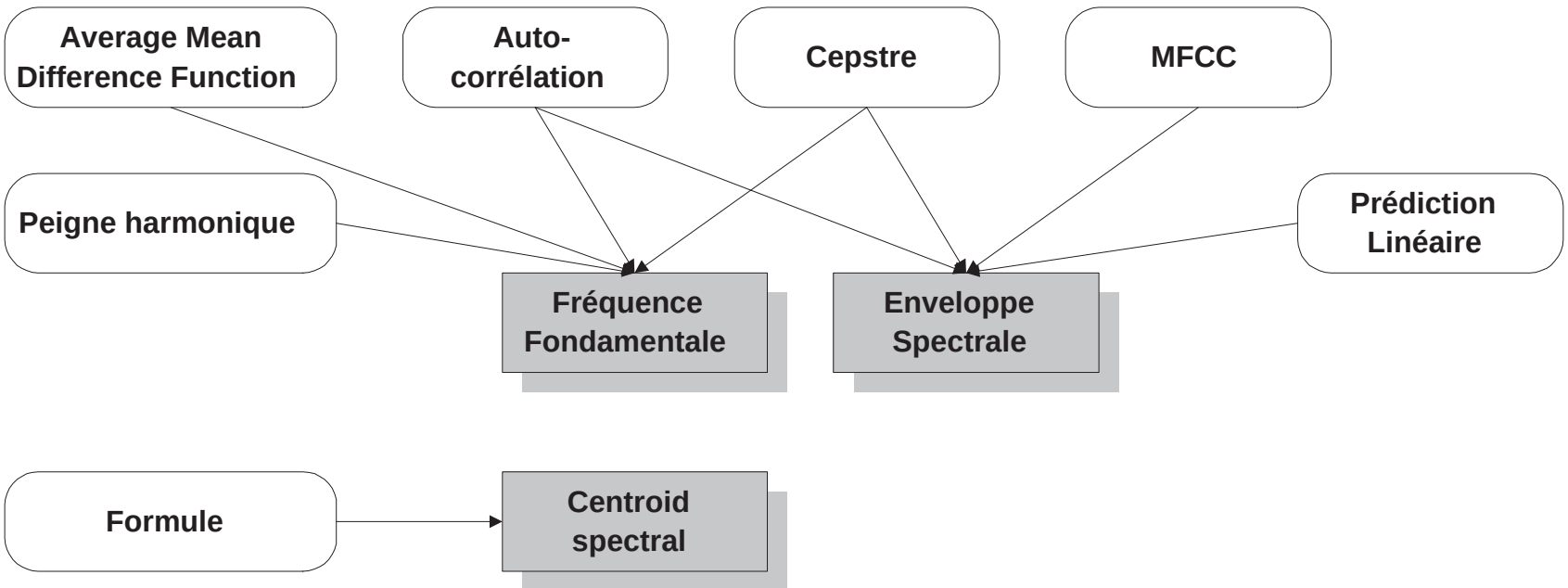


Table des descripteurs

	Echantillon	**Musique**
Bas-niveaux	MFCC SFM Chroma/PCP Centroid, Etendue, Pente, ... Sonie	MFCC SFM Chroma/PCP Centroid, Etendue, Pente, ... Sonie
Modèles temporels	Attack, Sustain, Release Tremolo	
Modèles de signal	Hauteur Vibrato Harmonicité Taux de bruit Formants/Résonnance	Hauteur-Multilples
Haut-niveaux (modèle musical)		Melodie/ Bass Suite d'accords Tonalité Tempo, Métrique Rythme Structure temporel

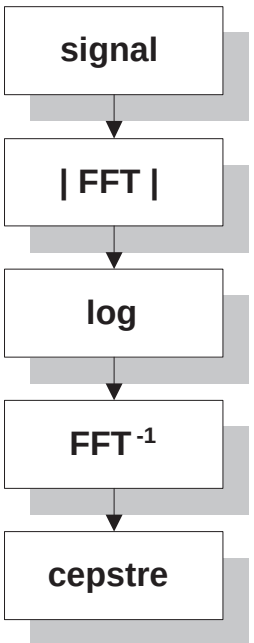
Le cepstre réel

- Cepstre ? Transformée de Fourier inverse du logarithme de la transformée de Fourier du signal

$$x(t) \rightarrow X(\omega) \rightarrow \log(X(\omega)) \rightarrow TF^{-1} \rightarrow cepstre \;\; c(l) \tag{11}$$

- Cepstre réel ? même chose sur le module de la TF
  $\log(|X(\omega)|)$
  $\log\left(A(\omega)e^{j\phi(\omega)}\right) = \log(A(\omega)) + j\phi(\omega)$
- A quoi ca sert ?

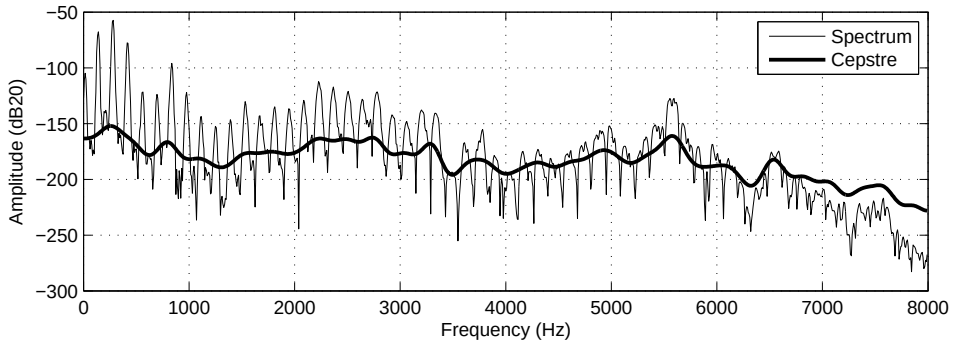
$$\begin{aligned} X(\omega) &= S(\omega)G(\omega) \\ log(X(\omega)) &= log(S(\omega)) + log(G(\omega)) \\ c(l) &= TF^{-1}(log(S(\omega))) + TF^{-1}(log(G(\omega))) \end{aligned} \tag{12}$$

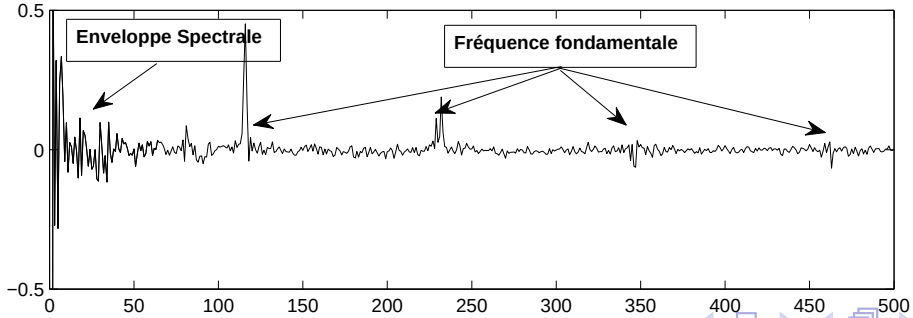


  - $S(\omega)$= spectre de la source (fréquence fondamentale), $\rightarrow$ $TF^{-1}(log(S(\omega)))$ varie rapidement à travers ω
  - $G(\omega)$ spectre du filtre (résonances/ anti-résonances), $\rightarrow$

