

M2 Data Science

DS-telecom-15 "Audio and Music Information Retrieval"



Geoffroy Peeters

contact: geoffroy.peeters@telecom-paris.fr

Télécom-Paris, IP-Paris, France

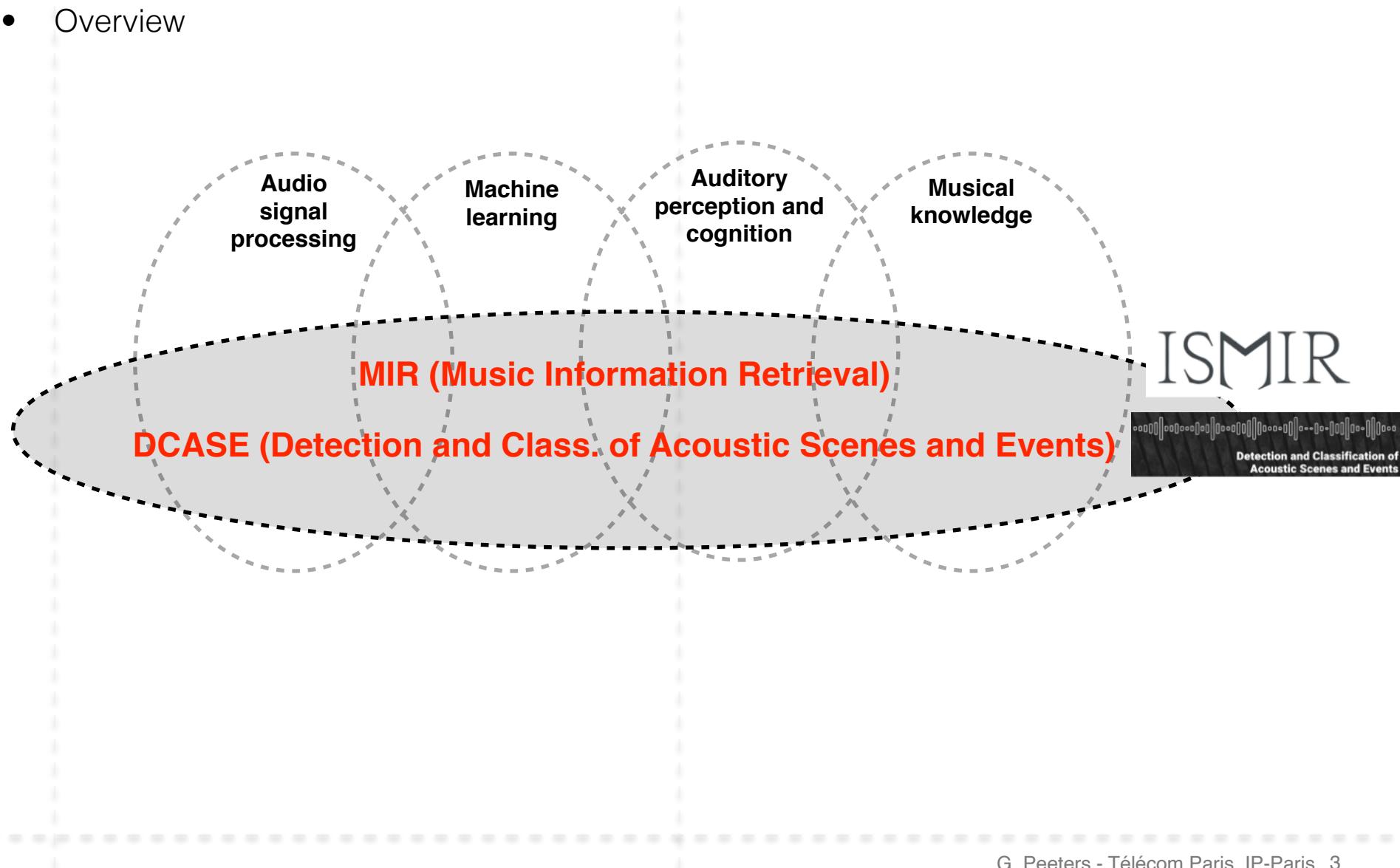
DS-telecom-15 "Audio and Music Information Retrieval"

| Slot | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
|----------|------------------|------------------|-------------------|--------------------|------------------|------------------|------------------|------------------|---------------------|------------|
| Date | 2022/01/18 | 2022-01-25 | 2022/02/01 | 2022/02/08 | 2022/02/15 | 2022/03/01 | 2022/03/08 | 2022/03/15 | 2022/03/22 | 2022/03/29 |
| Time | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | 14:00 - 18:00 | |
| Location | Télécom, Amphi 7 | Télécom, Amphi 7 | Télécom, Amphi 7 | X, salle 67 | Télécom, Amphi 7 | |
| Speaker | G. Peeters | G. Peeters | G. Peeters | G. Peeters | G. Peeters | G. Richard | G. Richard | G. Richard | G. Peeters, G. Rich | Exam |
| A | TFCT | Musique Chromas | Musique Structure | Deep Learning Clas | Deep Learning Me | Multi-Pitch | Sinusoidal | DCASE | Conférence | |
| B | Lab TFCT | Lab Chord | Lab Structure | Lab Class | Lab Similarity | Lab Multi-Pitch | Lab Shazam | Lab DCASE | Conférence | |

- **Evaluation:** 30% labs/project, 70% written exam
- **Site:** <https://moodle.polytechnique.fr/course/view.php?id=13194>

Audio and Music Information Retrieval is a multi-domain field

- Overview



Various categories of audio content and related specificities

- Categories of sound

- Speech
- Music
- Scene (Environmental sounds)

Image

Speech (object)



Audio



Mostly **single** speaker

A **structured** sound source

Music (artwork)



Multi timbre (texture) **polyphonic**

Structure of sound sources

Dynamic control (mixing)

Scene



Multi source **polyphonic**

Unstructured sound sources

source : S. Dieleman, J. Pons, J. Lee, Waveform-based music processing with deep learning, tutoriel, ISMIR, 2019

Various categories of audio content and related applications

Signal Processing

Representations: Fourier transform, Constant-Q-Transform

Signal models: sinusoidal model, source/filter model

Description: Audio features (MFCC, Chroma)

Transformation: Phase vocoder, P-sola

Separation: NMF, HPSS, REPET



Speech

Source separation,
denoising,
processing



Music

Source separation,
denoising,
processing



Environmental Sounds

Source separation,
denoising,
processing

Transformation

Speech coding
Speech transformation

Description

Speaker recognition
Speaker diarization
Speech to text (ASR)

AudioID (Shazam)
Auto-tagging (genre, mood, instrum.)
Pitch estimation
Tempo/Chord estimation
Cover Detection
Structure estimation

Acoustic Scene Classification
Sound Event Localization and Detection

Generation

Text to speech

Sound generation (FM Synthesis, ...)
Music generation

Sound Texture Synthesis

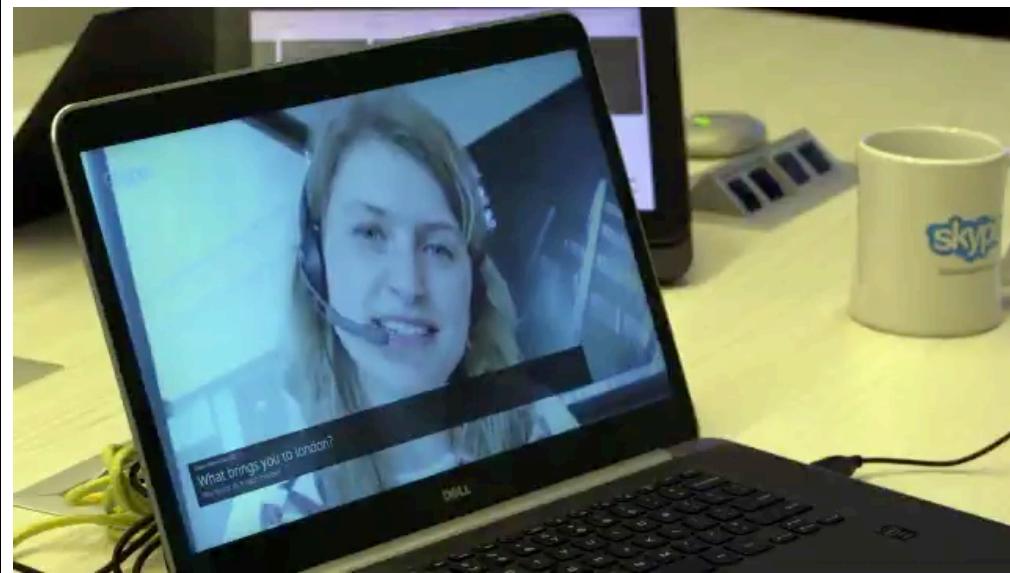
MIR, ISMIR
<http://ismir.net>

DCAESE
<http://dcase.community>

Various categories of audio content and related applications

- **Speech**

- Microsoft : https://www.youtube.com/watch?time_continue=544&v=Nu-nlQqFCKg
- Skype : <https://www.youtube.com/watch?v=JrlTzS7Fk6o>



- **Music**
 - Analysis: <https://chordify.net/>
 - Generation: <https://magenta.tensorflow.org/> <https://openai.com/blog/musenet/>
-

- **Environmental sounds**

- Sonyc : <https://wp.nyu.edu/sonyc/>
- Birdvox : <https://wp.nyu.edu/birdvox/>



What is Music Information Retrieval ?

Music Information Retrieval

MIR

- <http://ismir.net>
- https://www.music-ir.org/mirex/wiki/MIREX_HOME



- **Music Information Retrieval:**

- interdisciplinary research field dedicated to the **understanding**, **processing** and **generation** of music
- combines theories, concepts, and techniques from music theory, computer science, signal processing perception, and cognition.
- deals with the development of algorithms for
 - **describing** the content of the music from the analysis of its audio signal.
 - Examples:
 - estimation of the various pitches, chords, rhythm,
 - identification of the instruments being used in a music,
 - assignment of “tags” to a music (such as genres, mood or usage)
 - recommending music from catalogues,
 - detection of cover/plagiarism in catalogue or in user-generated contents.
 - **processing** the content of the music.
 - Examples: enhancement, source separation
 - **generating** new audio signals or music pieces, or **transferring** properties from one signal to another

Music has various representations

Partition



Musique

MIDI

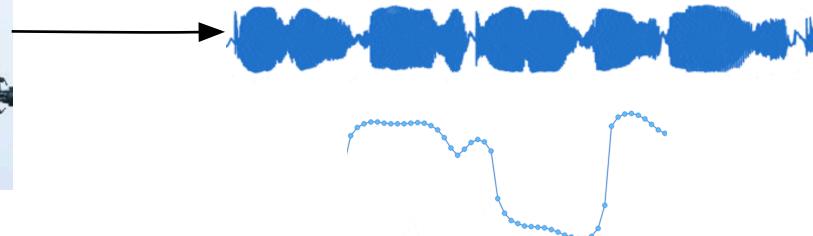
| Tick | Beat | Duration | Channel | Category | Type | Data |
|------|------|----------|---------|----------------|--------------------------------------|---------------------|
| 0 | 0 | 1:1 | 0:00 | Meta | MIDI port | Strings |
| 0 | 0 | 1:1 | 0:00 | Meta | Transmitter | |
| 0 | 1:1 | 0:00 | 4 Voice | Program Change | Ensemble Strings | Volume (coarse), 90 |
| 0 | 1:1 | 0:00 | 4 Voice | Control Change | Pan (coarse) | .89 |
| 0 | 1:1 | 0:00 | 4 Voice | Control Change | Bank Select (coarse) | , 0 |
| 12 | 1:1 | 0:00 | 4 Voice | Control Change | | |
| 36 | 1:1 | 0:00 | 4 Voice | Pitch Bend | 8192 | |
| 72 | 1:1 | 0:00 | 4 Voice | Control Change | Level Effect | .80 |
| 144 | 1:1 | 0:00 | 4 Voice | Control Change | Level Chorus | .15 |
| 96 | 1:1 | 0:00 | 4 Voice | Control Change | Registered Parameter Number (fine) | , 0 |
| 108 | 1:1 | 0:00 | 4 Voice | Control Change | Registered Parameter Number (coarse) | |
| 120 | 1:1 | 0:00 | 4 Voice | Control Change | Data Entry (coarse) | , 2 |
| 6912 | 10:1 | 0:31 | 4 Voice | Note On | D5, .79 | |
| 6912 | 10:1 | 0:31 | 4 Voice | Note On | D5, .79 | |
| 7104 | 10:2 | 0:32 | 4 Voice | Note On | F#5, .63 | |
| 7296 | 10:3 | 0:33 | 4 Voice | Note On | A#5, .71 | |
| 7684 | 10:4 | 0:35 | 4 Voice | Note On | F5, .0 | |
| 7680 | 11:1 | 0:35 | 4 Voice | Note On | D6, .67 | |
| 8064 | 11:3 | 0:37 | 4 Voice | Note On | A5, .84 | |
| 8064 | 11:3 | 0:37 | 4 Voice | Note On | C6, .107 | |
| 8078 | 11:3 | 0:37 | 4 Voice | Note On | A#5, .0 | |
| 8430 | 11:4 | 0:38 | 4 Voice | Note On | C6, .0 | |
| 8432 | 11:4 | 0:38 | 4 Voice | Note On | A5, .0 | |

Transcription

Enregistrement

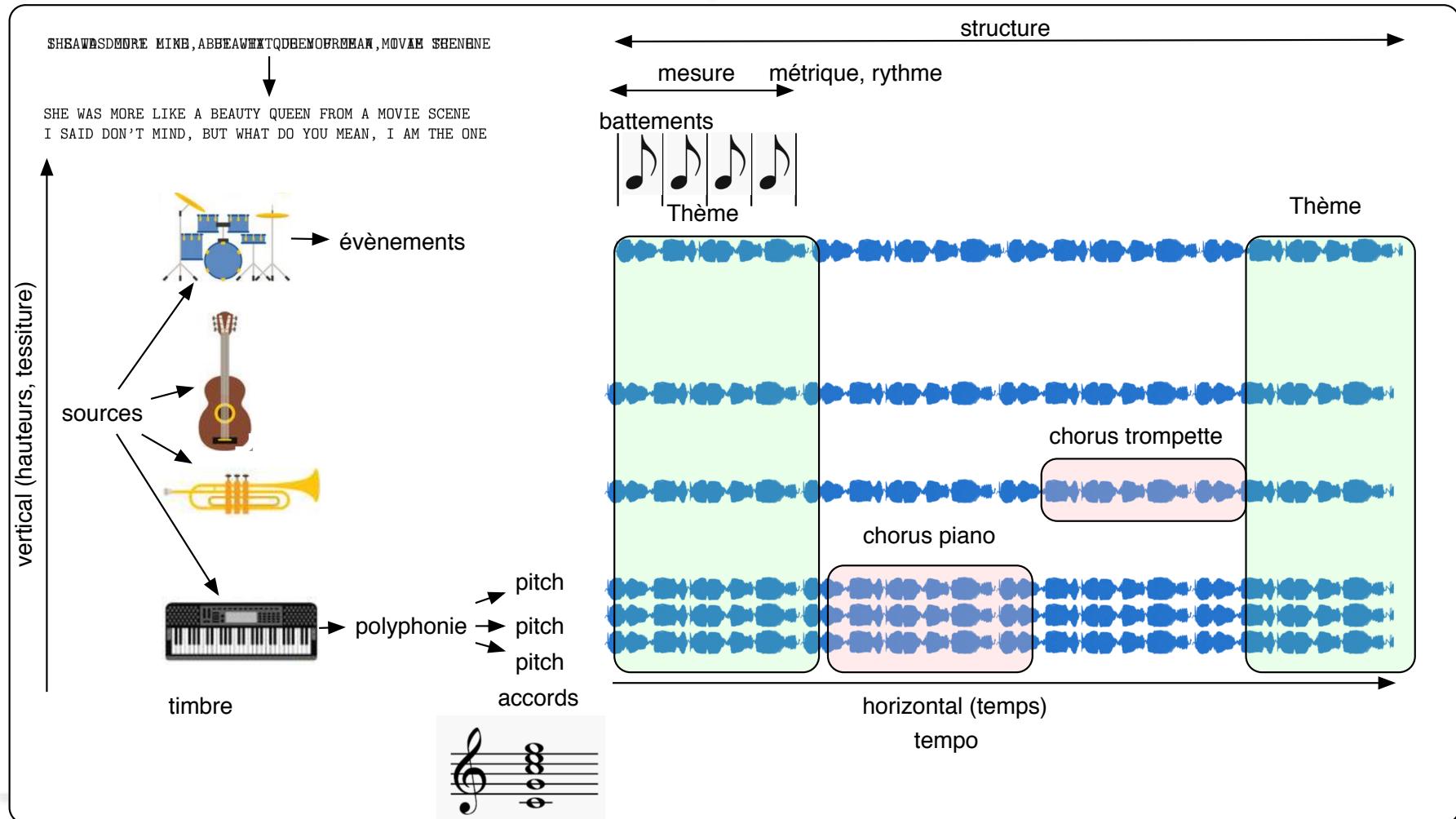


Audio

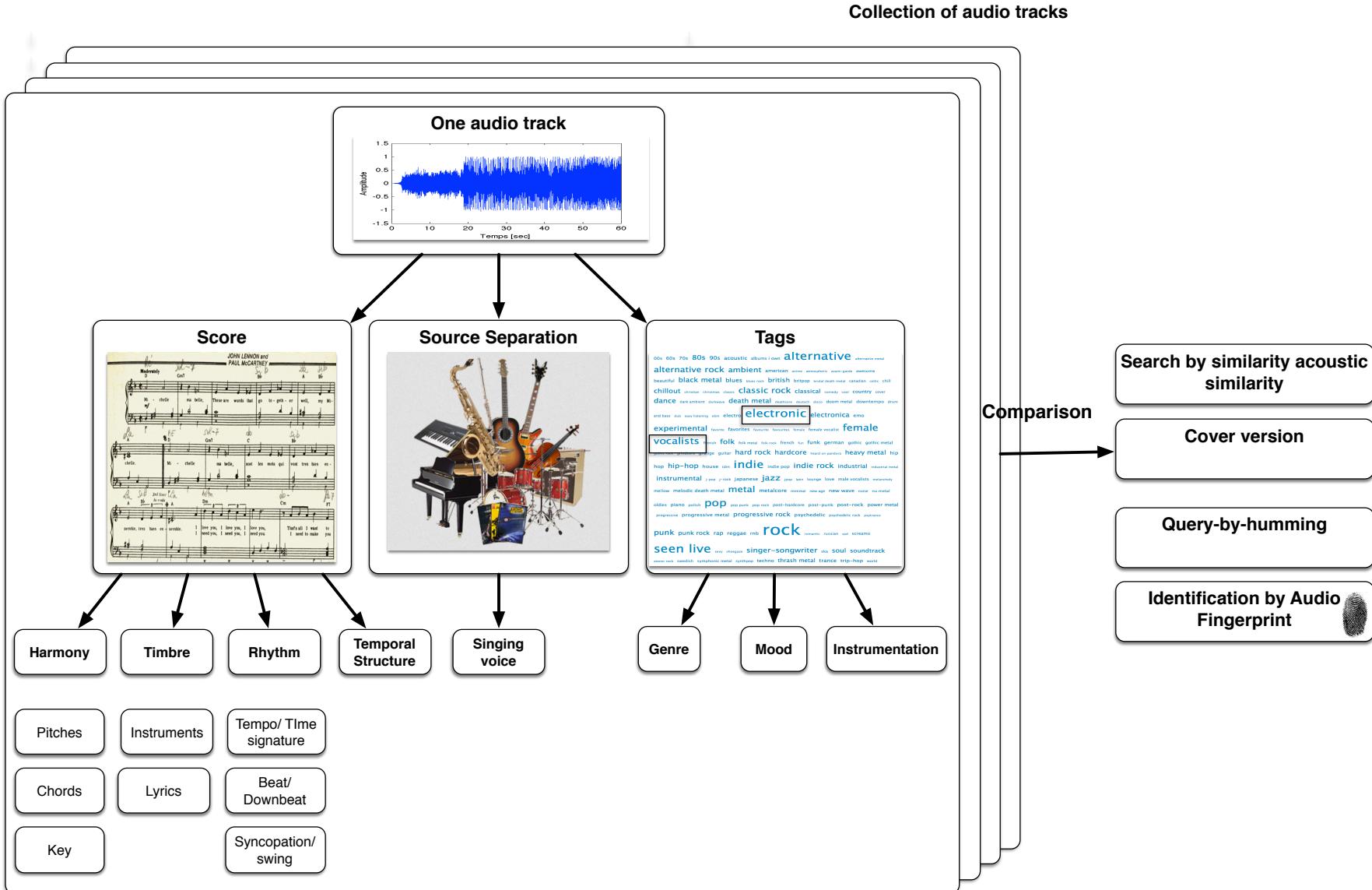


Music has both a vertical and horizontal organisation

- **Verticale** organization: sources, timbre, polyphony of pitches, chords, key
- **Horizontale** organization: structure, bar/measure, rhythm



Various types of music content that can be extracted from the audio



Music Information Retrieval: Real-world applications

Music Information Retrieval: Real-world applications

- **Automatic Music Transcription :**

- onset-detection
- pitch, multi-pitch, dominant melody
- chords/guitar tab, key
- rhythm : tempo, meter, beat, downbeat , musical instrument
- lyrics

- **Auto-tagging**

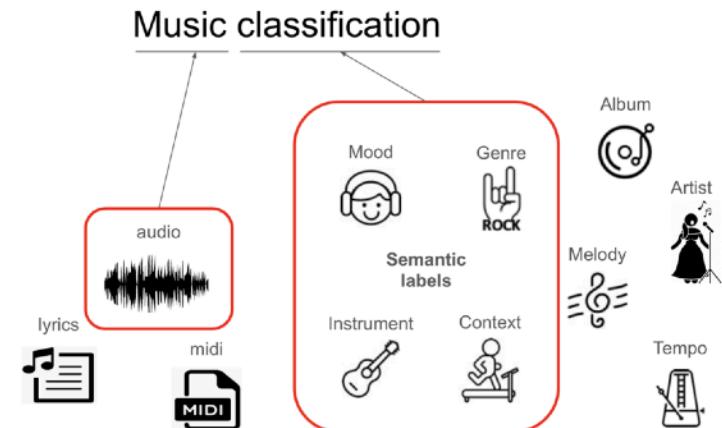
- genre, mood, context, ...