Estimation fréquence fondamentale/ enveloppe spectrale à partir du cepstre

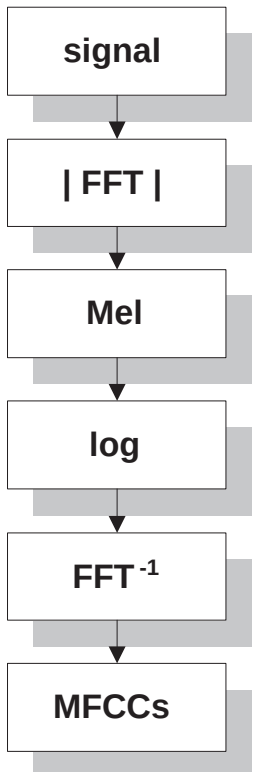
- Utilisation du cepstre $c(l)$? Si on prend la TF de $log(S(\omega)) + log(G(\omega))$ on va pouvoir
  - séparer ce qui varie rapidement (haute fréquence de la TF^{-1}): la fréquence fondamentale
  - séparer ce qui varie lentement (basse fréquence de la TF^{-1}): l'enveloppe spectrale





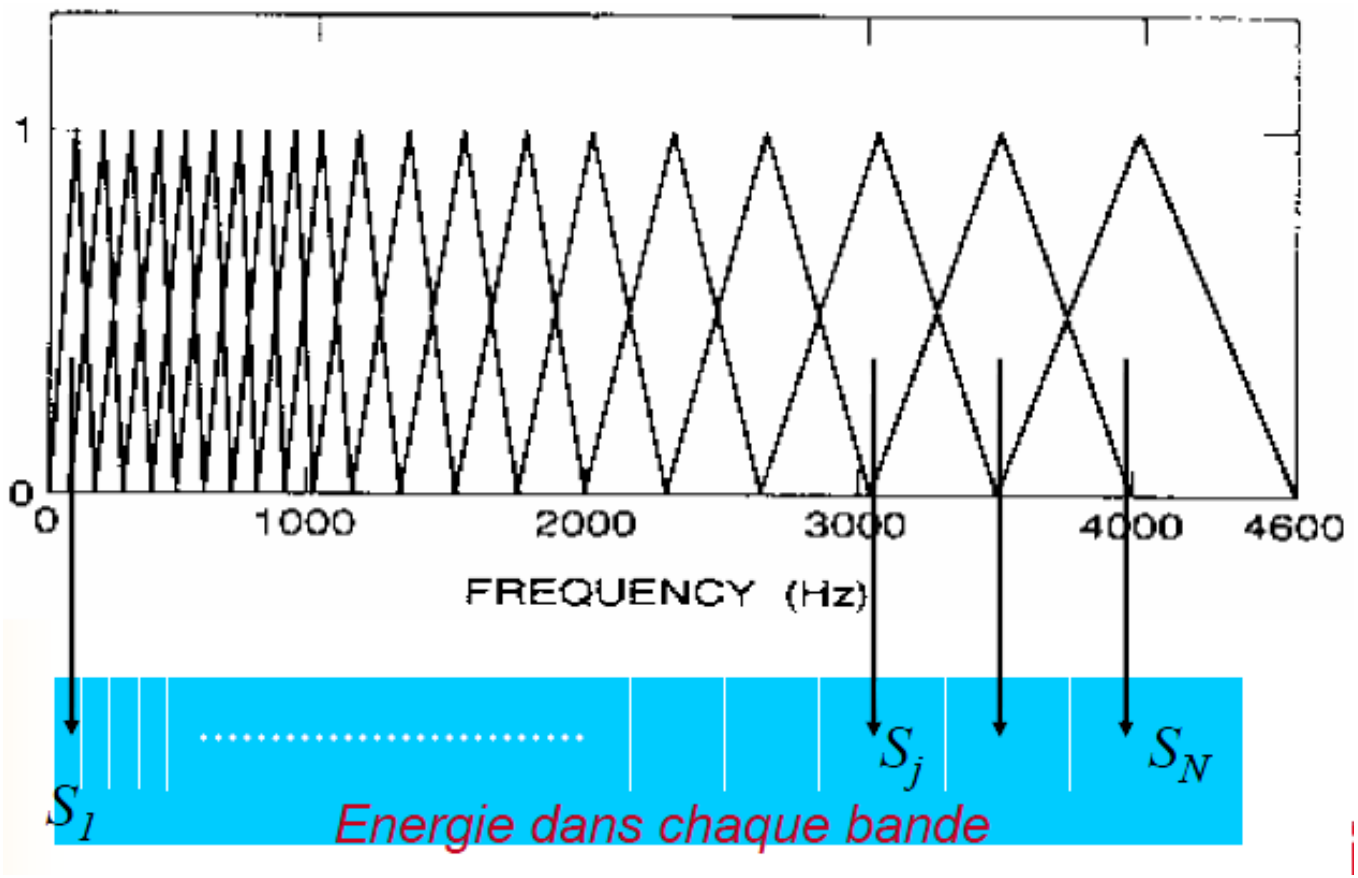
Les Mel Frequency Cepstral Coefficients (MFCCs)

- C'est quoi ? La même chose que le cepstre mais en convertissant le spectre $|X(\omega)|$ en échelle perceptive.
- Pouquoi ?
  - TFD: décomposition sur une série de sinusoides linéairement espacées (10Hz, 20Hz, 30Hz, ... Hz)
  - Oreille: décomposition sur une série de filtres de fréquences logarithmiquement espacé (10, 20, 40, 80, ... Hz).
  - Meilleure résolution en basses fréquences que en hautes fréquences.
  - Résonances de l'enveloppe spectrale sont plus rapprochées en basse fréquence.
- Comment ? On utilise des échelles dites perceptives: échelles de Mel, de Bark, filtres ERB.
- Utilisation ? Les coefficients les plus utilisés dans le monde de la reconnaissance audio: parole, musique, sons environnementaux, ...



$$x(t) \rightarrow X(\omega) \rightarrow X(mel) \rightarrow \log(X(mel)) \rightarrow TF^{-1} \rightarrow MFCCs$$

Banc de filtres



d'après Richard

Echelles perceptives

- Echelle de Mel

$$M = f \quad \text{pour} \quad f < 1000Hz$$

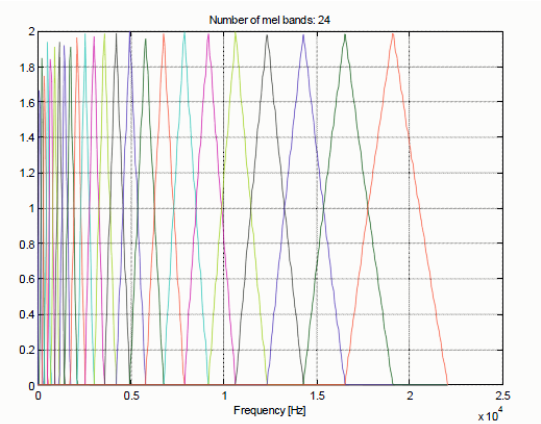
$$M = f_c \left(1 + \log_{10}\left(\frac{f}{f_c}\right)\right) \quad \text{pour} \quad f \geq 1000Hz \tag{14}$$

- Echelle de Bark

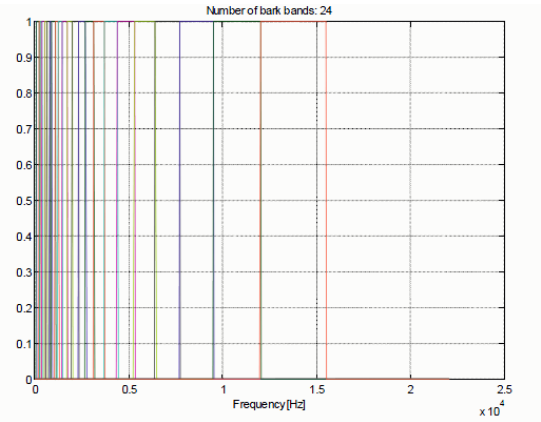
$$B = 13atan\left(\frac{f}{1315.8}\right) + 3.5atan\left(\frac{f}{7518}\right) \tag{15}$$

On somme alors l'énergie du spectre dans chaque bande de fréquence définit par les échelles

Banc de filtre Mel



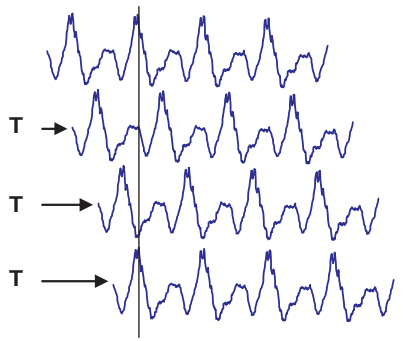
Banc de filtre Bark



L'auto-corrélation

- L'auto-corrélation ?
- Corrélation du signal $x(n)$ avec une version décalée de T

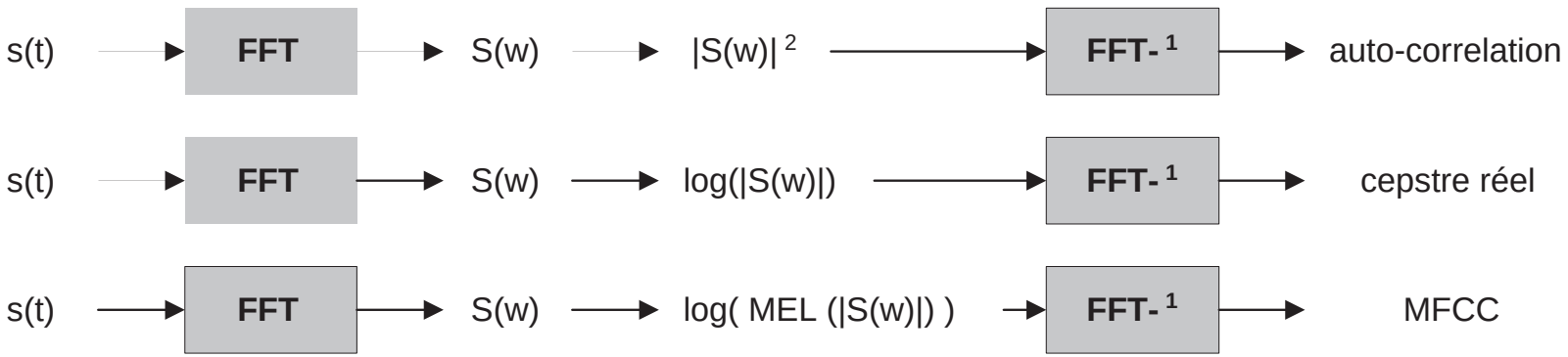
$$R(\tau) = \int_t x(t)x(t-\tau)dt$$
$$R(m) = \sum_{n=0}^{N-m-1} x(n)x(n-m) \quad (16)$$



Auto-corrélation = transformée de Fourier inverse du spectre de puissance $|X(\omega)^2$

$$R(\tau) = \int_t x(t)x(t-\tau)dt = \int_\omega S^2(\omega)e^{j\omega\tau}d\omega \quad (17)$$

Rapprochement



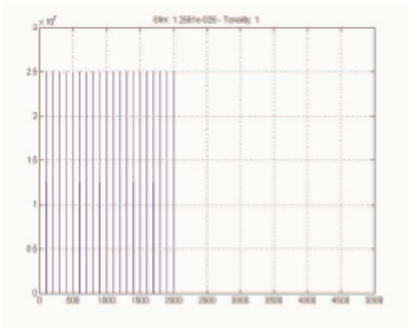
La Platitude Spectrale

- La platitude spectrale (spectral flatness) mesure le taux de composantes sinusoidales/ bruit dans chaque bande de fréquence B
- Méthode: comparaison de la moyenne géométrique à la moyenne arithmétique de spectre dans une bande de fréquence B

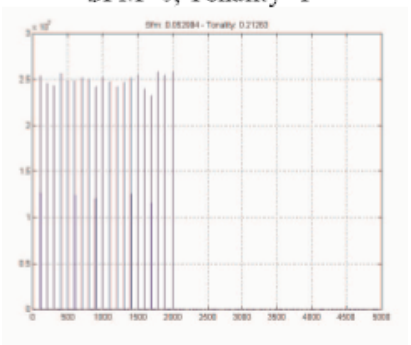
$$SFM(B) = \frac{(\Pi_{k \in B} A(k))^{1/K}}{\frac{1}{K}\left(\sum_{k \in B} A(k)\right)} \qquad (18)$$

Le SFM

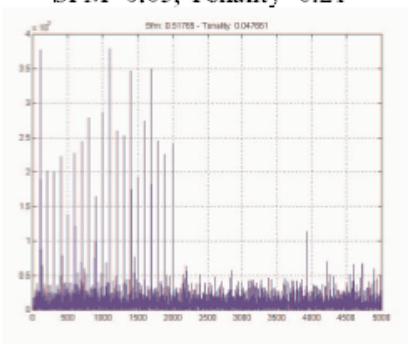
- prend des valeurs faibles quand une bande de fréquence B a beaucoup de "peaks", i.e. elle contient une quantité importante de composante sinusoidale
- prend des valeurs élevées quand une bande de fréquence B est plate, i.e. elle contient une quantité importante de bruit