- **Music Recommendation**

- similarity, cover-detection

- **Source separation**

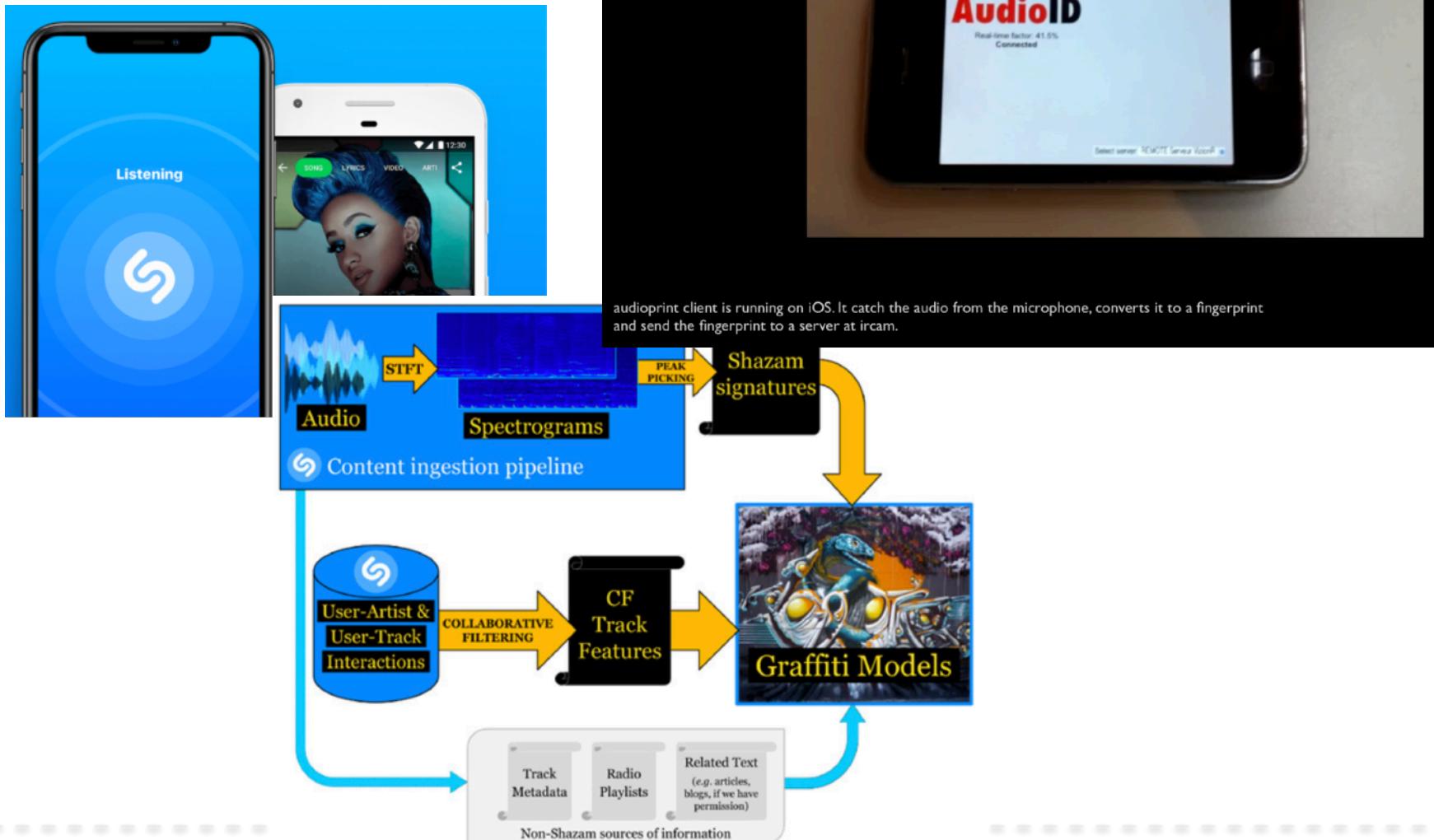
- Audio/ Music generation



Celma, Oscar, 2010 "Music recommendation", *Music recommendation and discovery*, Springer.

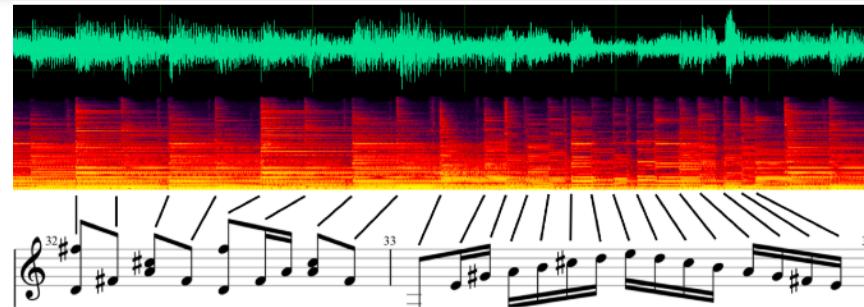
Music Information Retrieval: Real-world applications

- **Audio Identification**



Music Information Retrieval: Real-world applications

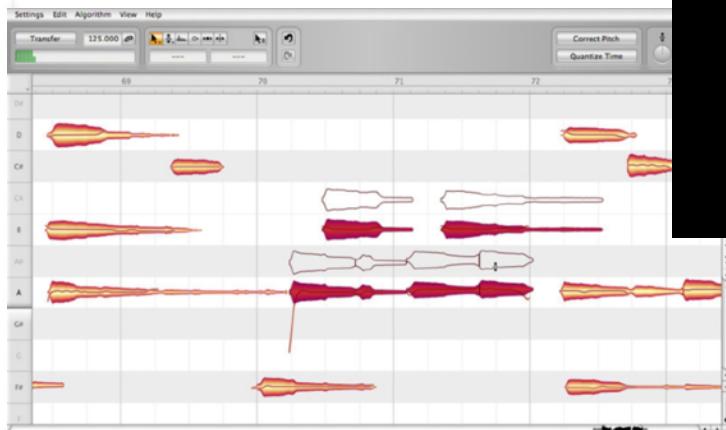
- **Music transcription**
- **Notes-manipulation**



Live Performance
Piano Transcription
with Onsets and Frames



g.co/magenta/onsets-frames



Music Information Retrieval: Real-world applications

- **Chord recognition**

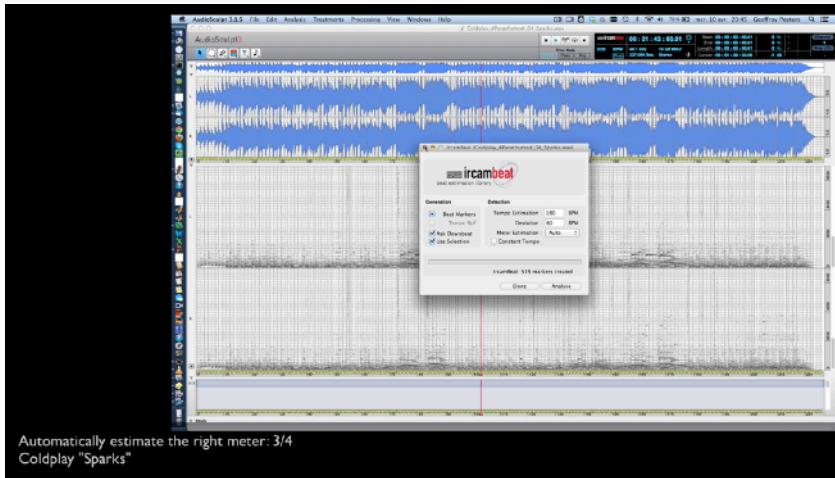
- Analysis <https://chordify.net/>

The screenshot shows the Chordify interface for the song "The Sound of Silence" by Simon & Garfunkel. At the top, there's a search bar with "RECHERCHER UNE CHANSON" and a file upload button "TÉLÉVÉRER UN FICHIER". Below the search bar are buttons for "DÉCOUVRIR", "CRÉER UN COMPTE", and "TARIFS". The main title "Simon & Garfunkel - The Sound of Silence (from The Concert in Central Park)" is displayed, along with a "Problème avec les accords ?" link. The navigation menu includes "GRILLES" (selected), "APERÇU", "AMÉLIORER", "TRANSPoser", "CAPO", "CHANSON", "ACCORDS" (selected), "TEMPO", "BOUCLE", "MIDI", and "IMPRIMER". A "SIMPPLIFIER LES ACCORDS" checkbox is also present.

The main content area shows a timeline with chords: Mi_m, Re, and Mi_m. Below the timeline are four chord diagrams for "GUITARE", "UKULÉLE", and "PIANO". The first diagram for Mi_m shows a 6-string guitar neck with three dots (1, 2, 3) indicating finger placement. The other three diagrams show ukulele and piano chords with similar dot patterns. The "ANIMÉ" and "APERÇU" buttons are located to the right of the piano diagram.

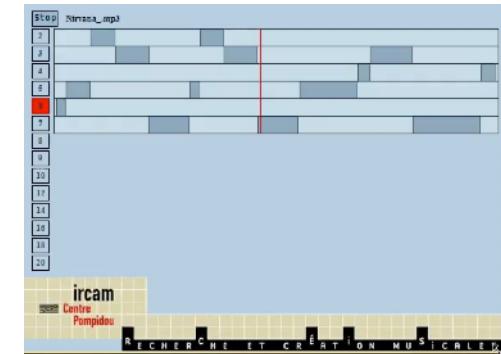
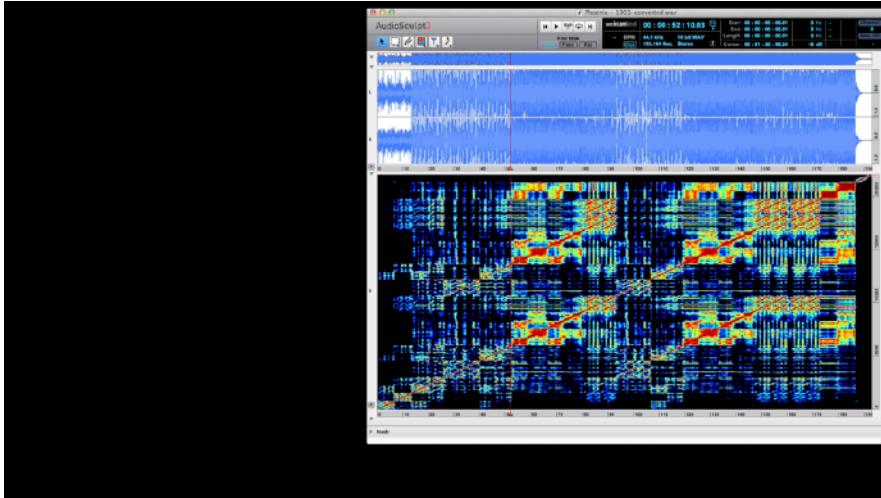
Music Information Retrieval: Real-world applications

- **Beat tracking**



- **Structure estimation**

- audio summary generation



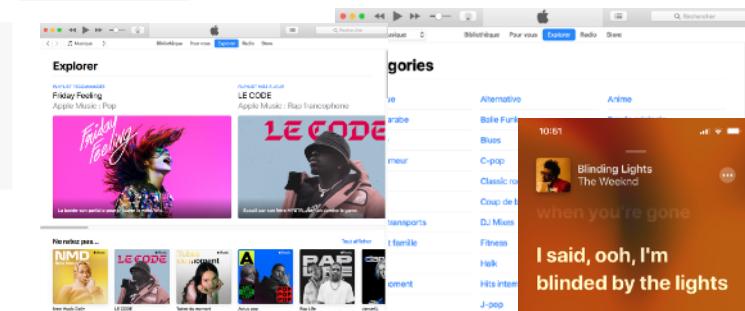
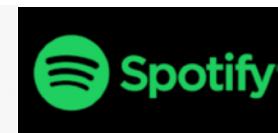
Music Information Retrieval: Real-world applications

- **Music search engine**

- **Streaming services:**

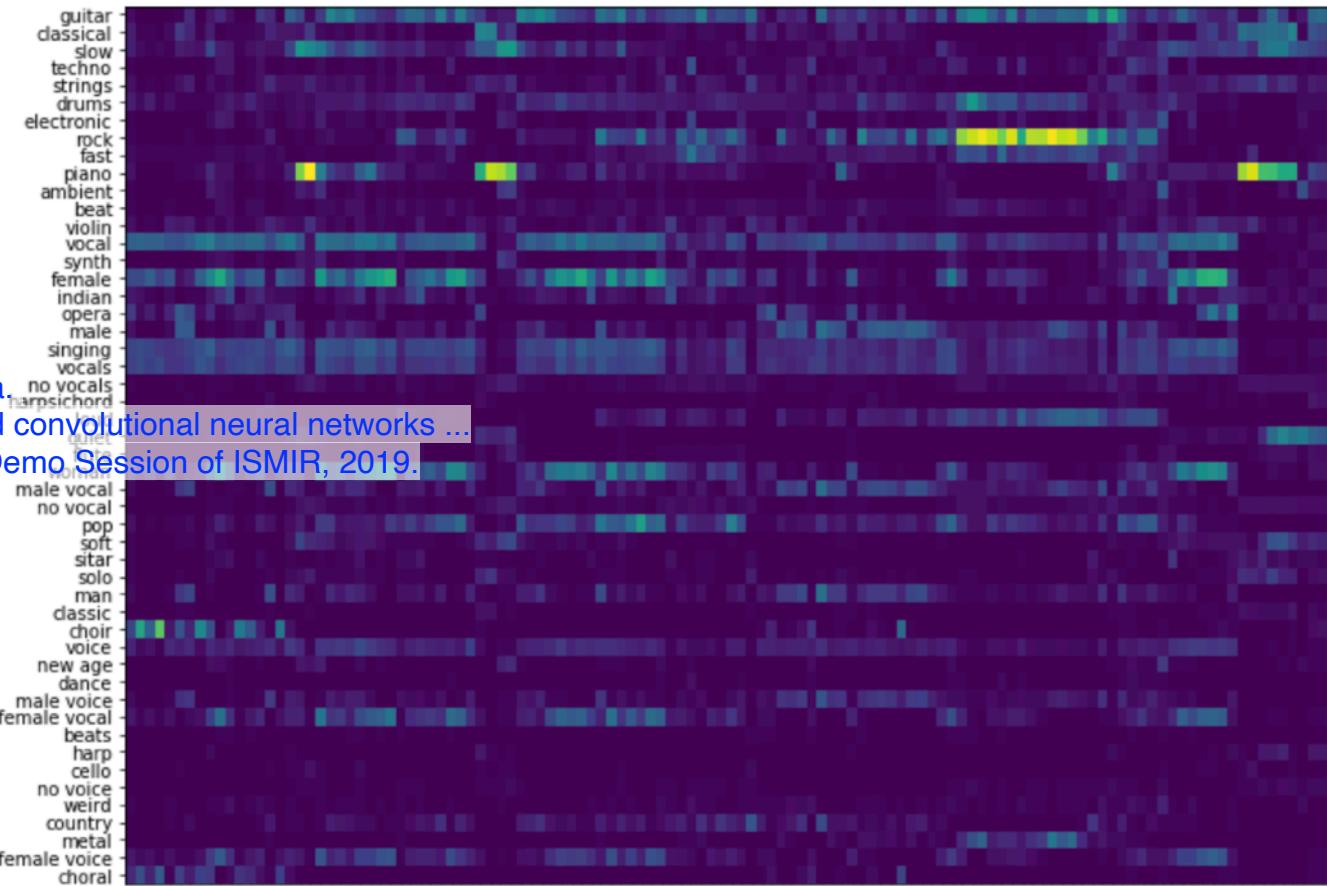
- YouTube-Music, Pandora, Spotify, Apple Music, Deezer, ...

- Auto-tagging
 - Lyrics synchronization
 - Recommendation
 - Search-by-similarity
 - Collaborative Filtering



The screenshot shows a search interface for 'maceo parker'. At the top, there's a search bar with the query 'maceo parker' and a 'Rechercher' button. Below the search bar, there's a media player area with playback controls and a volume slider. The main area is titled 'RÉSULTATS (6)' and lists six songs by Maceo Parker from the album 'Life on Planet Groove'. Each result includes the title, artist, album, and duration. To the right of the results, there are sections for 'HUMEURS' (Joyeux, Calme, Dynamique) and 'GENRES' (PopRock, Blues, Electronique, SoulFunk, Reggae, Classique, Jazz, Rap, Latin, R&B). Further down, there are sections for 'INSTRUMENTATIONS' (Guitare électrique, Guitare acoustique, Batterie, Cuivres, Orchestre à cordes, Piano, Acoustique) and 'ENREGISTREMENTS' (Studio, Live). At the bottom, there's a 'MES PLAYLISTS' section with a link to 'nouvelle playlist'.

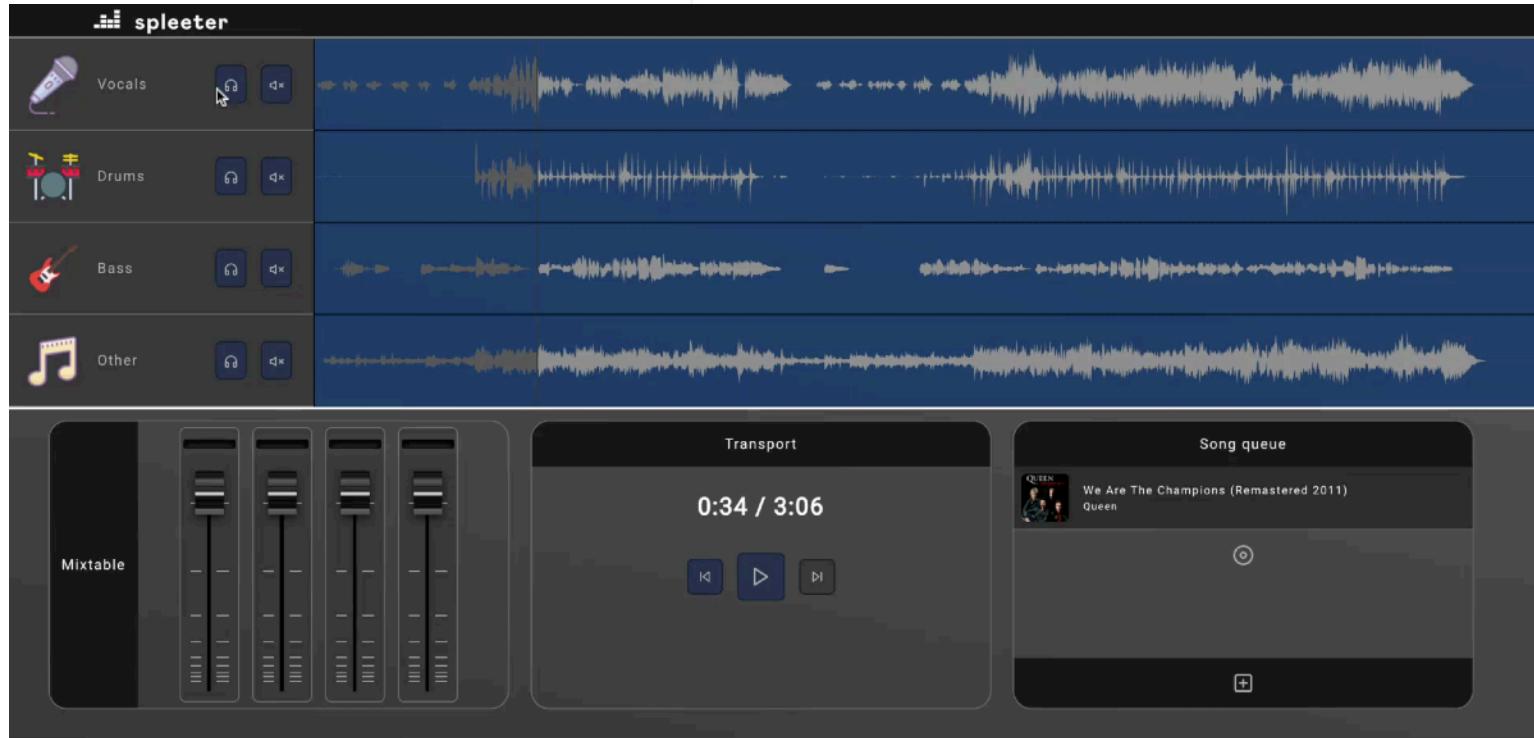
Music classification



J. Pons and X. Serra,
Musicnn: Pre-trained convolutional neural networks ...
In Late-Breaking/Demo Session of ISMIR, 2019.

Music Information Retrieval: Real-world applications

- **Source separation**
 - Spleeter by Deezer



Music Information Retrieval: Real-world applications

- **Music Generation**

- <https://magenta.tensorflow.org/>
- <https://openai.com/blog/jukebox/>



OpenAI

Curated samples

Provided with genre, artist, and lyrics as input, Jukebox outputs a new music sample produced from scratch. Below, we show some of our favorite samples.

[Unseen lyrics](#) [Re-renditions](#) [Completions](#) [Fun songs](#)

Jukebox produces a wide range of music and singing styles, and generalizes to lyrics not seen during training. All the lyrics below have been co-written by a language model and OpenAI researchers.

| | | |
|--|--|--|
| | Country, in the style of Alan Jackson – Jukebox | |
| | Rock, in the style of Elvis Presley – Jukebox | |
| | Pop, in the style of Katy Perry – Jukebox | |
| | Blues Rock, in the style of Joe Bonamassa – Jukebox | |
| | Heavy Metal, in the style of Rage – Jukebox | |
| | Classic Pop, in the style of Frank Sinatra – Jukebox | |

Detection and Classification of Acoustic Scenes and Events

DCASE

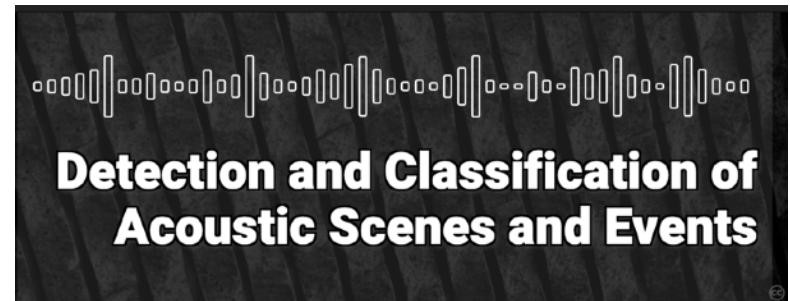
Detection and Classification of Acoustic Scenes and Events

DCASE

- <http://dcase.community>
- Sounds carry a large amount of information about our everyday environment and physical events that take place in it.

We can perceive the sound scene we are within (busy street, office, etc.), and recognize individual sound sources (car passing by, footsteps, etc.). Developing signal processing methods to automatically extract this information has huge potential in several applications, for example searching for multimedia based on its audio content, making context-aware mobile devices, robots, cars etc., and intelligent monitoring systems to recognize activities in their environments using acoustic information.

However, a significant amount of research is still needed to reliably recognize sound scenes and individual sound sources in realistic soundscapes, where multiple sounds are present, often simultaneously, and distorted by the environment.



Challenge status

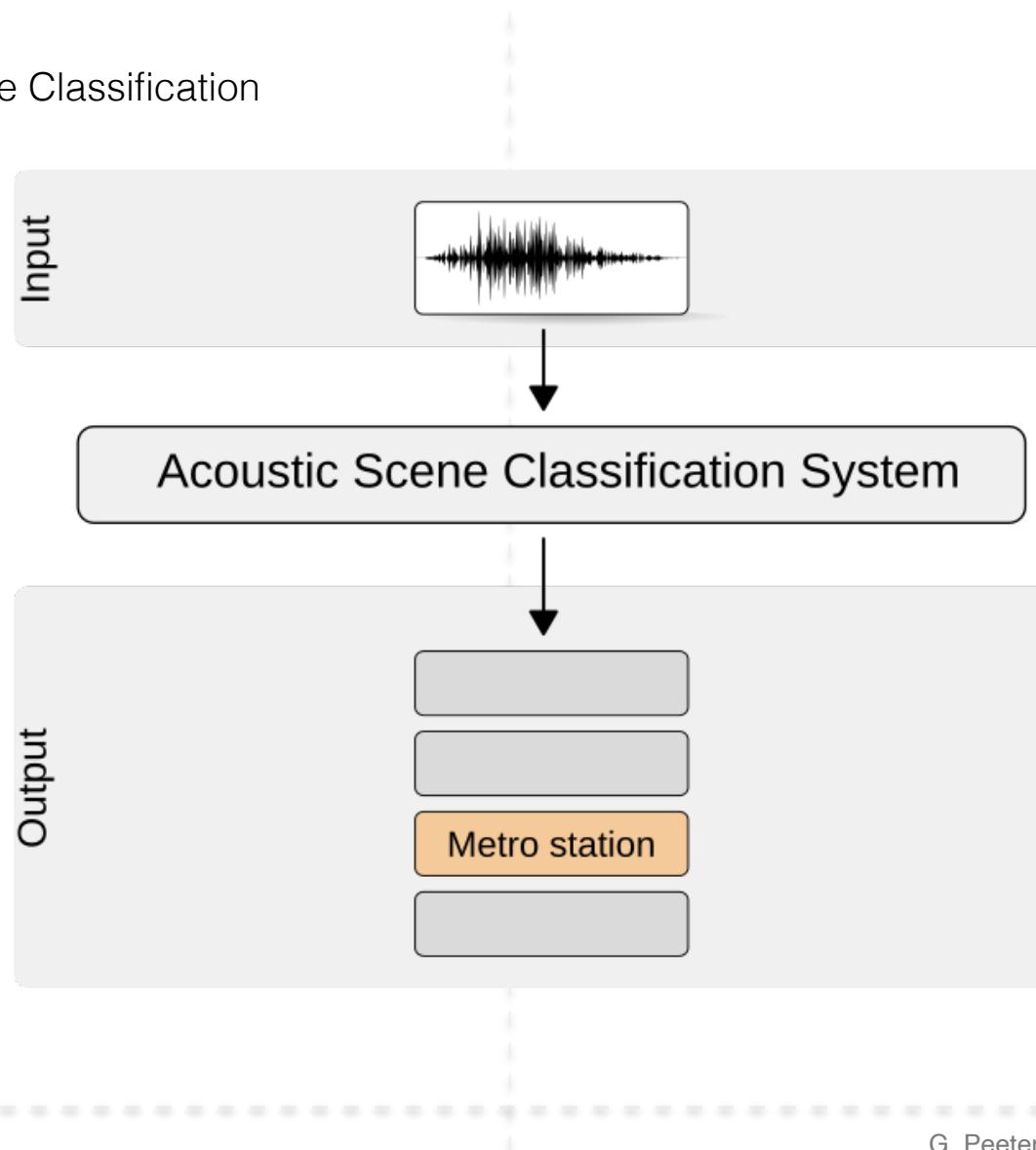
| Task | Task description | Development dataset | Baseline system | Evaluation dataset | Results |
|---|------------------|---------------------|-----------------|--------------------|---------|
| Task 1, Acoustic Scene Classification | Released | Released | Released | TBA | TBA |
| Task 2, Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions | Released | Released | Released | TBA | TBA |
| Task 3, Sound Event Localization and Detection with Directional Interference | Released | Released | Released | TBA | TBA |
| Task 4, Sound Event Detection and Separation in Domestic Environments | Released | Released | Released | TBA | TBA |
| Task 5, Few-shot Bioacoustic Event Detection | Released | Released | Released | TBA | TBA |
| Task 6, Automated Audio Captioning | Released | Released | Released | TBA | TBA |