SFM=0, Tonality=1



SFM=0.05, Tonality=0.21

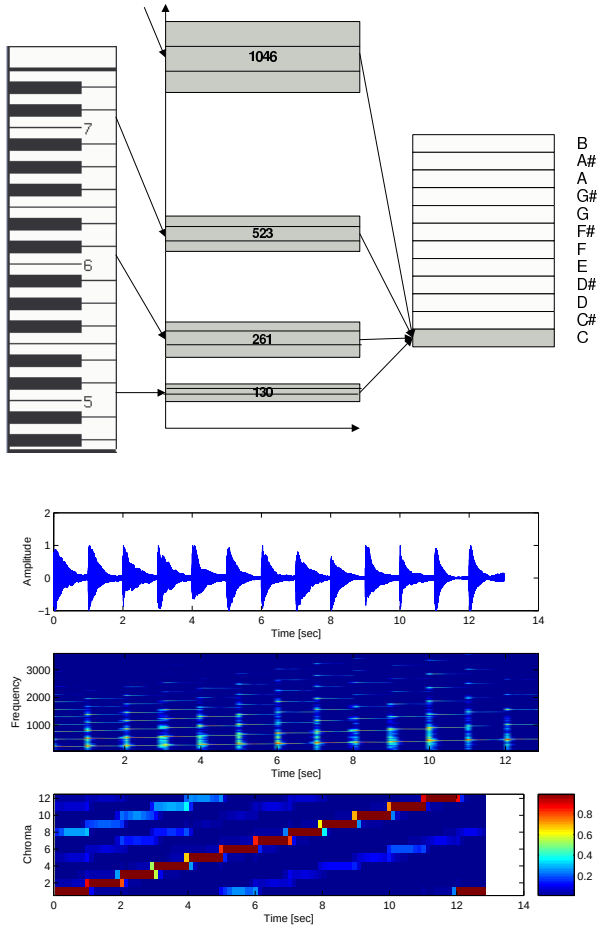


SFM=0.51, Tonality=0.047

Pitch Class Profile (PCP)/ Chroma

- **Objectif:** Représenter sous forme d'un vecteur le contenu harmonique du signal à un instant donné
- Très utilisé dans les applications de type
  - reconnaissance automatique de tonalité, de suite d'accords, de "cover versions"
- Shepard 1964: représenter la hauteur comme une structure bi-dimensionelles:
  - la hauteur tonale (numéro d'octave),
  - le chroma (classe de hauteur/ pitch class).
- Chroma spectrum/ Pitch Class Profile (PCP):
  - mapping entre les valeurs de la transformée de Fourier et les 12 classes de hauteurs de demi-tons C
- Mapping entre les fréquences f_k de la transformée de Fourier et l'échelle des hauteurs de demi-tons n (en échelle de notes MIDI):

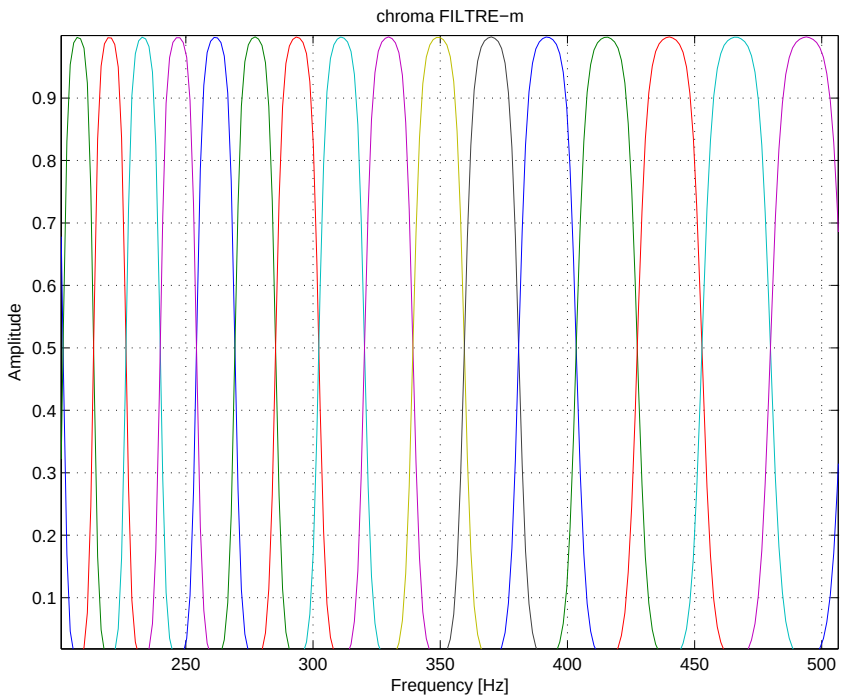
$$n(f_k) = 12 \log_2 \left(\frac{f_k}{440} \right) + 69 \quad n \in \mathbb{R}^+ \qquad (19)$$



Pitch Class Profile (PCP)/ Chroma

Méthode:

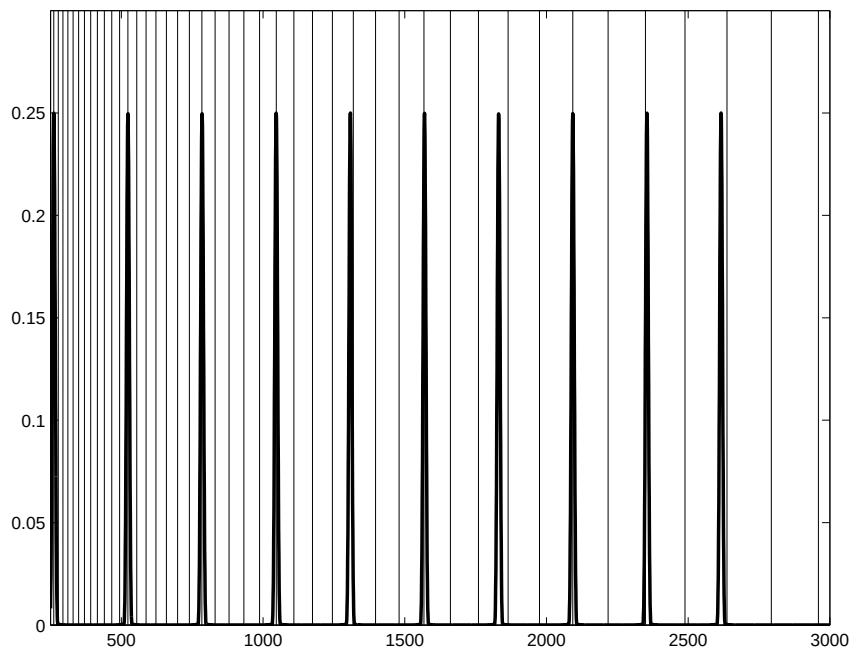
- Création d'un banc de filtres $H_n(\omega)$ centrés sur les hauteurs de demi-tons n
- La valeur du spectre de hauteur de demi-ton est obtenue en multipliant les valeurs de la transformée de Fourier $A(\omega)$ par l'ensemble des filtres $H_n(\omega)$
- On regroupe (additionne) ensuite les valeurs de filtres représentant les même chroma:
  - $C = C_1 + C_2 + C_3 + ...$
  - $C\# = C\#_1 + C\#_2 + C\#_3 + ...$



Pitch Class Profile (PCP)/ Chroma

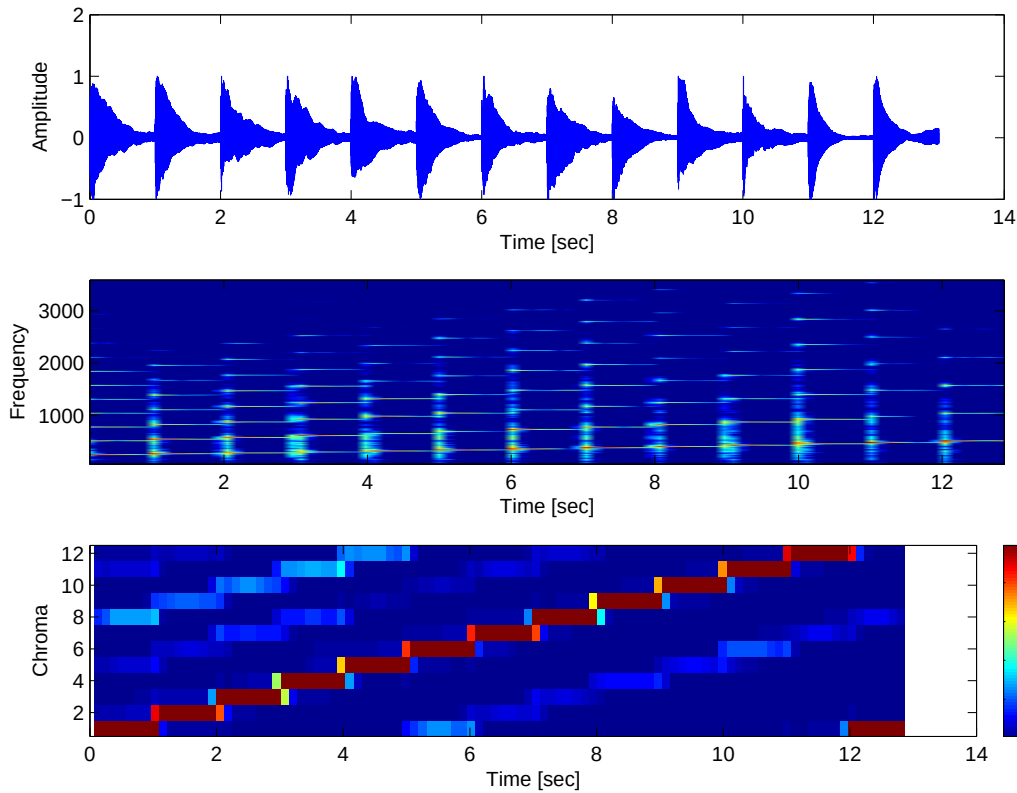
- Limitation de l'approche Pitch Class Profile/ Chroma: présence des harmoniques supérieures de chaque note
- Gamme parfaitement tempérée, gamme de Zarwin

numéro de l' harmonique	frequence	note la plus proche	interval
1	**261,6256**	**C4**	**0**
2	**523,2511**	**C5**	**0**
3	784,8766	G5	7 (7,01)
4	**1046,5021**	**C6**	**0**
5	1308,1276	E6	4 (3,86)
6	1569,7531	G6	7 (7,01)
7	*1831,3786*	*?*	*9,68*
8	**2093,0041**	**C7**	**0**
9	2354,6296	D7	2 (2,03)
10	2616,2551	E7	4 (3,86)
11	*2877,8806*	*?*	*5,51*

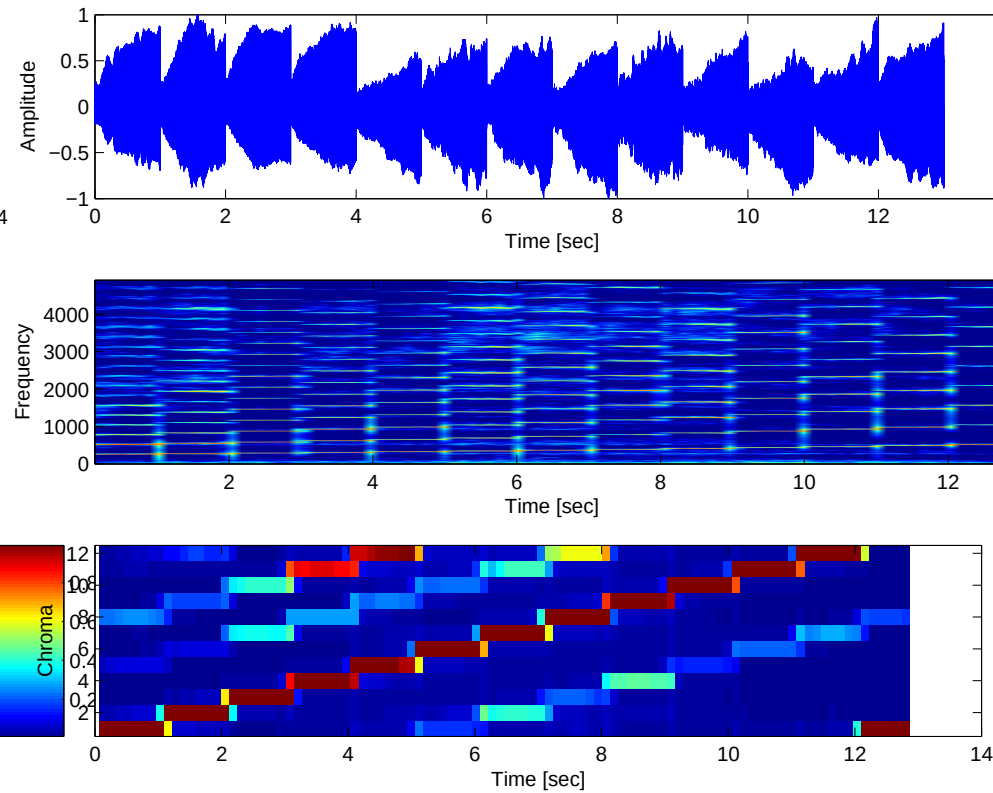


Pitch Class Profile (PCP)/ Chroma

Piano



Violon



Descripteurs audio temporels

- Le taux de passages par zéro (Zero Crossing Rate)

$$zcr = 0.5 \sum_{n=1}^{N} |sign(x(n)) - sign(x(n-1))| \quad (20)$$

Le zcr
  - prend des valeurs élevées quand le signal est bruité,
  - prend des valeurs faibles quand le signal est déterministe (sinusoide, harmonicité)