Detection and Classification of Acoustic Scenes and Events

DCASE

Tasks

- Acoustic Scene Classification

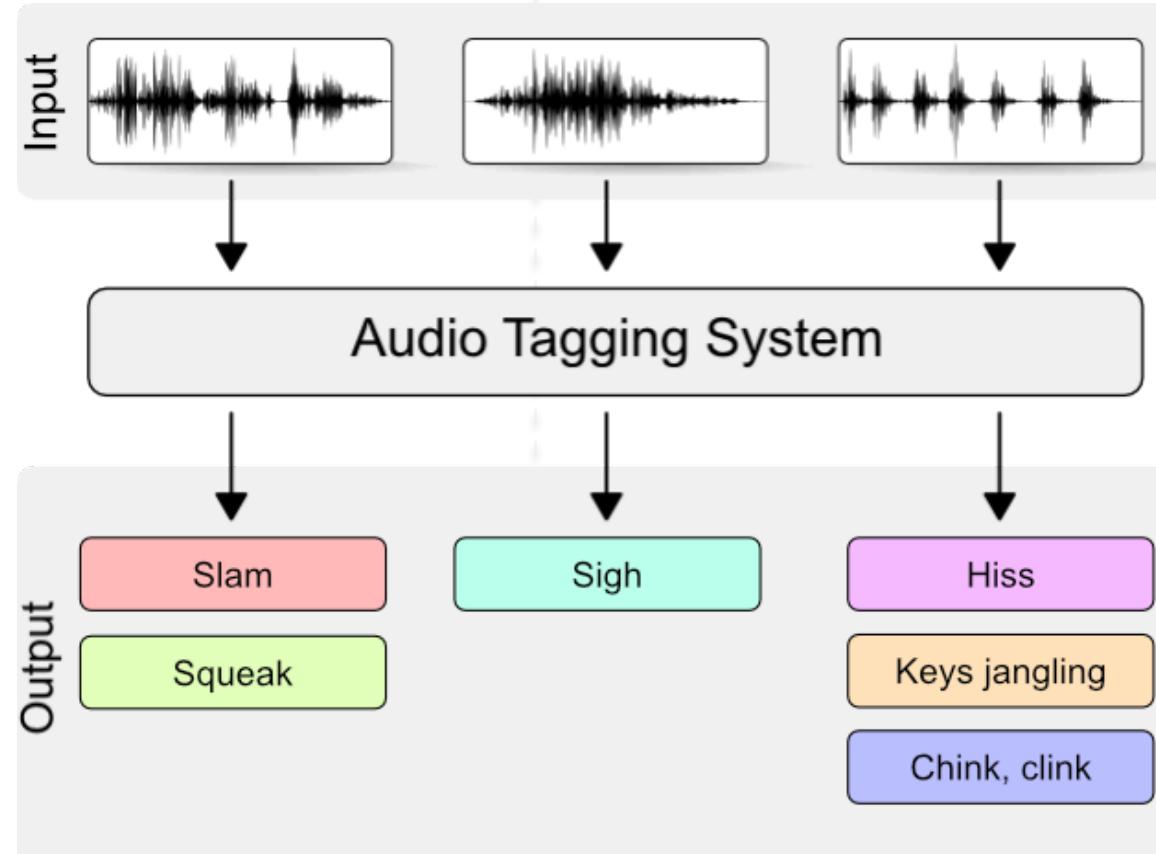


Detection and Classification of Acoustic Scenes and Events

DCASE

Tasks

- Audio tagging with noisy labels and minimal supervision

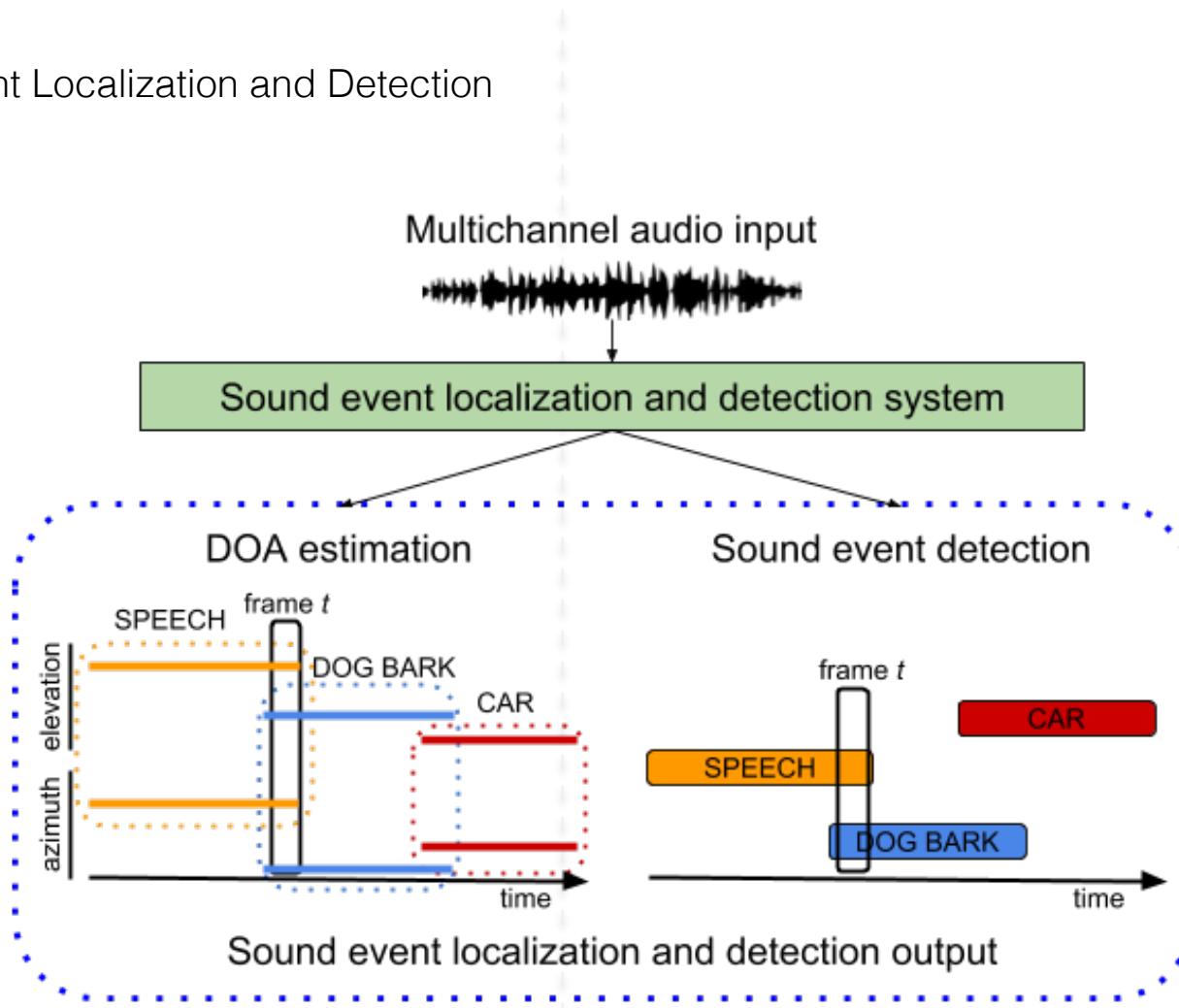


Detection and Classification of Acoustic Scenes and Events

DCASE

Tasks

- Sound Event Localization and Detection

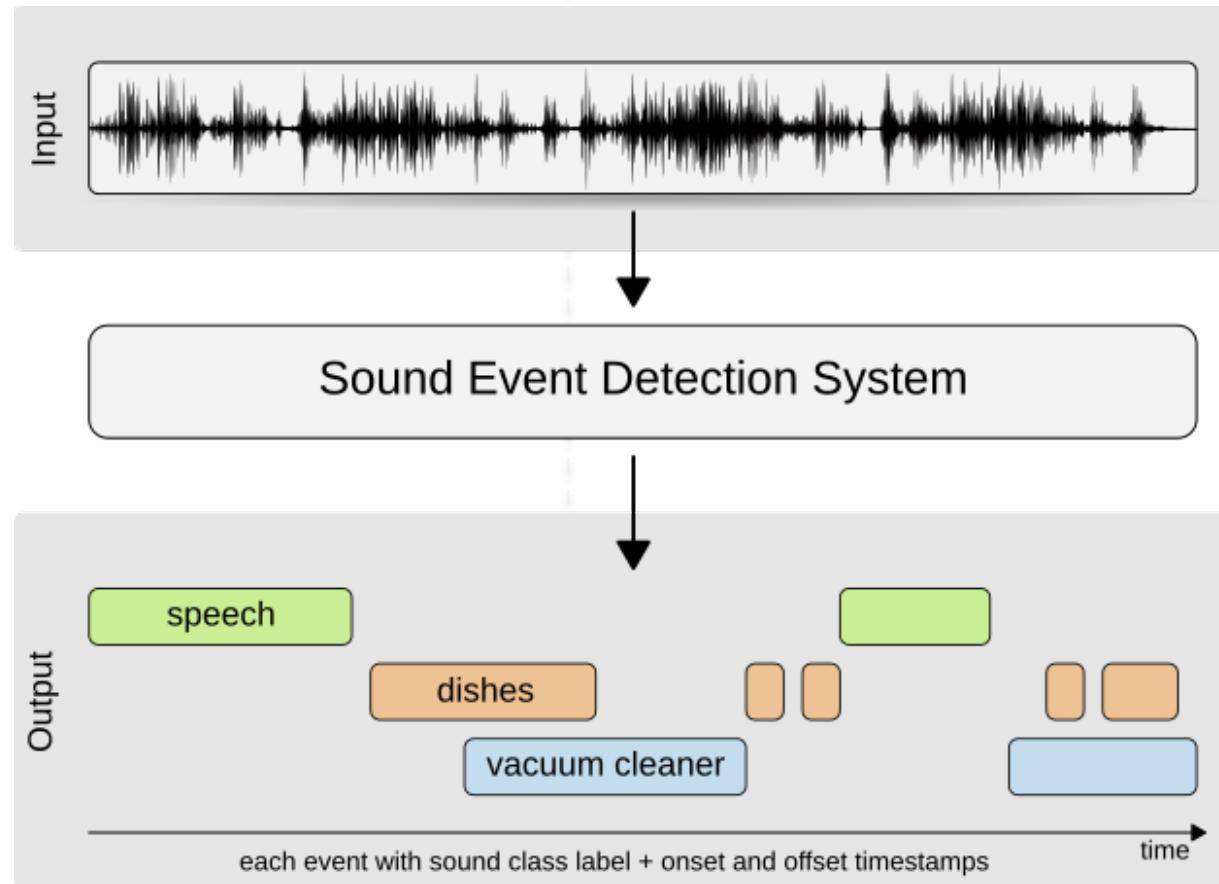


Detection and Classification of Acoustic Scenes and Events

DCASE

Tasks

- Sound event detection in domestic environments

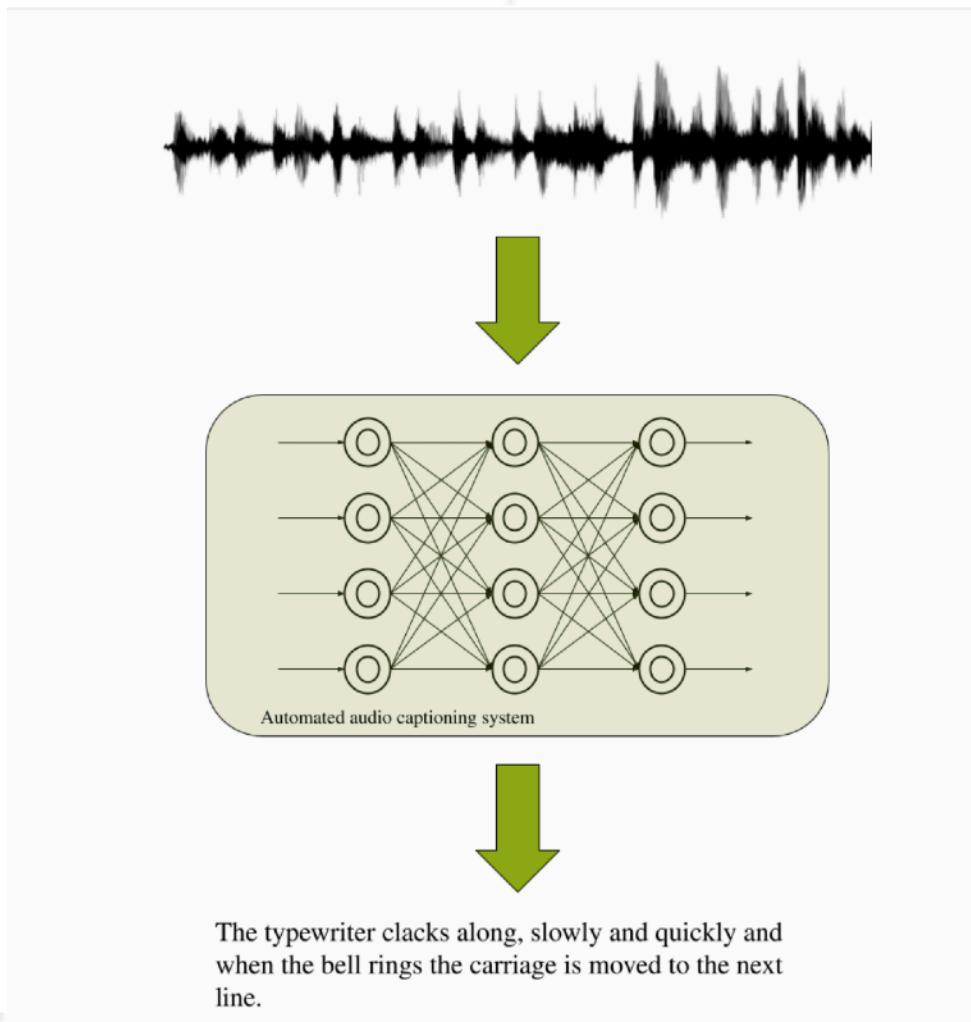


Detection and Classification of Acoustic Scenes and Events

DCASE

Tasks

- Automated audio captioning (AAC)



Transformée de Fourier

Transformée de Fourier temps et fréquences continus

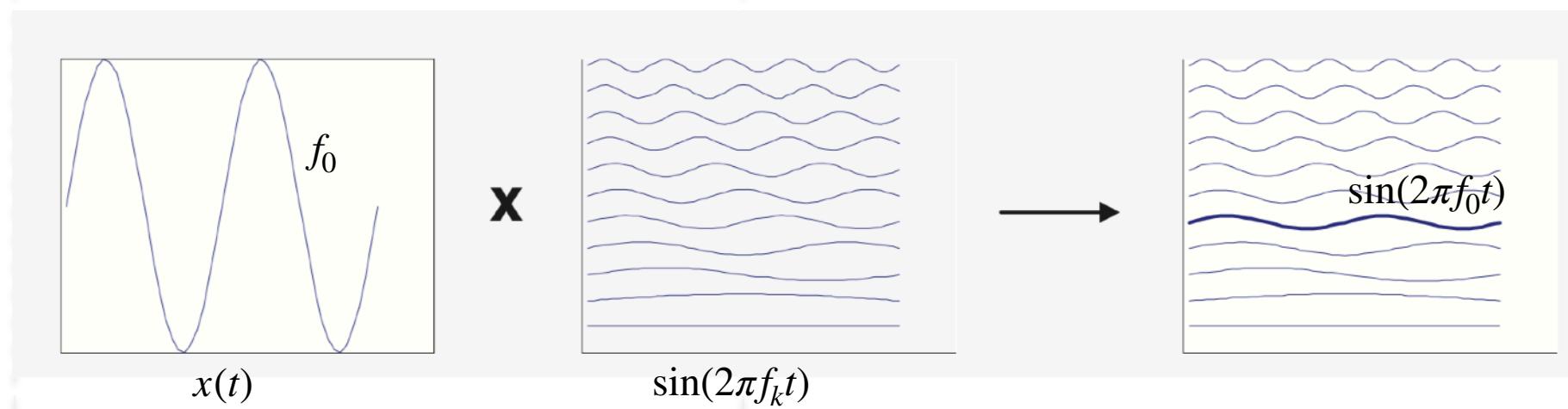
$$X(\omega) = \int_t x(t)e^{-j\omega t} dt \quad X(f) = \int_t x(t)e^{-j(2\pi f)t} dt$$

– Variables

- t : le **temps**
- $\omega = 2\pi f$: les **fréquences continues** exprimées en radian
- $\exp(j2\pi ft) = \cos(2\pi ft) + j \sin(2\pi ft)$

– Pourquoi la Transformée de Fourier ?

- Difficile d'extraire des observations directement à partir de la forme d'onde $x(t)$
- Reproduire la décomposition en fréquences de l'oreille humaine



Transformée de Fourier temps et fréquences continus

| Propriétés | $x(t)$ | $X(f)$ |
|-------------|--|--|
| Similitude | $x(at)$ | $\frac{1}{ a } X\left(\frac{f}{ a }\right)$ |
| Linéarité | $ax(t) + by(t)$ | $aX(f) + bY(f)$ |
| Translation | $x(t - t_0)$ | $X(f) \exp(-j2\pi f t_0)$ |
| Modulation | $x(t) \exp(j2\pi f_0 t)$ | $X(f - f_0)$ |
| Convolution | $x(t) \circledast y(t)$ | $X(f) Y(f)$ |
| Produit | $x(t)y(t)$ | $X(f) \circledast Y(f)$ |
| Parité | réelle paire réelle impaire imaginaire paire imaginaire impaire complexe paire complexe impaire réelle $x^*(t)$ | réelle paire imaginaire paire imaginaire paire réelle impaire complexe paire complexe impaire $X(f) = X^*(-f)$ $\Re(X(f))$ est paire $\Im(X(f))$ est impaire $X^*(f)$ |

Transformée de Fourier temps et fréquences discrets

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

- Variables

- m : le numéro d'**échantillon**
- k : les **fréquences discrètes**

- Fréquence d'échantillonnage (sample rate) sr

- sr définit à quelle fréquence le signal temporel va être échantillonné
- Exemple:

- Compact Disc (CD): $sr = 44100 \text{ Hz}$

- La distance temporelle entre deux échantillons (le pas d'échantillonnage) est de

$$\Delta t = \frac{1}{44100} = 0.000023 \text{ s}$$

- sr doit être $>$ à deux fois la f_{\max} présente dans le signal

- Sinon: repliement spectral
 - exemple: captation dans les films d'une roue d'une voiture qui accélère

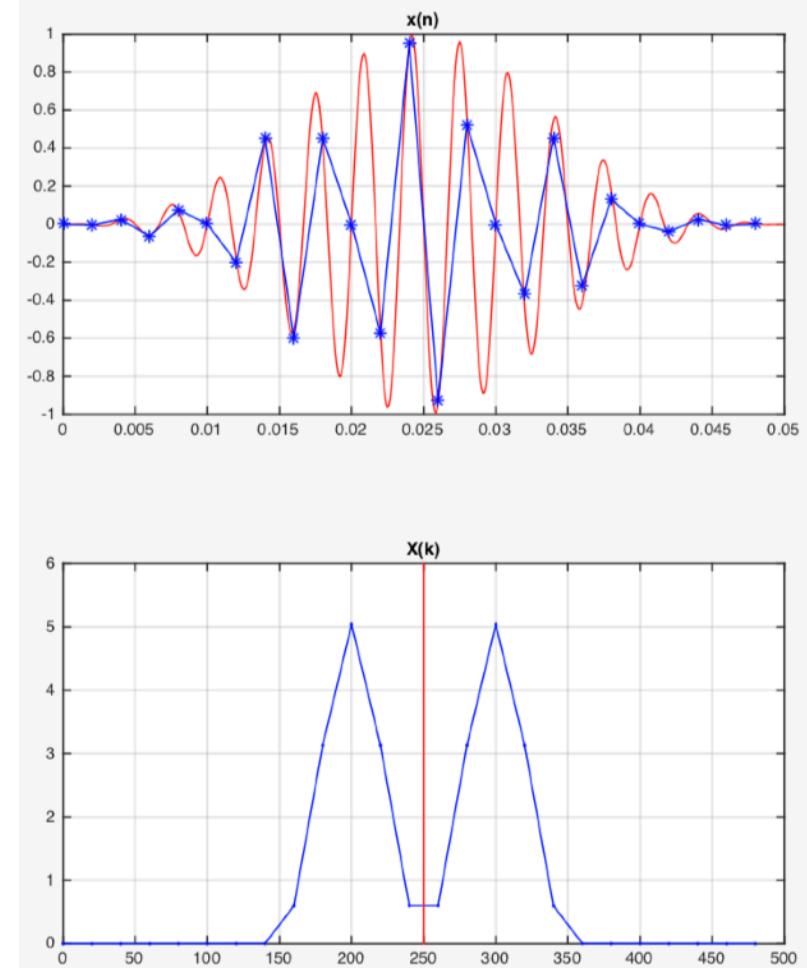
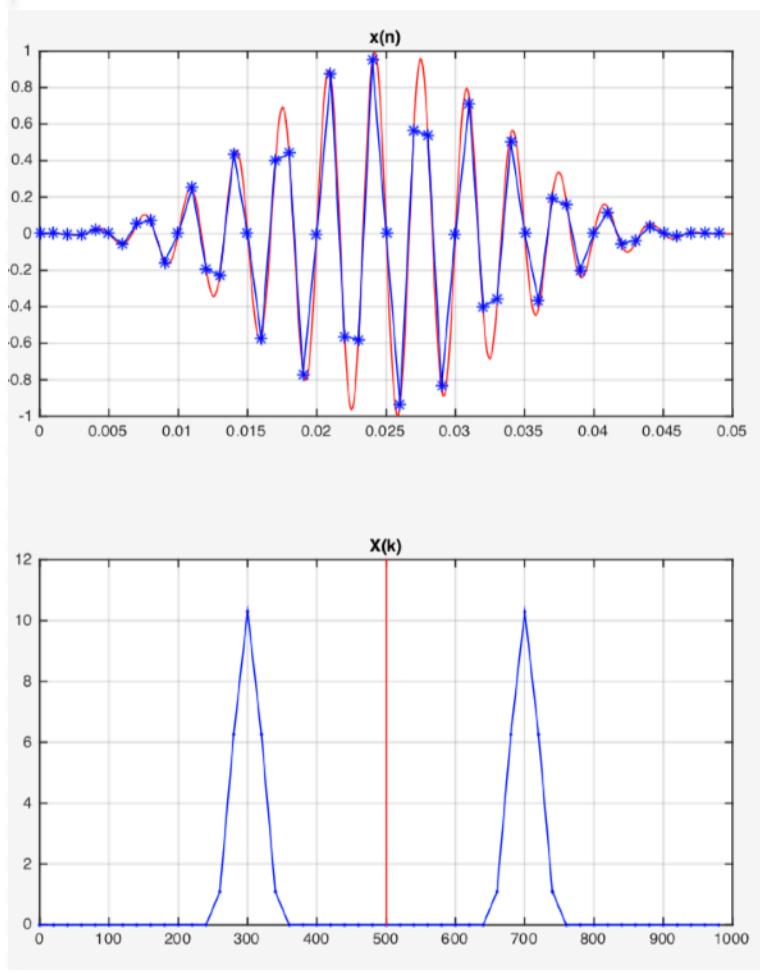
- **Fréquence de Nyquist**: $f_{Nyquist} = \frac{sr}{2} > f_{\max}$



Transformée de Fourier temps et fréquences discrets

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

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



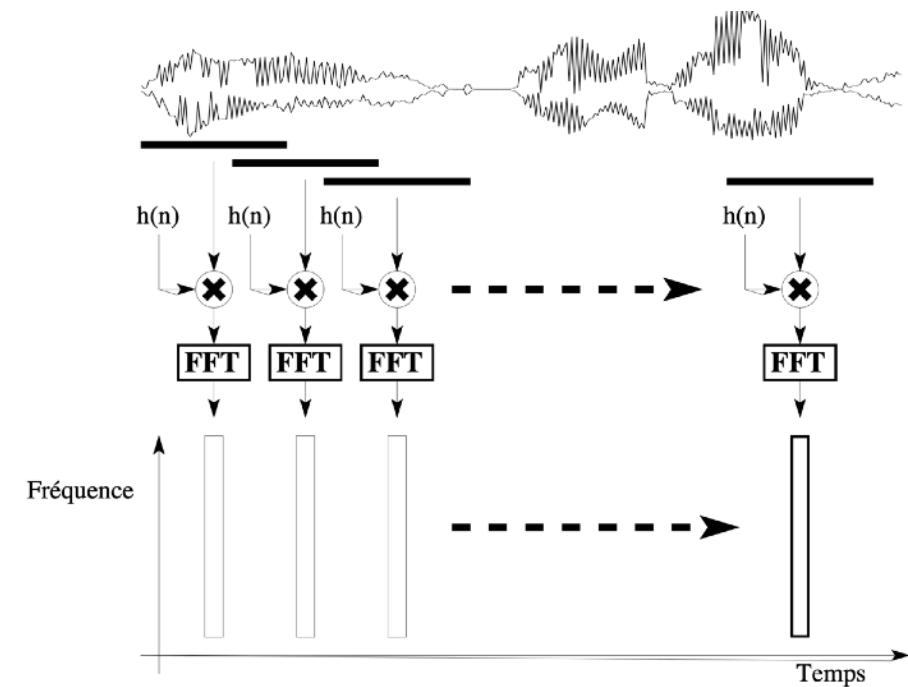
Transformée de Fourier à Court Terme (TFCT)

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

- Application de la TFD à une portion du signal centrée autour de l'échantillon n

- Pourquoi la TFCT ?

- Signal audio = non-stationnaire
 - ses propriétés varient au cours du temps
- Stationnarité "locale" (en temps)
 - sur une durée de $\pm 40 \text{ ms}$
- TFCT: suite d'analyse de Fourier sur des durées de $\pm 40 \text{ ms}$
 - = analyse à Court Terme ("trames/frames" en vidéo)



Transformée de Fourier à Court Terme (TFCT)

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

- Fenêtre de pondération $w(t)$

- $x(t) \cdot w(t) \rightleftharpoons X(f) \circledast W(f)$

- $w(t)$ est appelé "fenêtre de pondération"

- Paramètres:

- 1) **types** de la fenêtre $w(t)$
- 2) **longueur** temporelle L de la fenêtre $w(t)$

- Le type et la longueur détermines les caractéristiques spectrales

- Largeur du lobe principale (à $-6dB_{20}$): $B_w = \frac{C_w}{L}$
- Hauteur des lobes secondaires

| Propriétés | $x(t)$ | $X(f)$ |
|-------------|---|--|
| Similitude | $x(at)$ | $\frac{1}{ a } X(\frac{f}{ a })$ |
| Linéarité | $ax(t) + by(t)$ | $aX(f) + bY(f)$ |
| Translation | $x(t - t_0)$ | $X(f) \exp(-j2\pi f t_0)$ |
| Modulation | $x(t) \exp(j2\pi f_0 t)$ | $X(f - f_0)$ |
| Convolution | $x(t) \circledast y(t)$ | $X(f) Y(f)$ |
| Produit | $x(t) y(t)$ | $X(f) \circledast Y(f)$ |
| Parité | <p>réelle paire réelle impaire imaginaire paire imaginaire impaire complexe paire complexe impaire réelle</p> | <p>réelle paire imaginnaire paire imaginnaire paire réelle impaire complexe paire complexe impaire $X(f) = X^*(-f)$ $\Re(X(f))$ est paire $\Im(X(f))$ est impaire $X^*(f)$</p> |
| | $x^*(t)$ | |

Transformée de Fourier à Court Terme (TFCT)

- 1) **Type** de la fenêtre $w(t)$:

- Rectangulaire

- $w(t) = 1$ si $t \in [0, T]$

- $Cw = 1.21$

- Hann

- $w(t) = 0.5 - 0.5 \cos\left(2\pi \frac{t}{L}\right)$ si $t \in [0, T]$

- $Cw = 2$

- Hamming

- $w(t) = 0.54 - 0.46 \cos\left(2\pi \frac{t}{L}\right)$ si $t \in [0, T]$

- $Cw = 1.81$

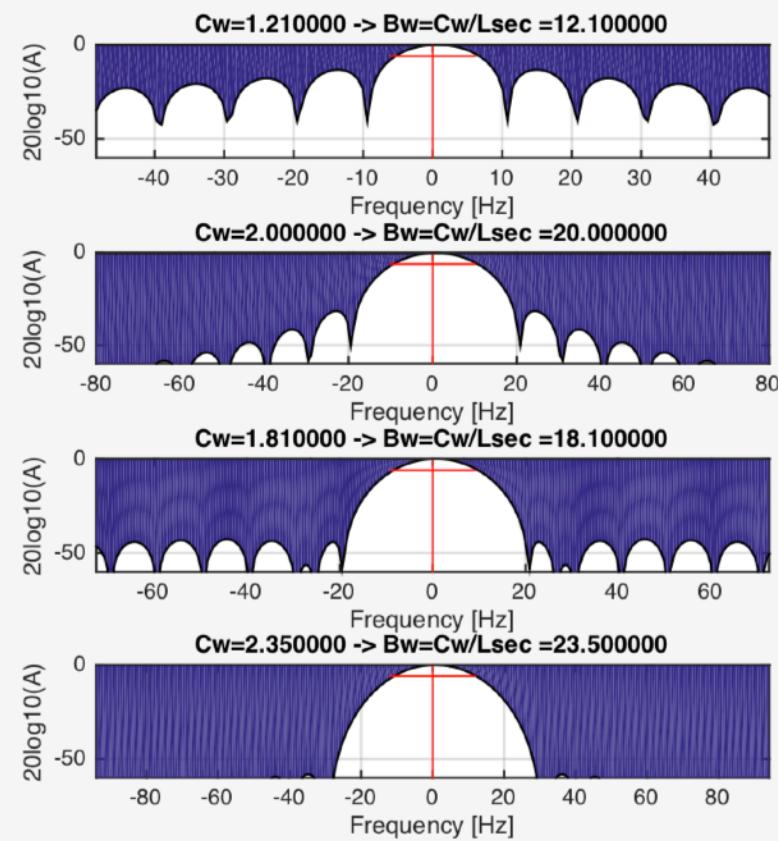
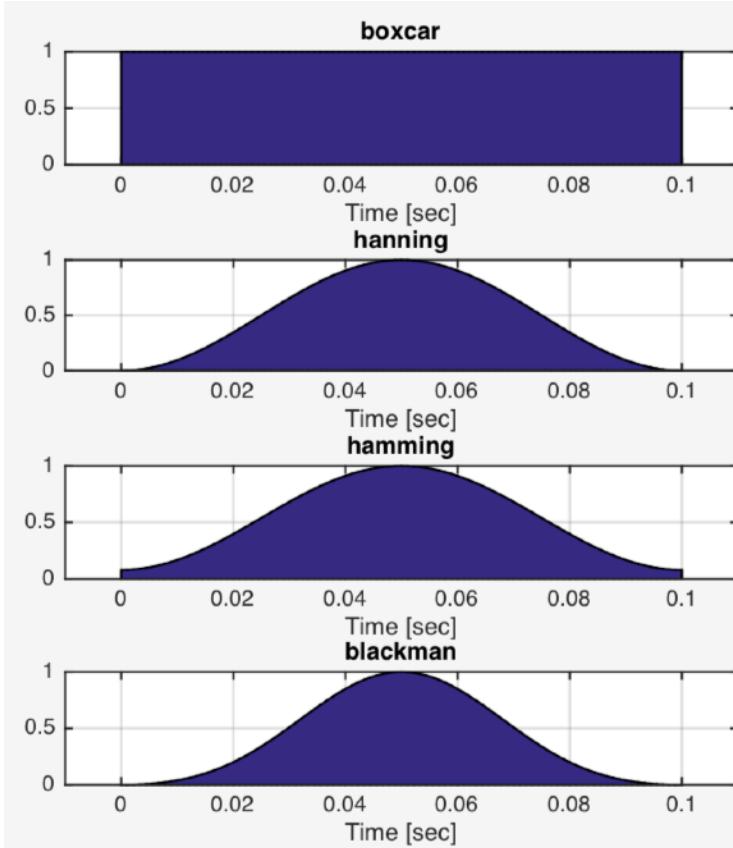
- Blackman

- $w(t) = 0.42 - 0.5 \cos\left(2\pi \frac{t}{L}\right) + 0.08 \cos\left(4\pi \frac{t}{L}\right)$ si $t \in [0, T]$

- $Cw = 2.35$

Transformée de Fourier à Court Terme (TFCT)

- 1) **Type** de la fenêtre $w(n)$



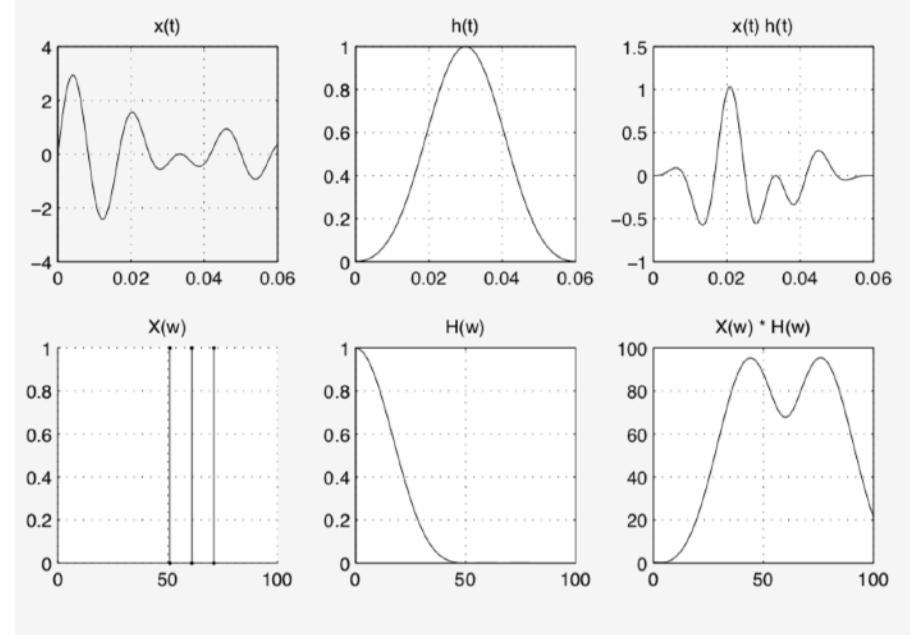
Transformée de Fourier à Court Terme (TFCT)

- 2) **Longueur** temporelle L de la fenêtre $w(n)$:
 - Au plus la fenêtre est courte ($L \ll$),
 - au plus on observe précisément les temps (mais pas les fréquences)
 - Au plus la fenêtre est longue ($L \gg$),
 - au plus on observe précisément les fréquences (mais pas les temps)

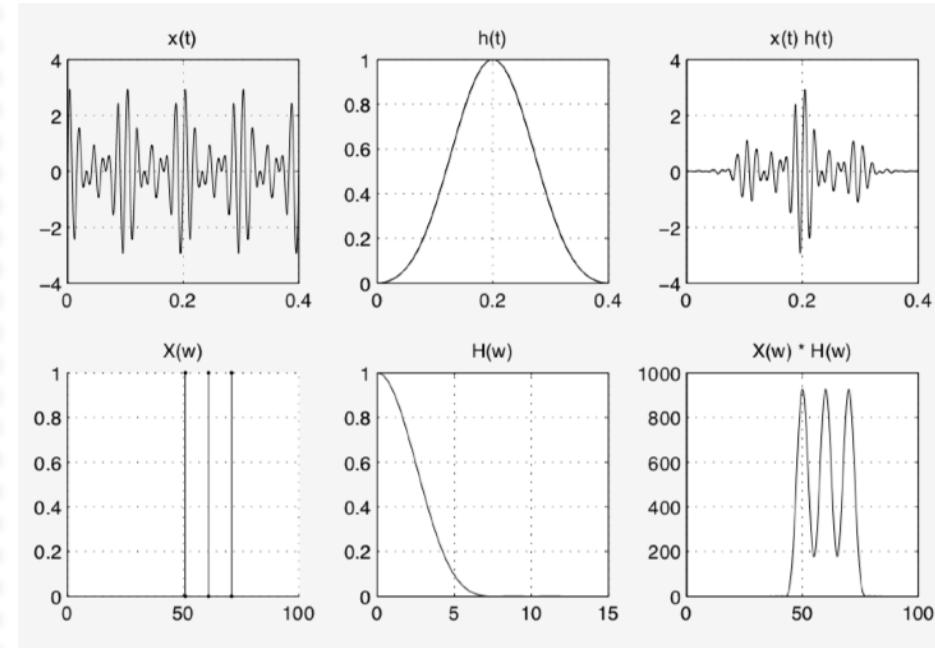
Transformée de Fourier à Court Terme (TFCT)

- 2) **Longueur** temporelle L de la fenêtre $w(n)$:

$$L = 0.06 \text{ s}$$



$$L = 0.4 \text{ s}$$

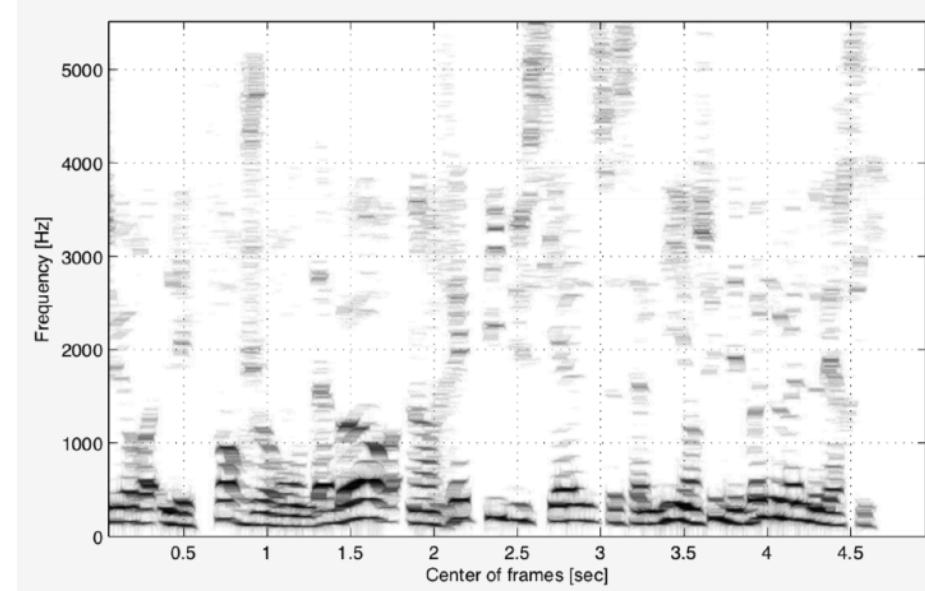
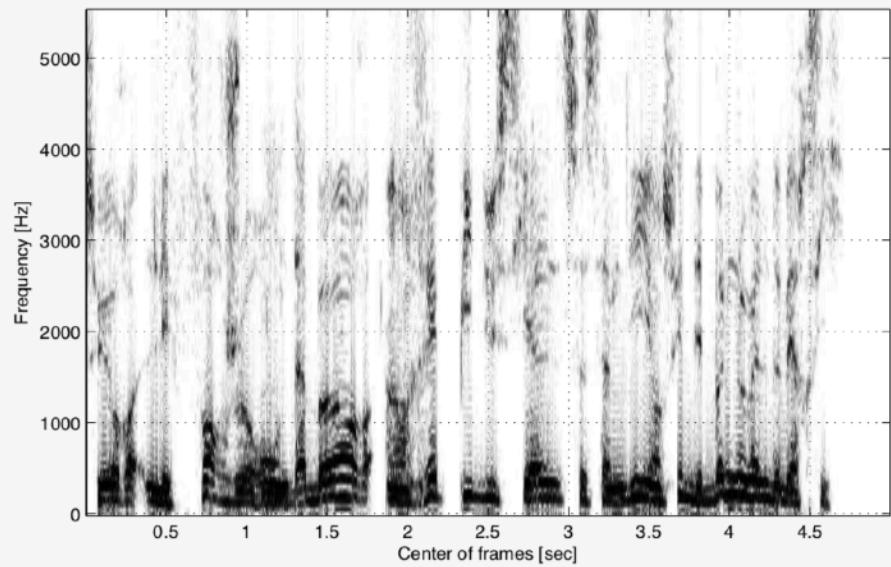


Transformée de Fourier à Court Terme (TFCT)

- 2) **Longueur** temporelle L de la fenêtre $w(n)$:

$$L = 0.01 \text{ s}$$

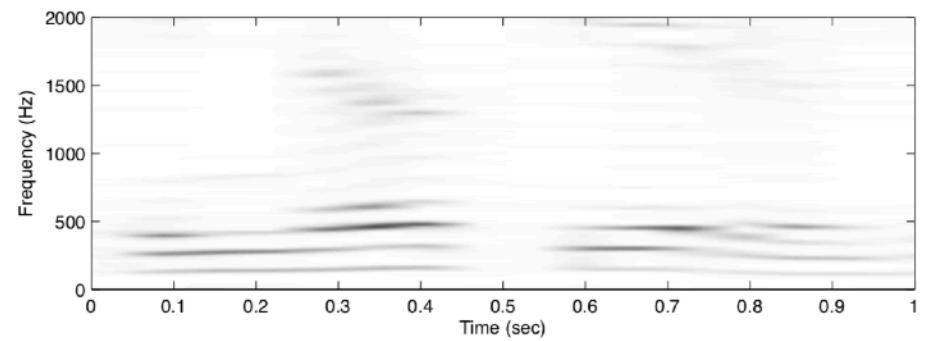
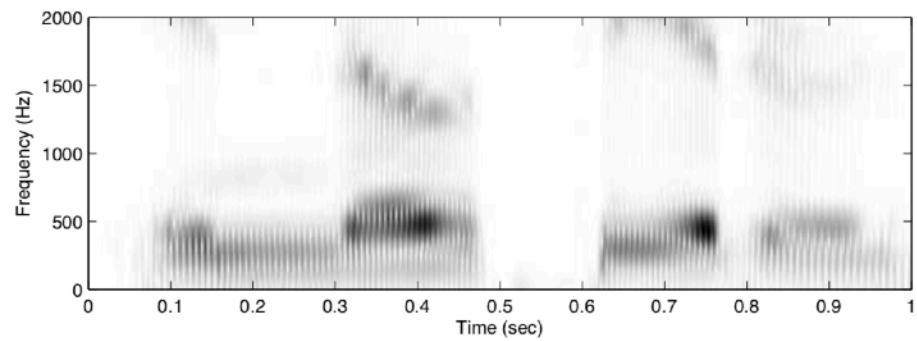
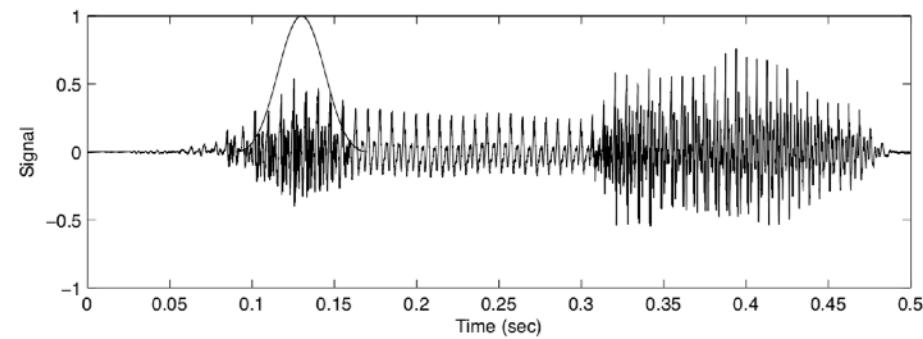
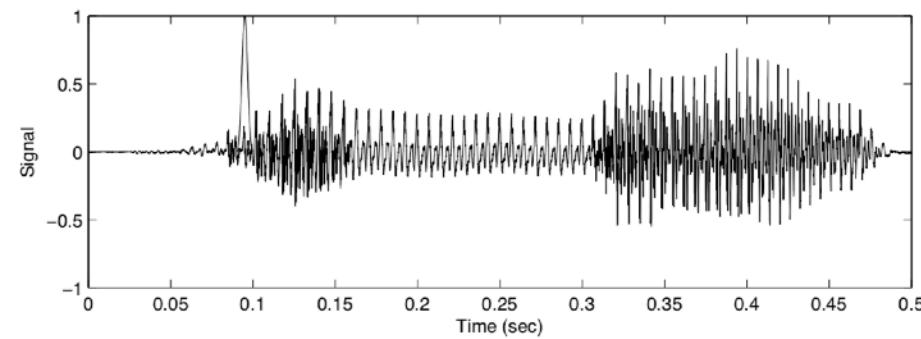
$$L = 0.1 \text{ s}$$



Transformée de Fourier à Court Terme (TFCT)

- **Paradoxe temps/ fréquence**

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



- **Comme résoudre ce problème ?**

- Utiliser d'autres transformées que celle de Fourier

Transformée de Fourier à Court Terme

Transformée de Fourier à Court Terme

Deux interprétations

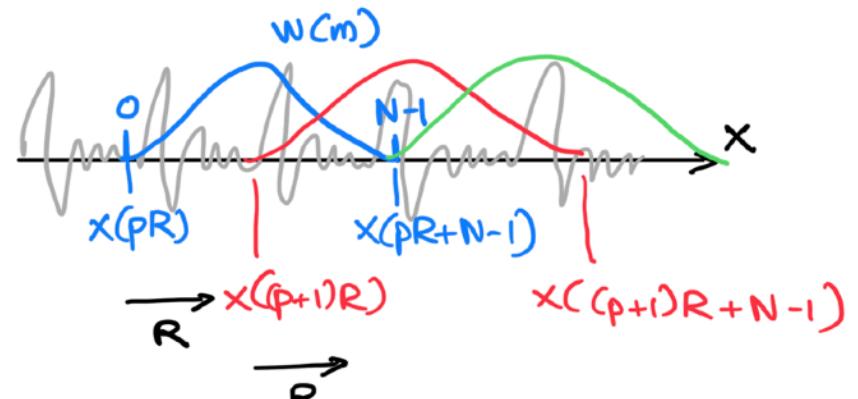
Jean Laroche

- **L'implémentation de la TFCT sur un CPU, GPU**

- utilise une longueur de fenêtre N
 - un pas d'avancement R
 - p : numéro de la trame
- Notation: $X(k, n) \rightarrow \tilde{X}(w_0, pR)$

$$\tilde{X}(\omega_0, p) = \sum_{m=0}^{N-1} w(-m)x(pR + m)e^{-j\omega_0 m}$$

- $[x(pR) \dots x(pR + N)]$ puis $[x((p + 1)R) \dots x((p + 1)R + N)]$



- **Deux interprétations possibles**

Transformée de Fourier à Court Terme

Deux interprétations

- **Convention passe-bande**

- correspond à l'implémentation de la TFCT:

- la référence du temps de la DFT est le début de la fenêtre m

$$\tilde{X}(\omega_0, p) = \sum_{m=0}^{N-1} w(-m)x(pR + m)e^{-j\omega_0 m}$$

- si $R = 1$ et en notant $m' = -m$

$$\begin{aligned}\tilde{X}(\omega_0, p) &= \sum_{m'=0}^{N-1} w(m')x(p - m')e^{-j\omega_0 m'} \\ &= \sum_{m'} w(m') \underbrace{e^{j\omega_0 m'}}_{w_0(m')} x(p - m') \\ &= \sum_{m'} w_0(m')x(p - m')\end{aligned}$$

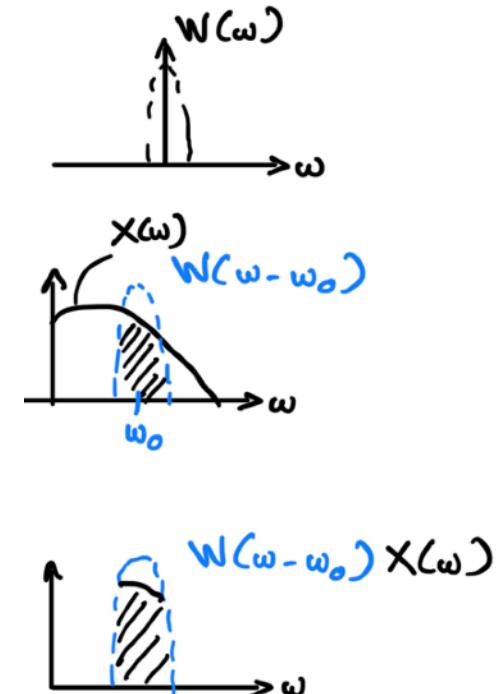
- $w_0(m')$: la fenêtre d'analyse modulée (passe-bande)

- $\tilde{X}(\omega_0, p)$ est le résultat de la convolution de $x(m)$ par filtre passe-bande centré autour de ω_0 : $w_0(m)$

- Interprétation de la TFCT en terme de banc de filtres (passe-bandes)

- si $R \neq 1$,

- $\tilde{X}(\omega_0, p)$ est une version sous-échantillonnée par un facteur R du signal passe-bande



| Propriétés | $x(t)$ | $X(f)$ |
|-------------|--|--|
| Similitude | $x(at)$ | $\frac{1}{ a } X\left(\frac{f}{ a }\right)$ |
| Linéarité | $ax(t) + by(t)$ | $aX(f) + bY(f)$ |
| Translation | $x(t - t_0)$ | $X(f) \exp(-j2\pi f t_0)$ |
| Modulation | $x(t) \exp(j2\pi f_0 t)$ | $X(f - f_0)$ |
| Convolution | $x(t) * y(t)$ | $X(f) Y(f)$ |
| Produit | $x(t)y(t)$ | $X(f) \otimes Y(f)$ |
| Parité | réelle paire réelle impaire imaginaire paire imaginaire impaire complexe paire complexe impaire réelle | réelle paire imaginaire paire imaginaire paire réelle impaire complexe paire complexe impaire $X(f) = X^*(-f)$ $\Re(X(f))$ est paire $\Im(X(f))$ est impaire $X^*(f)$ |

Transformée de Fourier à Court Terme

Deux interprétations

- Convention passe-bas**

- correspond à la définition de la TFCT**

- la référence du temps de la DFT est le début du signal $m' = pR + m$

$$\begin{aligned}\tilde{X}(\omega_0, p) &= \sum_{m=0}^{N-1} w(-m) x(pR + m) e^{-j\omega_0 m} \\ &= e^{j\omega_0 pR} \sum_{m'} w(pR - m') \underbrace{x(m') e^{-j\omega_0 m'}}_{x_0(m')} \\ &= e^{j\omega_0 pR} X(\omega_0, p)\end{aligned}$$

- $X(\omega_0, p)$ est maintenant un signal passe-bas (contenu fréquentiel est localisé autour de la fréquence nulle);
 - la convention passe-bas se réfère à $X(\omega_0, p)$
 - définition de la TFCT la plus souvent utilisée, en raison de la simplification des calculs qu'elle entraîne