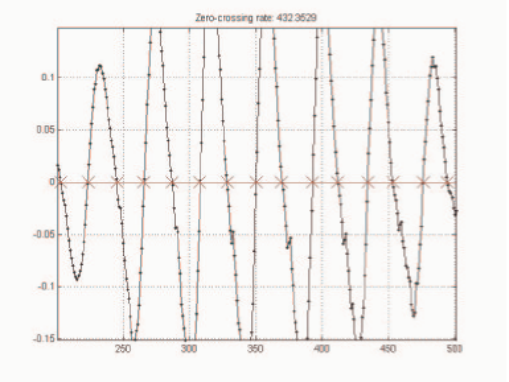


Figure 12 Zero-crossing rate (=432) during voiced speech region

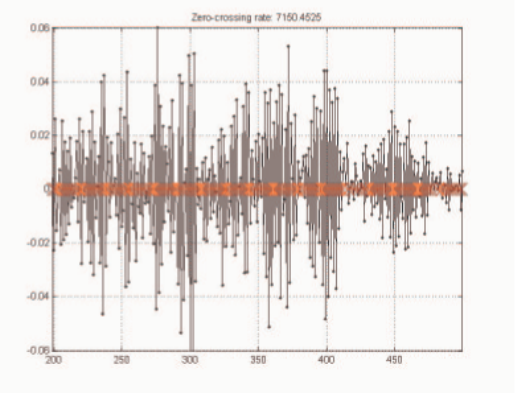


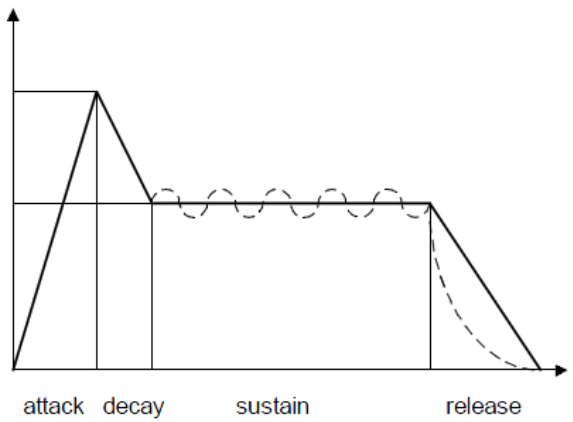
Figure 13 Zero-crossing rate (=7150) during unvoiced speech region

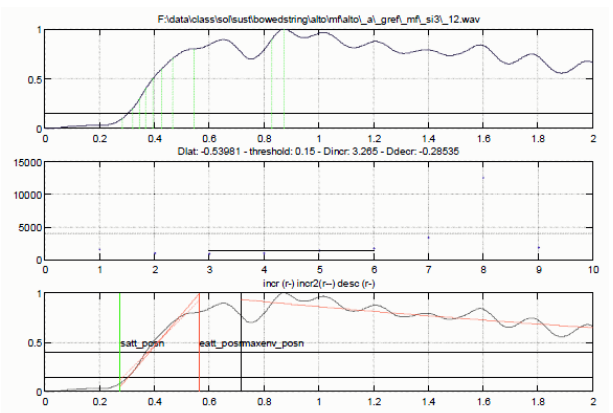
Descripteurs audio temporels

Modélisation de l'enveloppe temporelle: modèle Attack/ Decay/ Sustain/ Release (ADSR)

- Caractérisation du temps d'attaque: log-attack time
- Centroide temporelle
- Durée effective
- Taux de croissance
- Taux de décroissance
- Modulation d'énergie sur la partie tenue: tremolo

Egalement: modulation à 4 Hz de l'enveloppe d'énergie (parole/ musique)

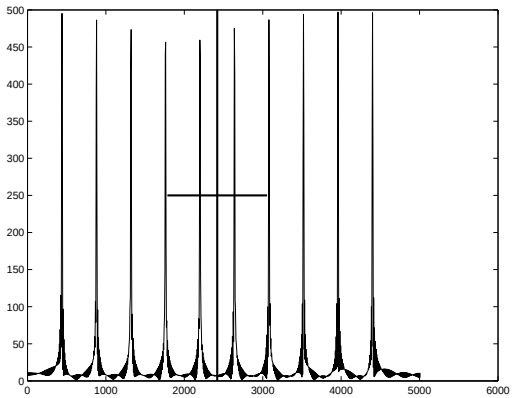




Descripteurs audio spectraux

- Le centroid spectral (spectral centroid)

$$cgs = \frac{\sum_{k=1}^{N} f_k A(f_k)}{\sum_{k=1}^{N} A(f_k)} \quad (21)$$

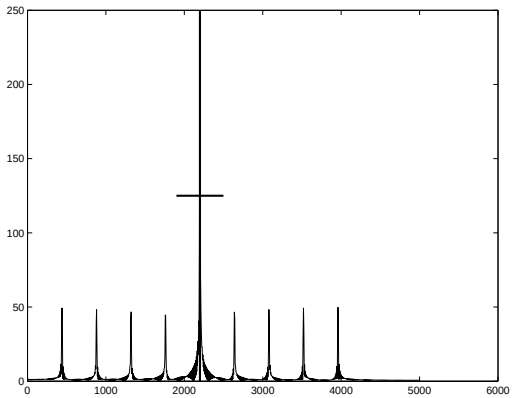


Le cgs

- prend des valeurs élevées pour les sons brillants (trompette, klaxon, ...)
- prend des valeurs faibles pour les sons ternes (bassons, grosse caisse, ...)