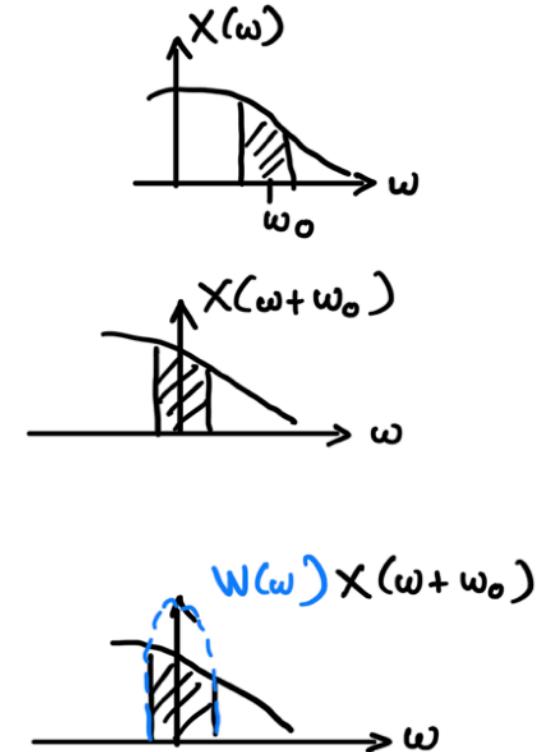
- si $R = 1$

$$X(\omega_0, p) = \sum_{n'} w(p - m') \underbrace{x(m') e^{-j\omega_0 m'}}_{x_0(m')}$$

- $X(\omega_0, p)$ est obtenu par filtrage passe-bas $w(m)$ du signal $x_0(m)$

- si $R \neq 1$

- $X(\omega_0, p)$ est une version sous-échantillonnée par un facteur R du signal passe-bande



| Propriétés | $x(t)$ | $X(f)$ |
|-------------|--|--|
| Similitude | $x(at)$ | $\frac{1}{ a } X(\frac{f}{ a })$ |
| Linéarité | $ax(t) + by(t)$ | $aX(f) + bY(f)$ |
| Translation | $x(t - t_0)$ | $X(f) \exp(-j2\pi f t_0)$ |
| Modulation | $x(t) \exp(j2\pi f_0 t)$ | $X(f - f_0)$ |
| Convolution | $x(t) * y(t)$ | $X(f) Y(f)$ |
| Produit | $x(t)y(t)$ | $X(f) \otimes Y(f)$ |
| Parité | réelle paire réelle impaire imaginaire paire imaginaire impaire complexe paire complexe impaire réelle | réelle paire imaginaire paire imaginaire paire réelle impaire complexe paire complexe impaire $X(f) = X^*(-f)$ $\Re(X(f))$ est paire $\Im(X(f))$ est impaire $X^*(f)$ |

Transformée de Fourier à Court Terme

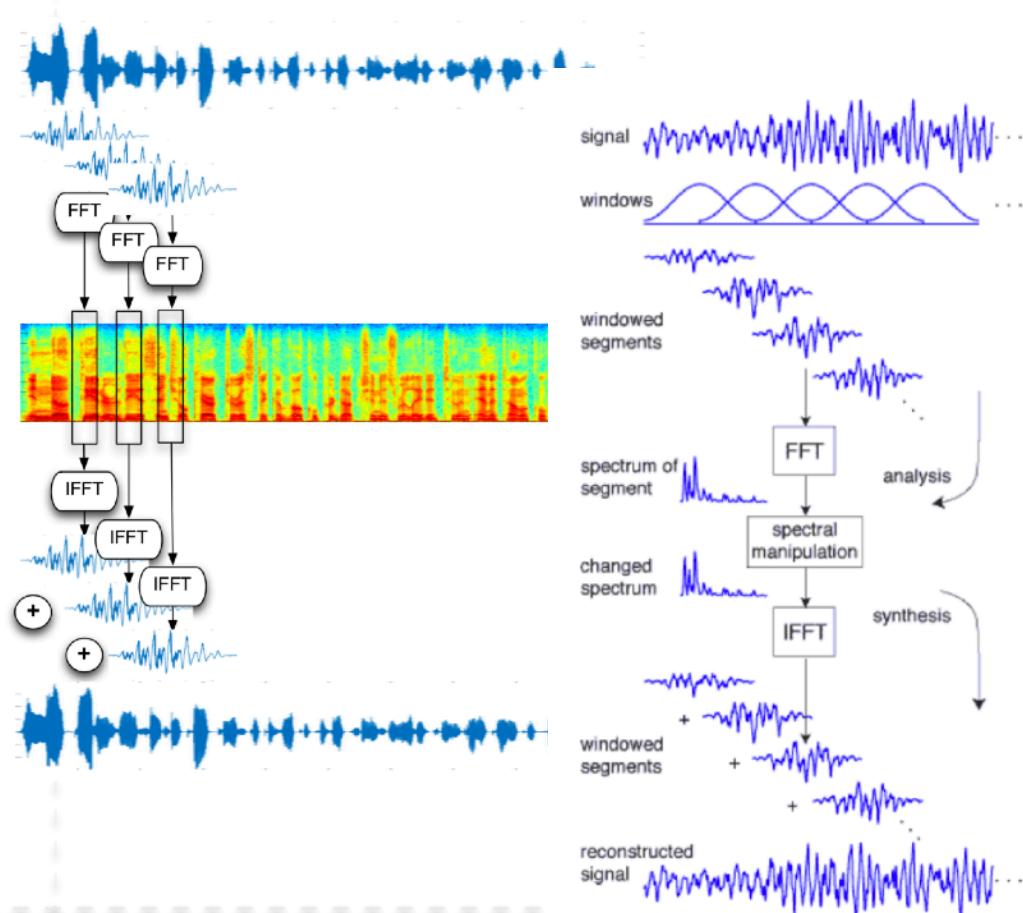
Reconstruction du signal par addition/ recouvrement (OLA)

- En l'absence de modification de la TFCT, on peut reconstruire le signal $x(n)$ à partir des informations de sa TFCT $X(k, pR)$ par addition/recouvrement (OverLap-Add, OLA)
 - Si on note $n = pR$, p le numéro de la trame d'analyse, R le pas d'avancement

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

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

$$\begin{aligned} \hat{y}(m) &= \sum_p y(m, pR) \\ &= \sum_p x(m)w(pR - m) \\ &= x(m) \sum_p w(pR - m) \\ x(m) &= \frac{\sum_p y(m, pR)}{\sum_r w(pR - m)} \end{aligned}$$



- **En présence de modification de la TFCT, il faut utiliser l'algorithme de Griffin & Lim**
 - **Redondance de la TFCT**
 - TFCT avec recouvrement de 50%
 - → volumes de données de la TFCT est double !
 - les informations contenues dans la TFCT ne sont pas indépendantes entre elles
 - **Modification du module de la TFCT (filtrage, denoising, ...)**
 - il n'est pas forcément possible de trouver un signal temporel correspondant à une TFCT arbitraire
 - **Algorithme de Griffin & Lim**
 - **à partir d'une TFCT quelconque (non nécessairement valide), il est possible de la modifier itérativement pour obtenir un signal temporel qui lui corresponde** (en changeant la phase par exemple)

Transformée de Fourier à Court Terme

Reconstruction du signal par algorithme de Griffin & Lim (II)

- Objectif:**

- estimer $x(n)$ tel que sa TFCT se rapproche le plus de $|Y_w(mS, \omega)|$
- minimiser la différence entre
 - $|X_w(mS, \omega)|$ (une TFCT valide) et
 - $|Y_w(mS, \omega)|$ (une TFCT non nécessairement valide)

- Algorithme itératif**

- à chaque itération, calculer la TFCT de $x^i(n) \rightarrow X_w^i(mS, \omega)$
- remplacer l'amplitude de $X_w^i(mS, \omega)$ par $|Y_w(mS, \omega)|$

$$\begin{aligned}\hat{X}_w^{i+1}(mS, \omega) &= Y_w^w(mS, \omega) | \cdot e^{j\angle X_w^i(mS, \omega)} \\ &= |Y_w^w(mS, \omega)| \frac{X_w^i(mS, \omega)}{|X_w^i(mS, \omega)|}\end{aligned}$$

- recalculer (overlap and add), le signal correspondant

$$x^{i+1}(n) = \frac{\sum_m w(mS - n) \int_{\omega=-\pi}^{\pi} \hat{X}_w^i(mS, \omega) e^{j\omega n} d\omega}{\sum_m w^2(mS - n)}$$

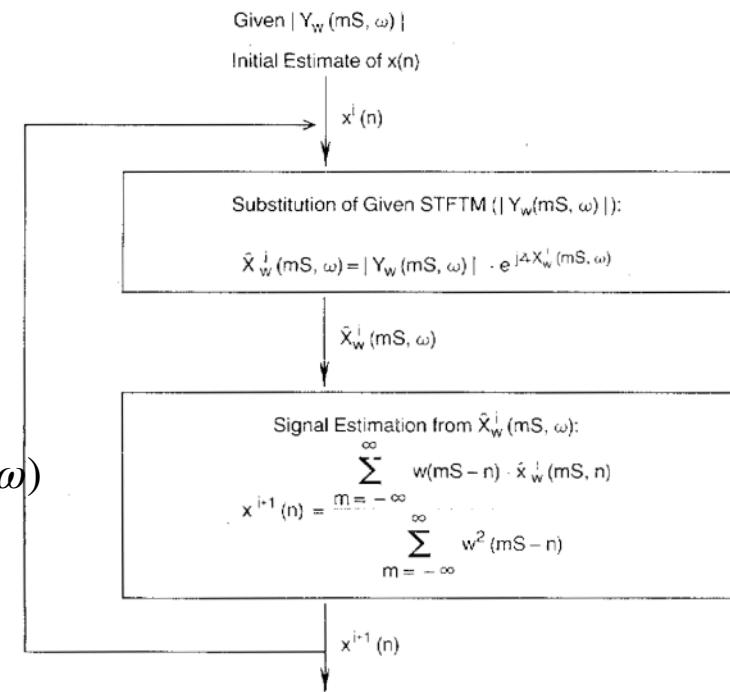


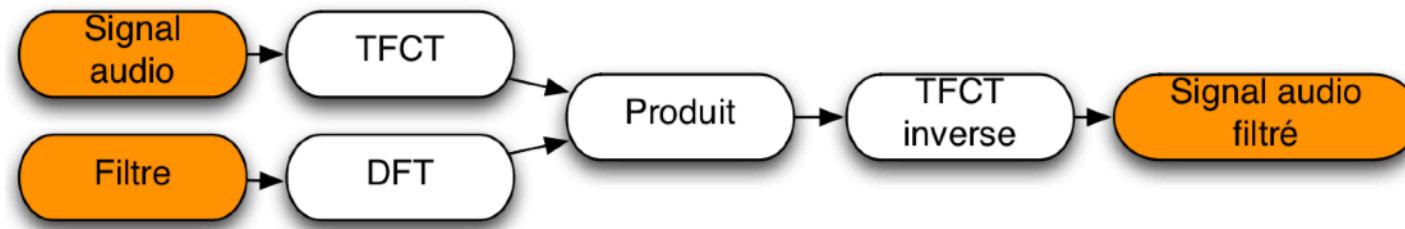
Fig. 1. LSEE-MSTFTM algorithm.

Filtrage par TFCT

Application

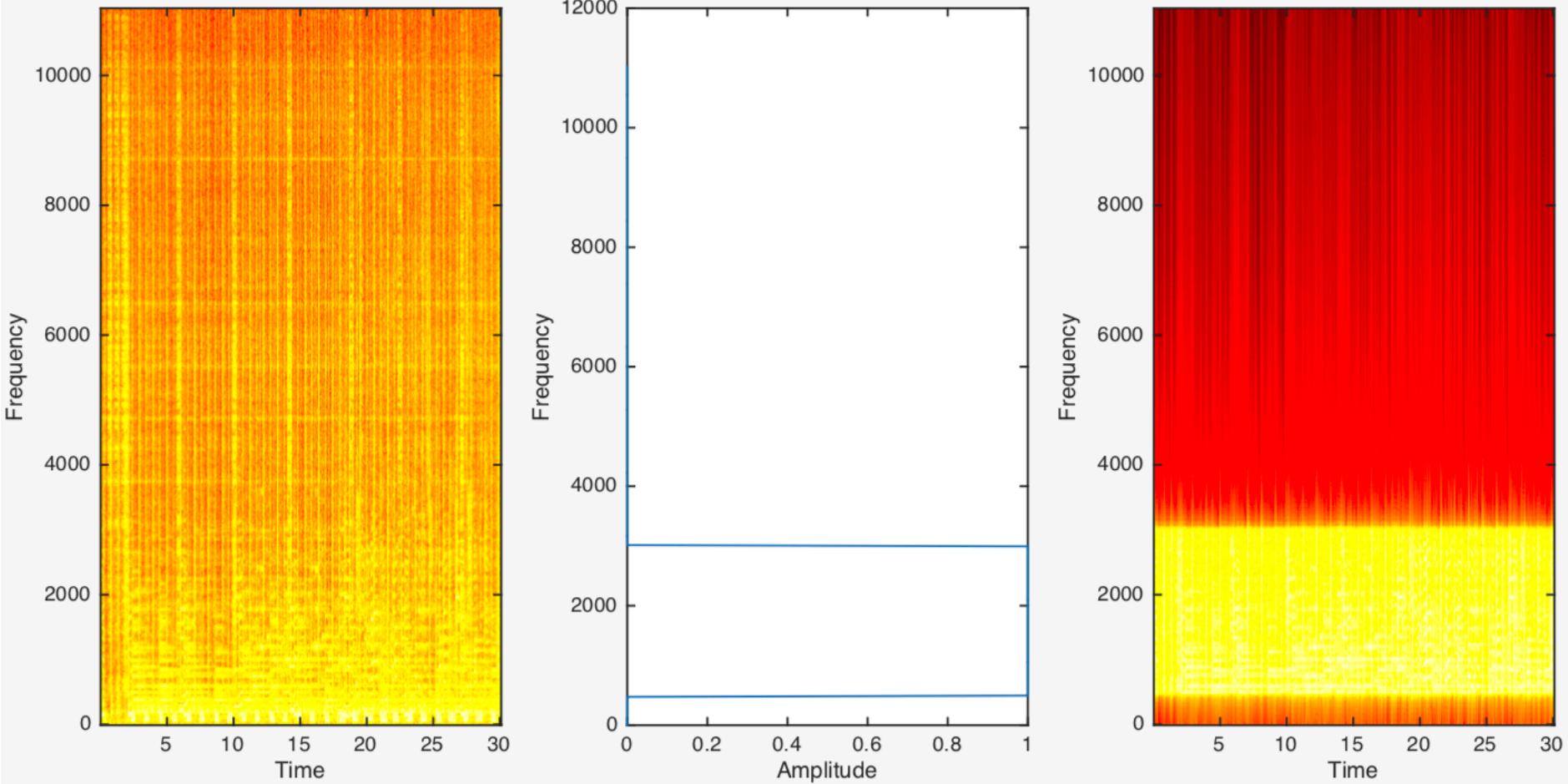
Filtrage constant au cours du temps

- Filtrage dans le domaine fréquentiel = très économique en coût de calcul
 - $x(t) \otimes h(t) \Leftrightarrow X(\omega)H(\omega)$
 - convolution temps \Leftrightarrow produit en fréquence
 - utilisation de l'algorithme de FFT



Application

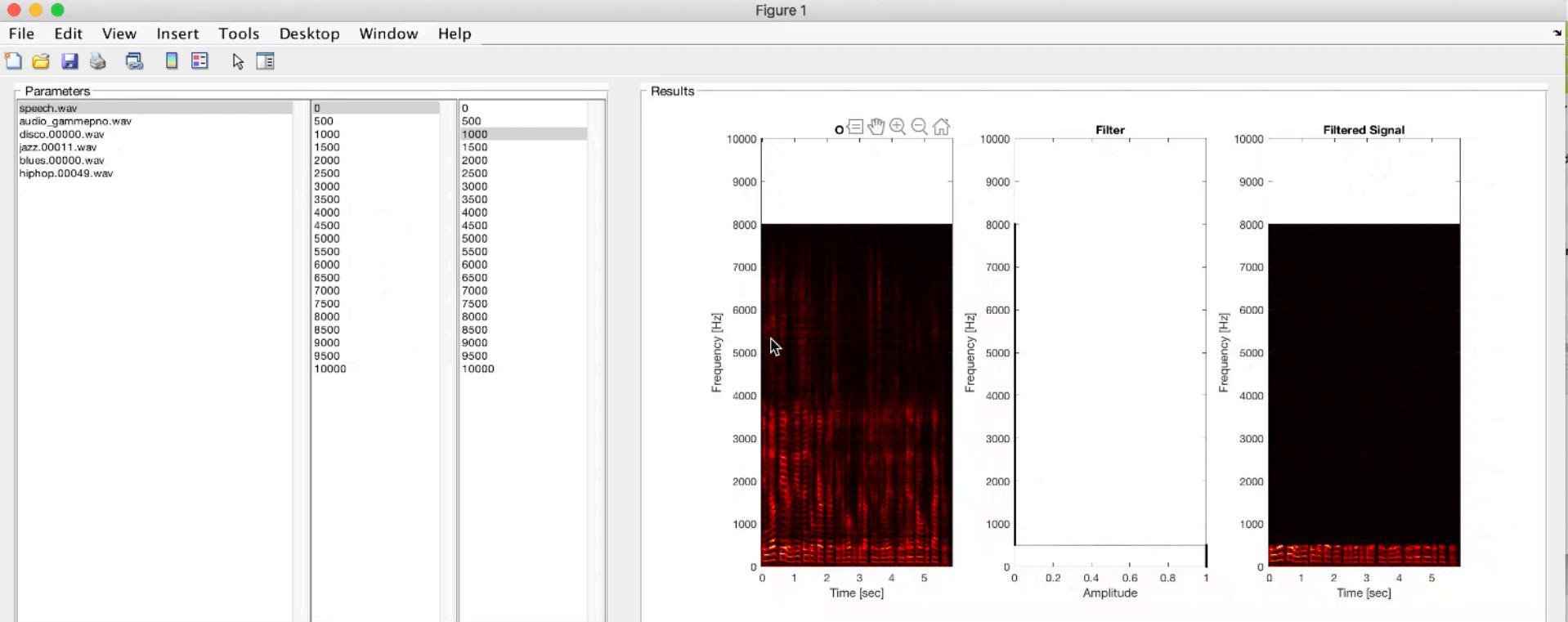
Filtrage constant au cours du temps



Application

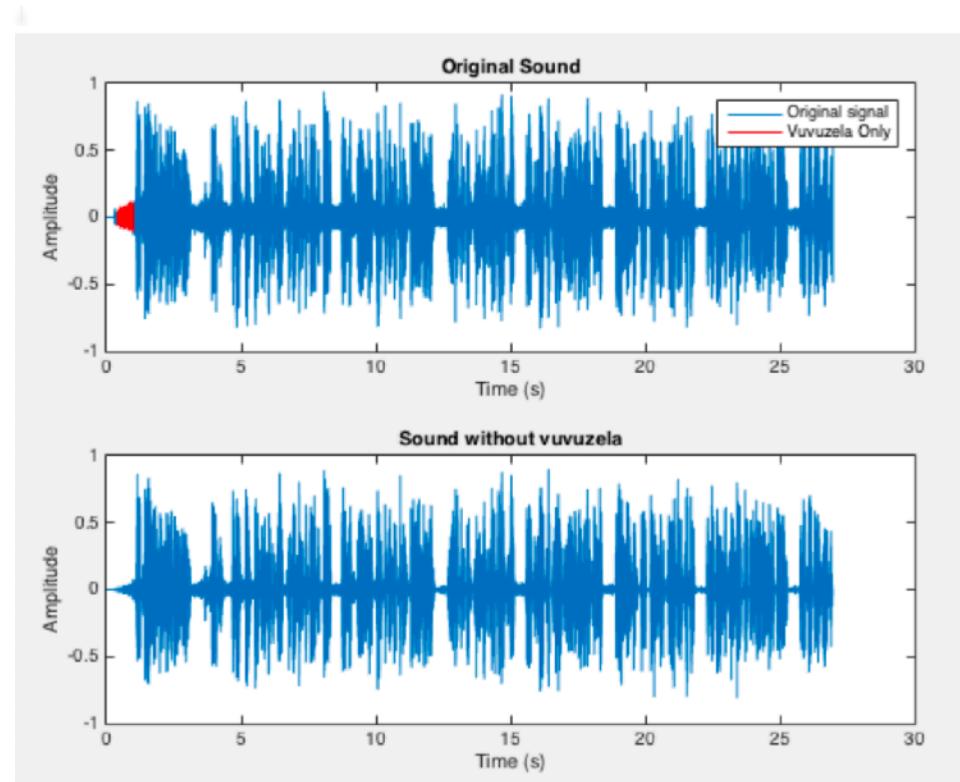
Filtrage constant au cours du temps

Figure 1



Application Débruitage par soustraction spectrale

- Soit $x(t) = s(t) + n(t)$
 - $s(t)$ est un signal de parole
 - $n(t)$ est un bruit blanc additif
 - on peut écrire le modèle
$$X(\omega, \tau) = S(\omega, \tau) + N(\omega, \tau)$$
- **Méthode**
 - On cherche un filtre fréquentiel $H(\omega, \tau)$ permettant de retirer le bruit additif
 - Amplitude de ce filtre
 - = valeur moyenne de $|N(\omega, \tau)|^2$ calculée sur un segment T ne contenant que ce bruit
 - $\mu(\omega) = \mathbb{E}_{\tau \in T} \{ |N(\omega, \tau)| \}$
 - Phase
 - = la phase de X : $\phi_x(\omega, \tau)$