- L'étendue spectrale (spectral spread)

$$std = \sqrt{\frac{\sum_{k=1}^{N} (f_k - cgs)^2 A(f_k)}{\sum_{k=1}^{N} A(f_k)}} \quad (22)$$



Le std

- prend des valeurs élevées quand le signal a un spectre étendu (très riche, une trompette),
- prend des valeurs faibles quand le signal a un spectre étroit (peu riche, une flute),

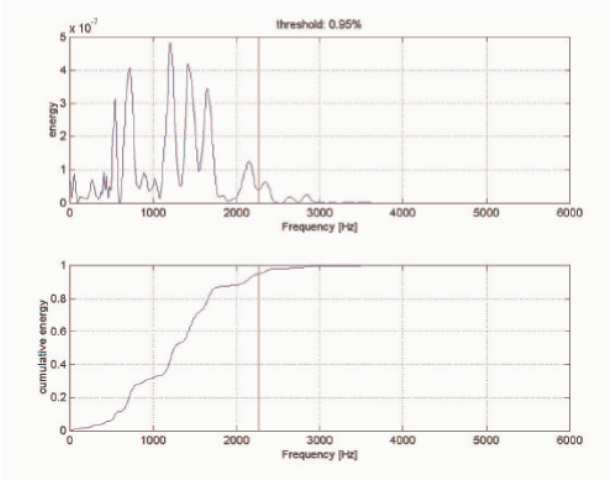
Descripteurs audio spectraux

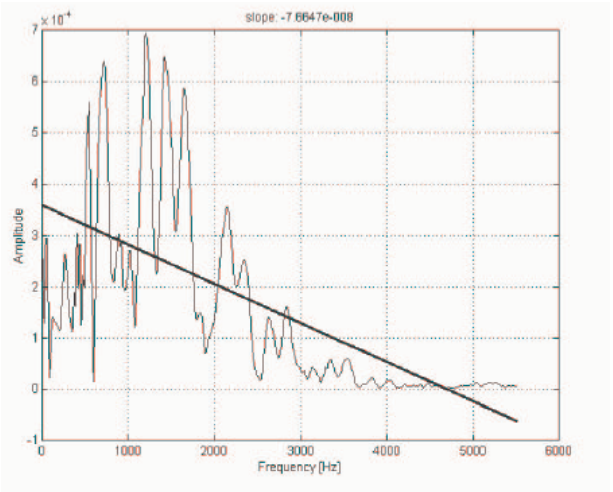
- La coupure spectrale (spectral roll-off)
  La fréquence F telle que 95% de l'énergie du spectre est en-dessous

$$\sum_{f=0}^{F} A^2(f) = 0.95 \sum_{f=0}^{sr/2} A^2(f) \qquad (23)$$

- La pente spectrale (spectral slope)
  La pente spectrale représente le taux de décroissance en fréquence du spectre d'amplitude

$$\hat{A}(f) = slope \cdot f + const \qquad (24)$$





Descripteurs audio harmoniques

Descripteurs issus de la modélisation sinusoidale harmonique

- Fréquence fondamentale

$$f_0 \quad (25)$$

- Noisiness (exemple: flûte)

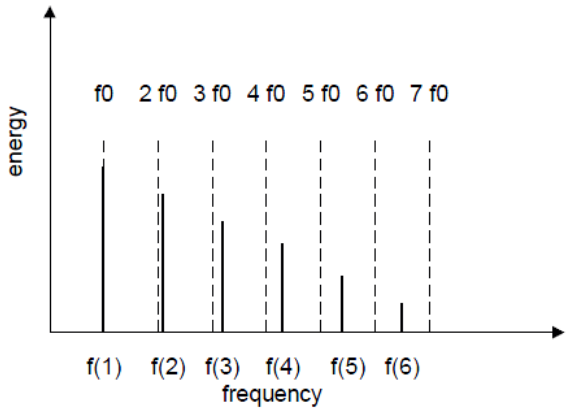
$$noisiness = \frac{ener\ bruit}{ener\ totale} \quad (26)$$

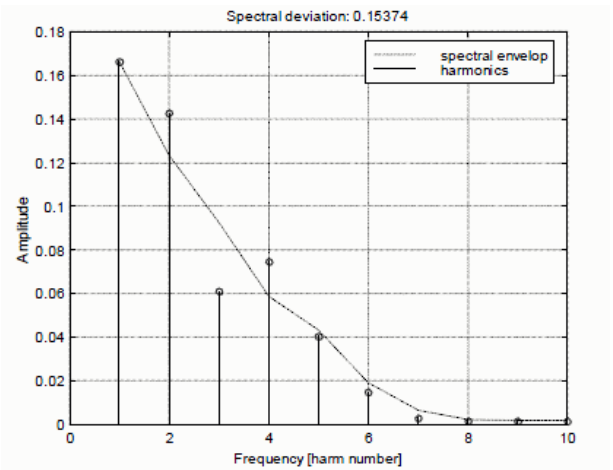
- In-harmonicité (exemple: piano, sitar)

$$inharmo = \frac{2}{f_0} \frac{\sum_h |f(h) - hf_0| a^2(h)}{\sum_h a^2(h)} \quad (27)$$

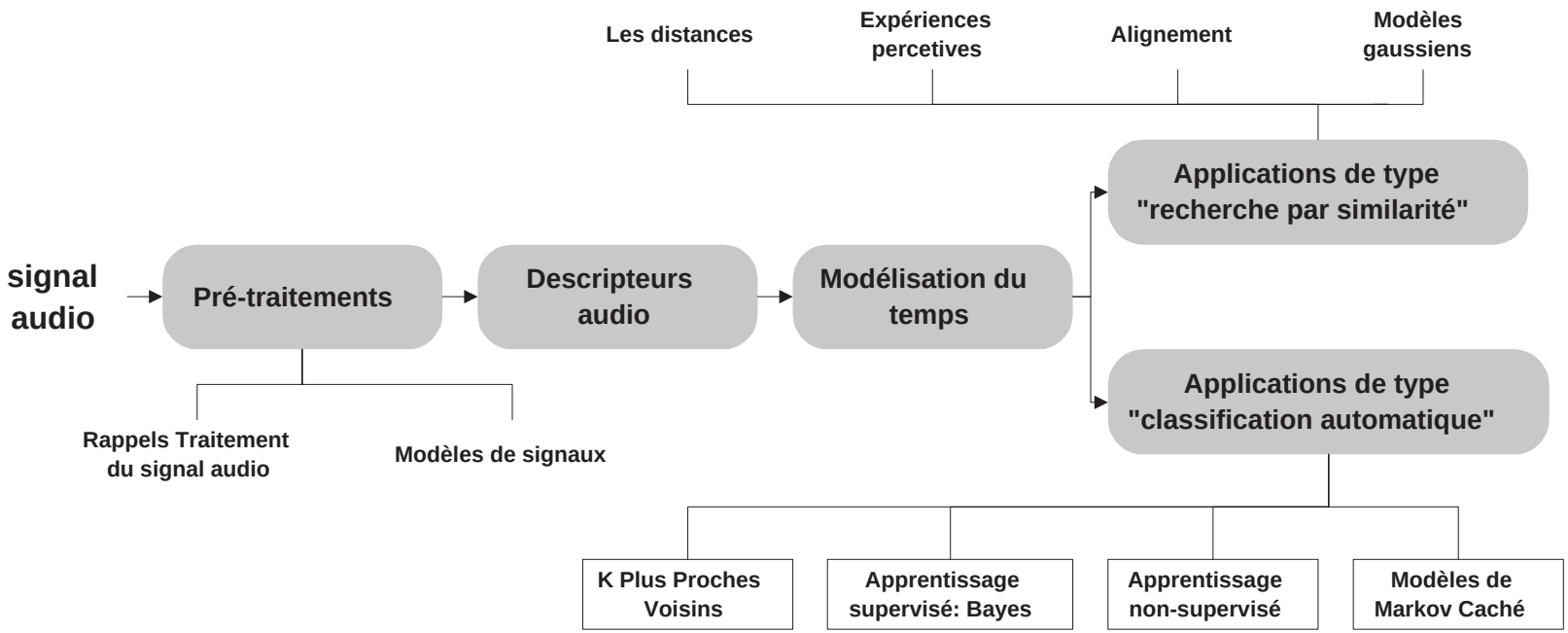
- Déviation spectrale (exemple: clarinette)

$$devsepc = \frac{1}{H} \sum_h (a(h) - SE(h)) \quad (28)$$





Plans


Les distances
Expériences percetives
Alignement
Modèles gaussiens
Applications de type "recherche par similarité"
signal audio
Pré-traitements
Descripteurs audio
Modélisation du temps
Applications de type "classification automatique"
Rappels Traitement du signal audio
Modèles de signaux
K Plus Proches Voisins
Apprentissage supervisé: Bayes
Apprentissage non-supervisé
Modèles de Markov Caché

Modélisation de l'évolution temporelle des descripteurs

- La plupart des descripteurs audio représentent le comportement du signal autour d'un temps m (la trame d'analyse de la TFCT) et sur une durée de 40ms (la durée de la fenêtre de pondération de la TFCT)
- On note $d(m)$ la valeur du descripteur extrait à la trame m
- Le son est rarement stationnaire $\rightarrow$ il faut représenter les temps $\rightarrow$ Difficile pour un ordinateur !
- Trois grandes tendances
  1. créer des descripteurs audio de plus haut-niveau qui représentent l'évolution temporelle
     - A. Les Delta descripteurs
     - B. Les "Texture windows"
  2. modéliser le temps par des modèles statistiques (globale: GM, GMM ou locale HMM) $\rightarrow$ pour plus tard

A. Les Delta descripteurs

- Delta (vitesse) Pour un descripteurs, $d(m)$ on calcule la dérivée de sa trajectoire temporelle: $\frac{\delta d(t)}{\delta t}$

$$\Delta d(m) = d(m) - d(m-1) \tag{29}$$

  Pour plus de précisions, on prend la dérivée de l'approximation polynomiale de sa trajectoire locale.
- Les Delta-Delta descripteurs (accélération): Pour un descripteurs, $d(m)$ on calcule la dérivée seconde de sa trajectoire temporelle: $\frac{\delta^2 d(t)}{\delta^2 t}$

$$\Delta\Delta d(m) = \Delta d(m) - \Delta d(m-1) \tag{30}$$

  Pour plus de précisions, on prend la dérivée seconde de l'approximation polynomiale de sa trajectoire locale.
- très utilisé en traitement de la parole (reconnaissance de la parole)

B. Les "Texture windows"

- Les "texture windows": on modélise le comportement temporel des descripteurs $d(m)$ sur une certaine durée M
  - durée M: l'ensemble du fichier
  - durée M: un segment d'une durée autour de 500ms

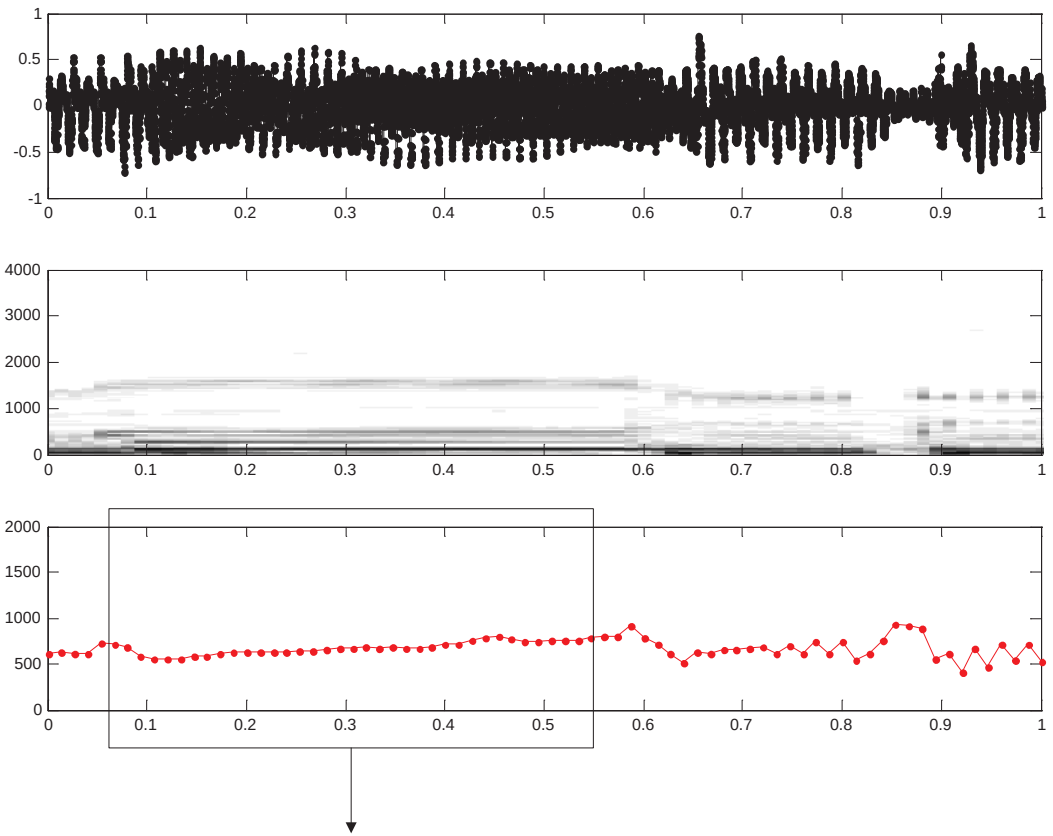
  Comment ?
  - deux premiers moments statistiques (moyenne et écart-type) d'un descripteur. On considère les valeurs du descripteur comme des "observations".

$$\mu(d(m')) = \frac{1}{M} \sum_{m=m'-M/2}^{m'+M/2} d(m) \tag{31}$$

$$\sigma^2(d(m')) = \frac{1}{M} \sum_{m=m'-M/2}^{m'+M/2} (d(m) - \mu(d(m'))^2 \tag{32}$$

  - modèles auto-régressifs appliqués au trajet temporel du signal $d(m)$,
  - amplitude de la TFCT de $d(m)$ considéré comme un signal temporel

Modélisation de l'évolution temporelle des descripteurs



Signal audio échantilloné
Distance entre deux points= 1/44100

TFCT
Distance entre deux FFT: 20ms
Une FFT représente: 40ms

Centroid Spectral
Distance entre deux valeurs: 20ms
Une valeur représente: 40ms

moyenne et variance du descriptor sur une durée de 500ms