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

Application Débruitage par soustraction spectrale

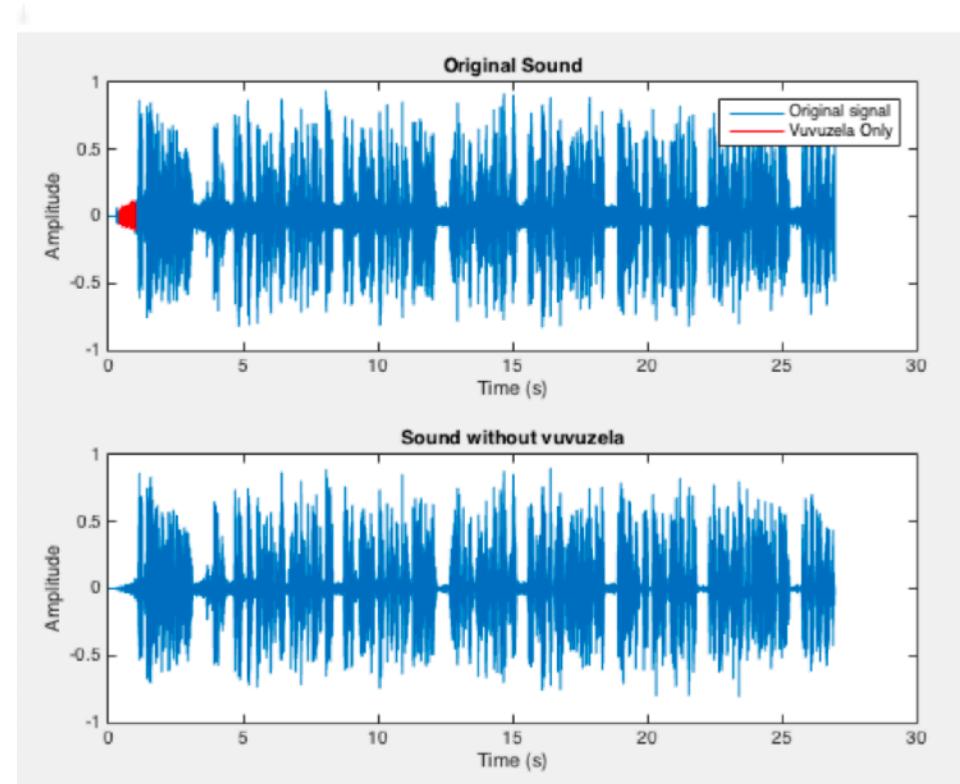
- **Soustraction**

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

- **Amélioration:**

- pour éviter des problèmes lorsque $|X(\omega, \tau)| < \mu(\omega, \tau)$
 - i.e. quand le spectre d'amplitude est < au spectre moyen du bruit
- on applique une rectification demi-onde (half-wave rectification):

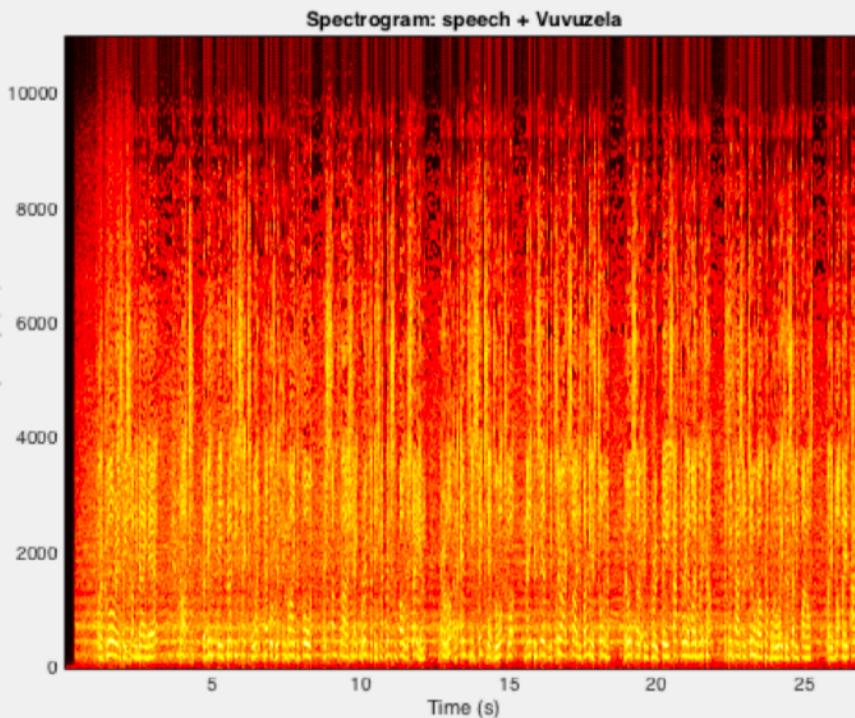
$$H_R(\omega, \tau) = \frac{H(\omega, \tau) + |H(\omega, \tau)|}{2}$$



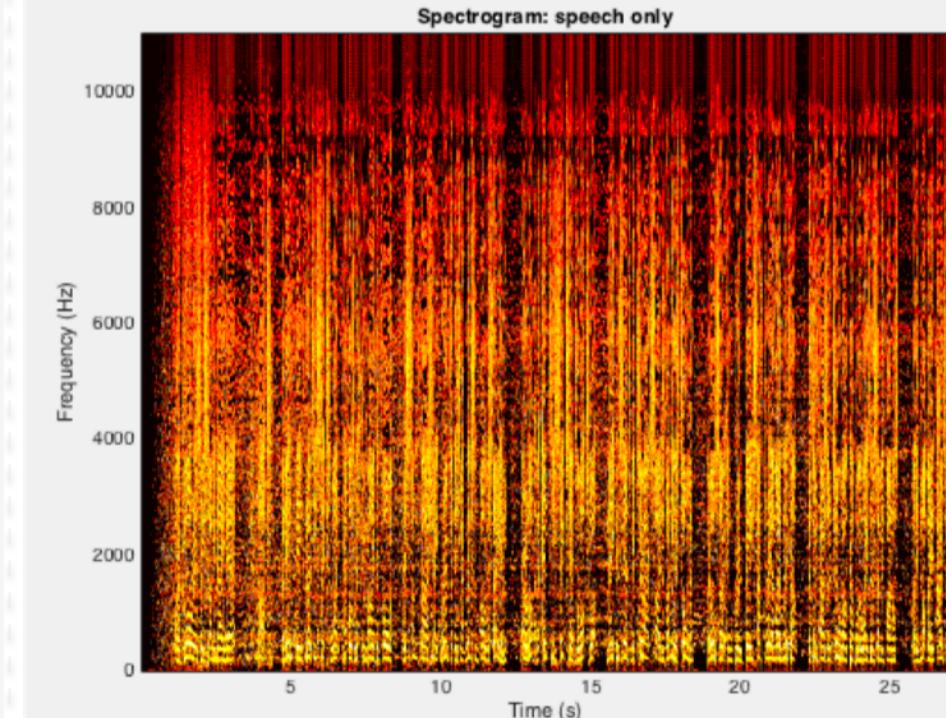
[S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. Acoustics, Speech and Signal Processing, IEEE Transactions on, 27(2) :113–120, 1979.]

Application Débruitage par soustraction spectrale

Spectrogramme speech+noise



Spectrogramme speech

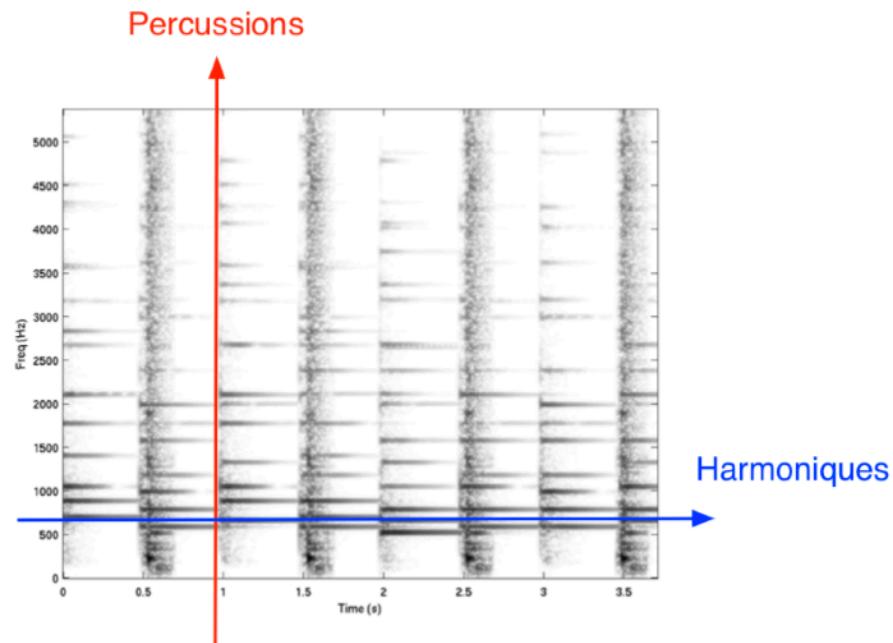


Séparation de source par TFCT

Séparation de sources

Séparation Harmonique Percussive (HPS)

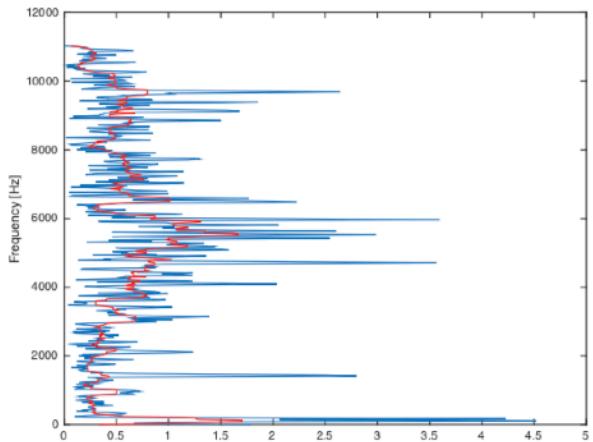
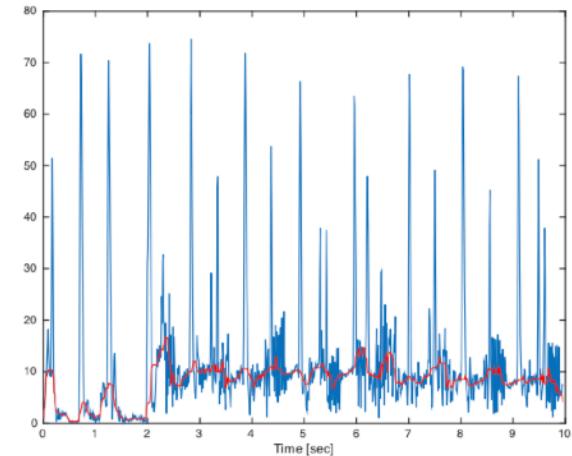
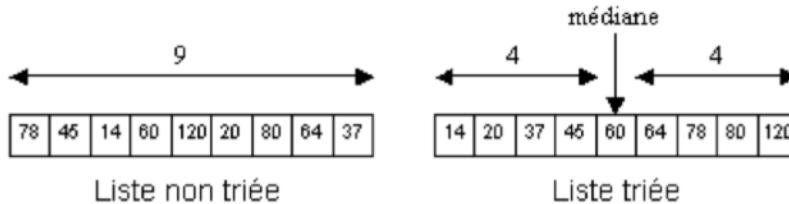
- **HPSS: Harmonic-Percussive Source Separation**
 - séparation de la partie harmonique et percussive d'un morceau de musique
- On considère la TFCT $X(f, n)$ comme le résultat de l'addition de composantes harmoniques $H(f, n)$ et percussives $P(f, n)$
 - $X(f, n) = H(f, n) + P(f, n)$
- Morphologie en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



Séparation de sources

Séparation Harmonique Percussive (HPS)

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



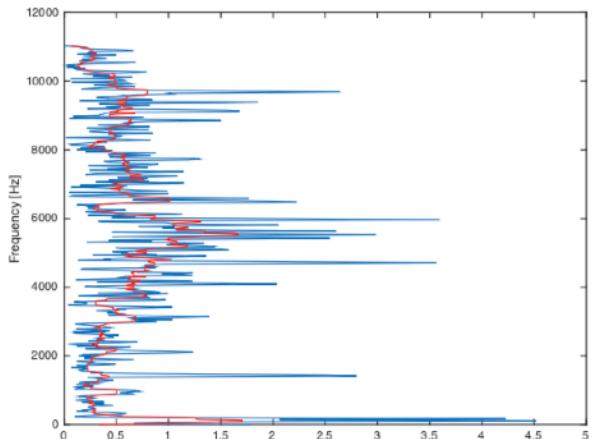
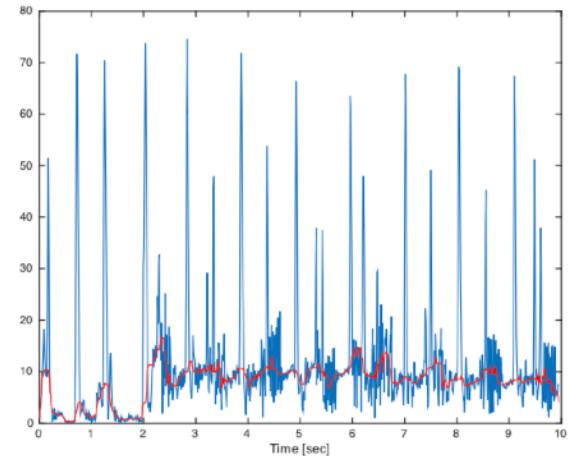
Séparation de sources

Séparation Harmonique Percussive (HPS)

- Création d'un **masque harmonique**
 - $M_H(f, n) = \begin{cases} 1 & \text{si } H(f, n) > P(f, n) \\ 0 & \text{sinon} \end{cases}$
- Création d'un **masque percussif**
 - $M_P(f, n) = \begin{cases} 1 & \text{si } P(f, n) > H(f, n) \\ 0 & \text{sinon} \end{cases}$
- **Re-création de la TFCT**

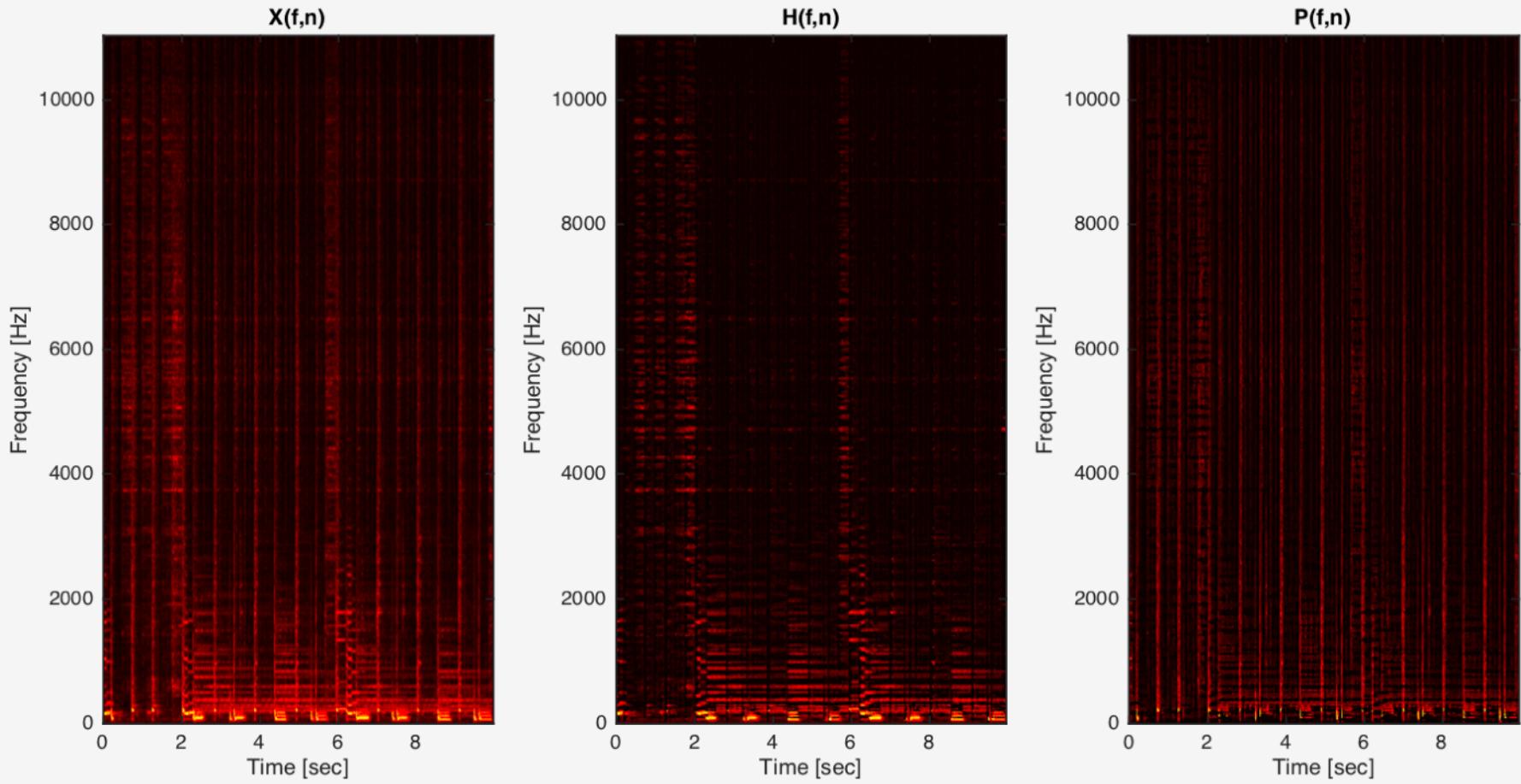
$$H(f, n) = X(f, n) \cdot M_H(f, n)$$

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



Séparation de sources

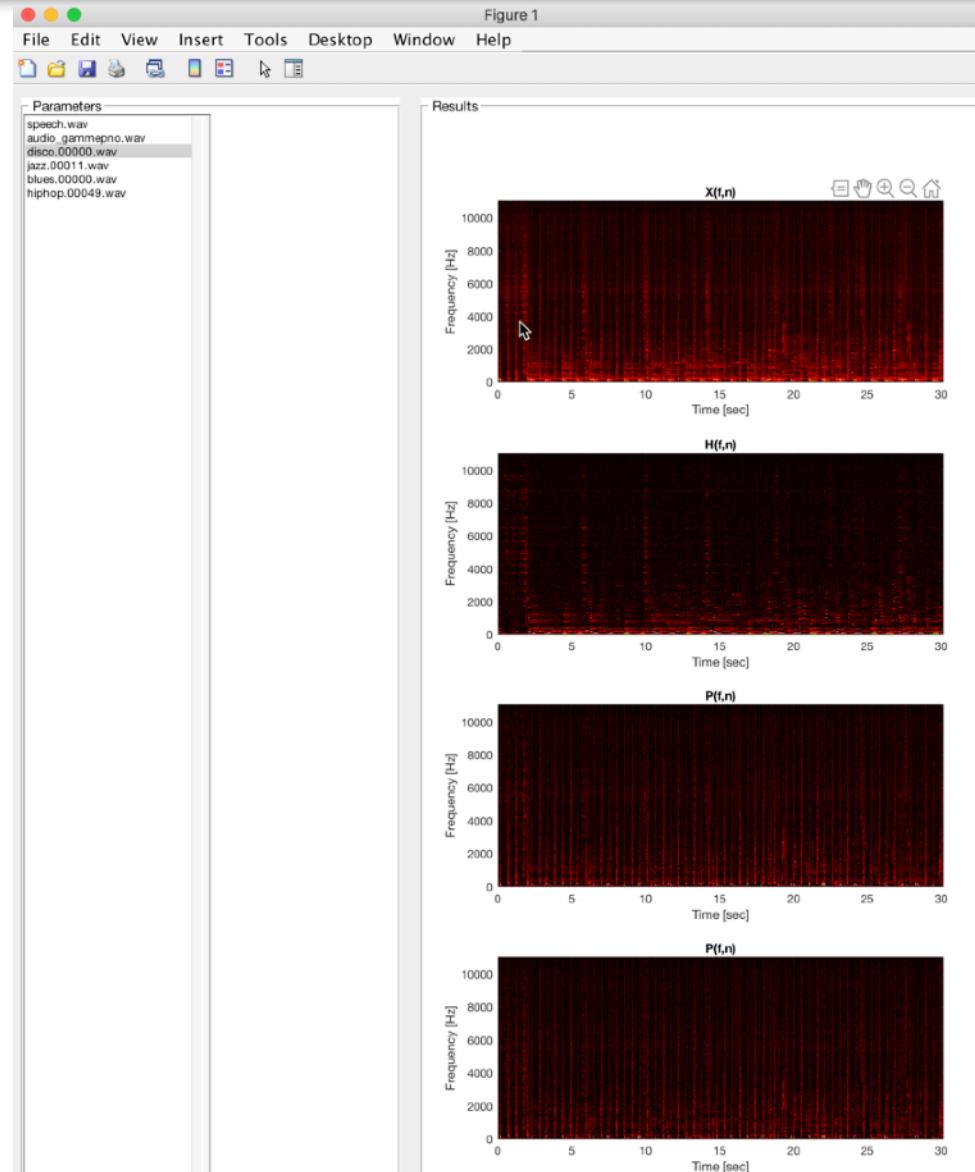
Séparation Harmonique Percussive (HPS)



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

Séparation de sources

Séparation Harmonique Percussive (HPS)



Vocodeur de phase

Vocodeur de phase

Dilatation/Contraction du temps

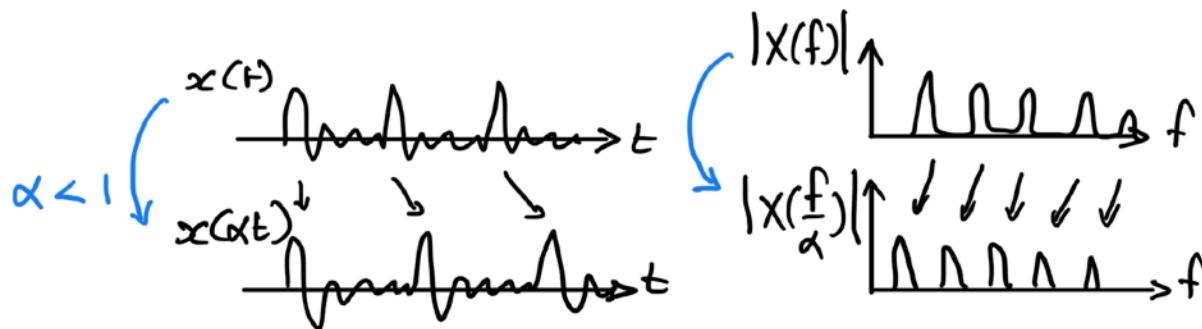
- **Technique de DJ pour changer le tempo**

- Ralentir/accélérer la vitesse de lecture (du vinyle, de la bande magnétique)

$$x(at) \Leftrightarrow \frac{1}{|a|} X\left(\frac{f}{|a|}\right)$$

- Ralentir le temps: $a < 1$
 - mais contracte aussi les fréquences (on abaisse les hauteurs)
 - Accélérer le temps: $a > 1$
 - mais étend aussi les fréquences (on augmente les hauteurs)

| Propriétés | $x(t)$ | $X(f)$ |
|-------------|--|---|
| Similitude | $x(at)$ | $\frac{1}{ a } X\left(\frac{f}{ a }\right)$ |
| Linéarité | $ax(t) + by(t)$ | $aX(f) + bY(f)$ |
| Translation | $x(t - t_0)$ | $X(f) \exp(-j2\pi f t_0)$ |
| Modulation | $x(t) \exp(j2\pi f_0 t)$ | $X(f - f_0)$ |
| Convolution | $x(t) * y(t)$ | $X(f)Y(f)$ |
| Produit | $x(t)y(t)$ | $X(f) \otimes Y(f)$ |
| Parité | <ul style="list-style-type: none">réelle paireréelle impaireimaginaires paireimaginaires impairecomplexe pairecomplexe impaireréelle | <ul style="list-style-type: none">réelle paireimaginaires paireimaginaires paireréelle impairecomplexe pairecomplexe impaire$X(f) = X^*(-f)$$\Re(X(f))$ est paire$\Im(X(f))$ est impaire$X^*(f)$ |
| | $x^*(t)$ | |

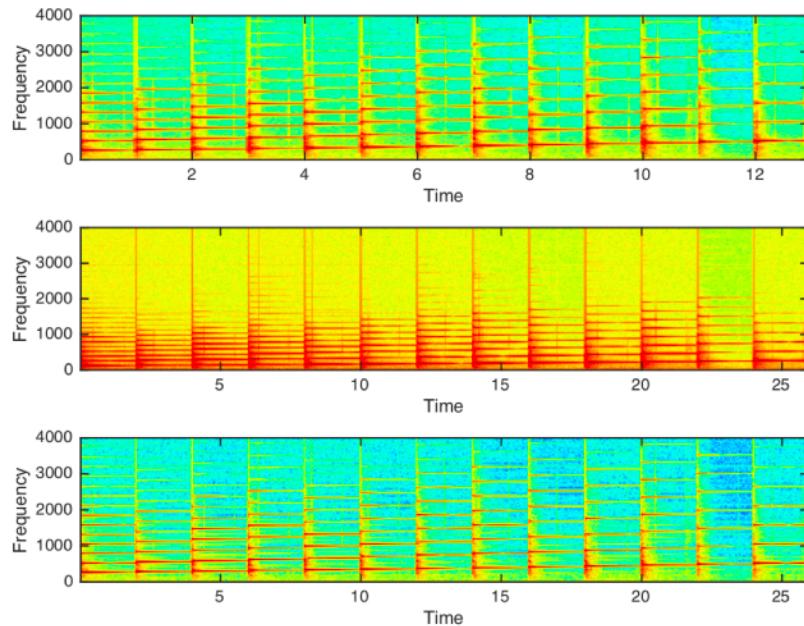


- **Objectif du vocodeur de phase**

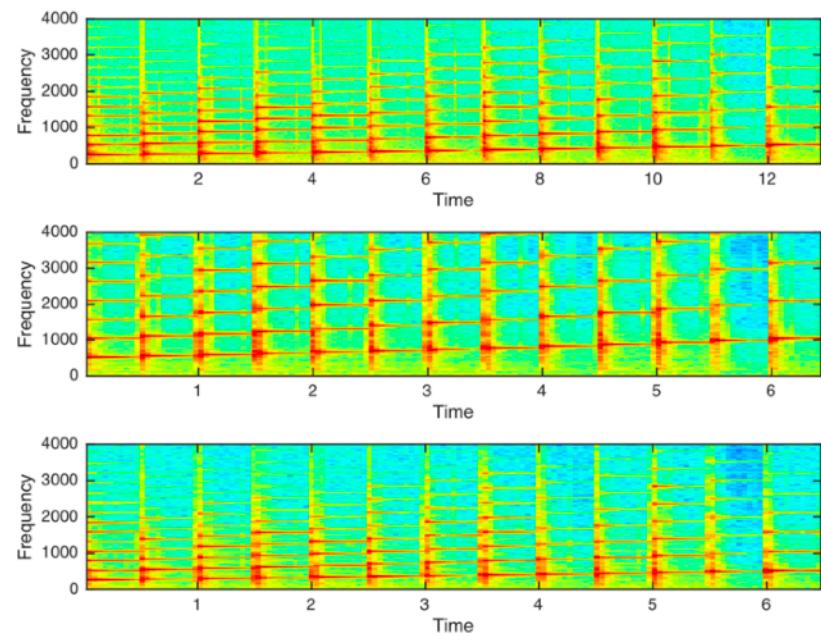
- Changer le temps et les hauteurs de manière **indépendante**

Vocodeur de phase

Dilatation/Contraction du temps



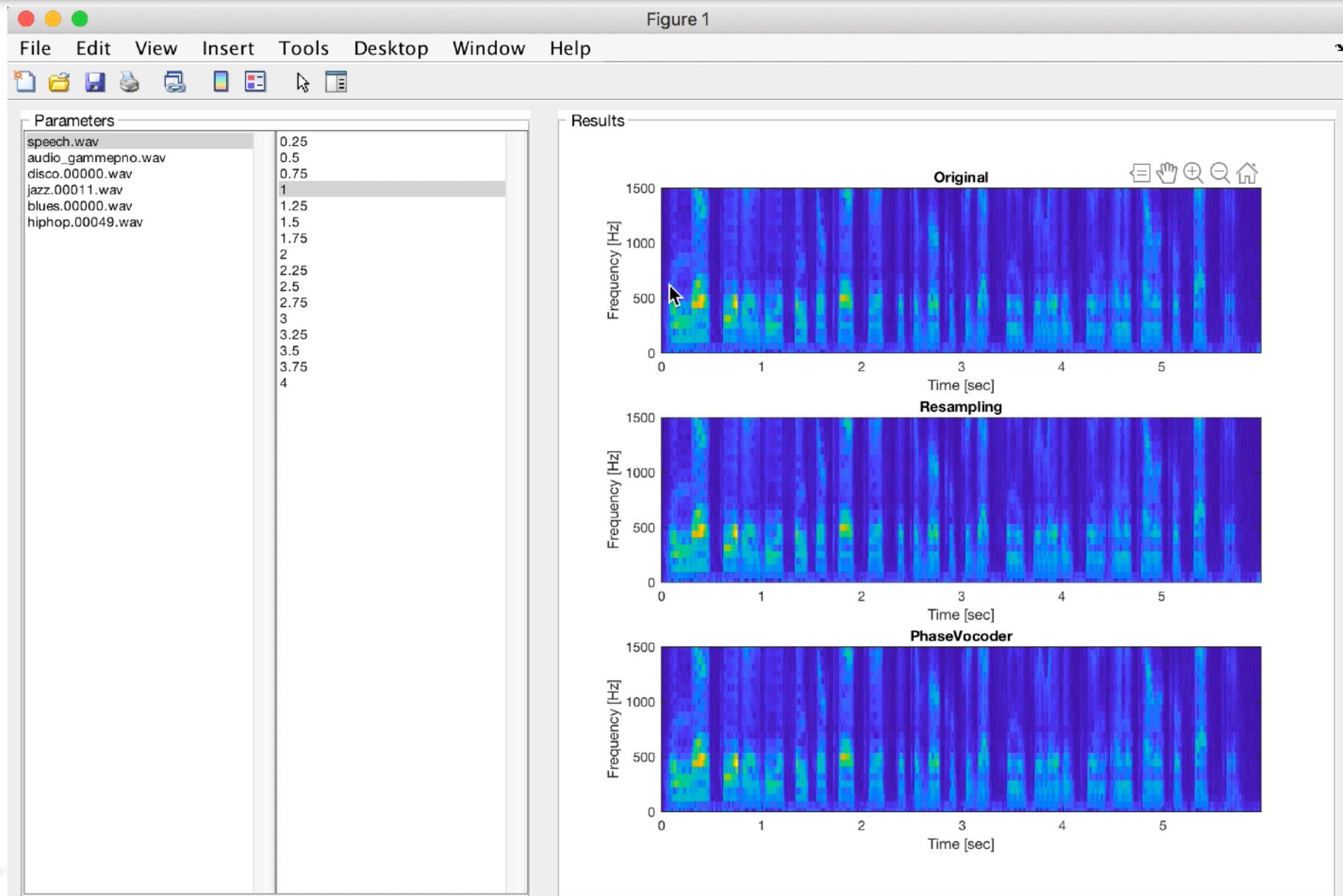
[haut] : signal original, **[milieu]** $\alpha < 1$ par ré-échantillonnage, **[bas]** : $\alpha < 1$ par vocodeur de phase



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

Vocodeur de phase

Dilatation/Contraction du temps

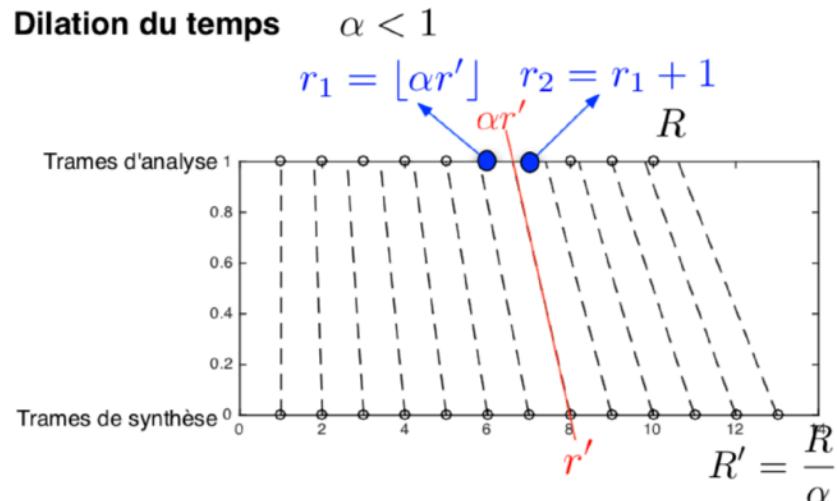


Vocodeur de phase

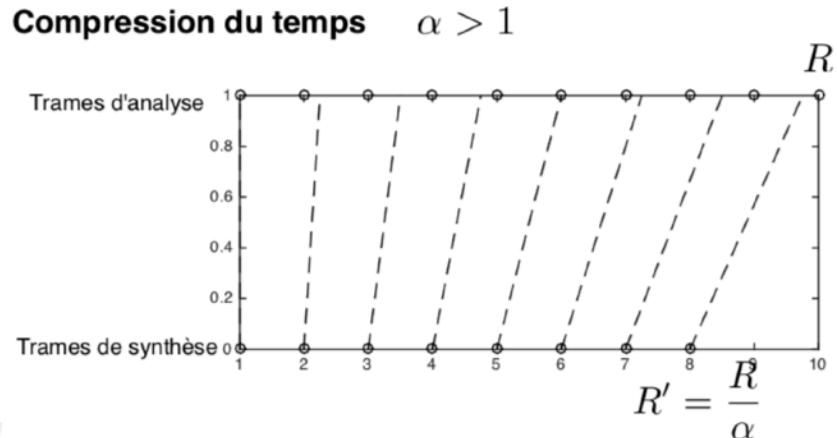
Dilatation/Contraction du temps

- Pour rallonger/raccourcir le signal:
 - on change le nombre de trames utilisées pour la re-synthèse par TFCT-1
 - Soit R :
 - le nombre de trames d'**analyse** de la TFCT
 - Soit R' :
 - le nombre de trames de **synthèse** (utilisées pour la re-synthèse par TFCT-1)
 - $R' = \frac{R}{\alpha}$
 - $\alpha < 1 \rightarrow$ on dilate (ralentit) le temps
 - $\alpha > 1 \rightarrow$ on compresse (accélère) le temps
 - Le contenu d'une trame de synthèse $r' \in \{1, \dots, R'\}$ est obtenu en recherchant les trames d'analyse r correspondantes les plus proches
 - $r_1 = \lfloor \alpha r' \rfloor$ et
 - $r_2 = r_1 + 1$

Dilation du temps



Compression du temps

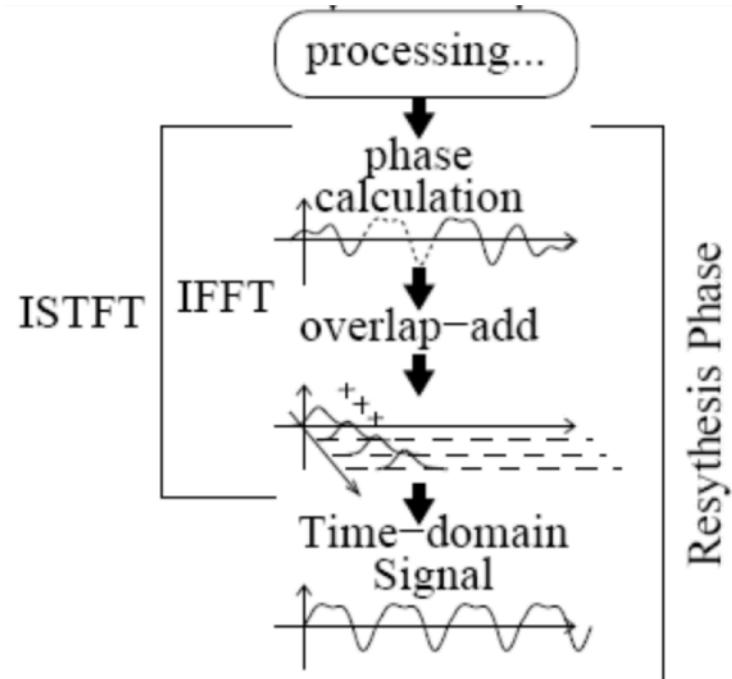
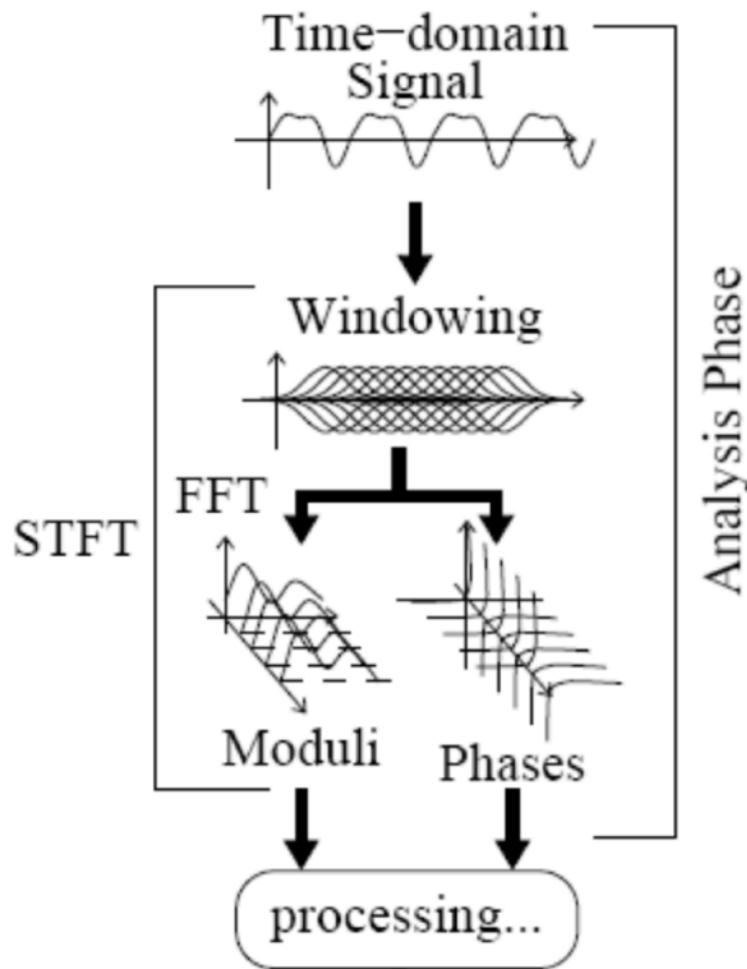


Vocodeur de phase

Dilatation/Contraction du temps

Analyse

Synthèse



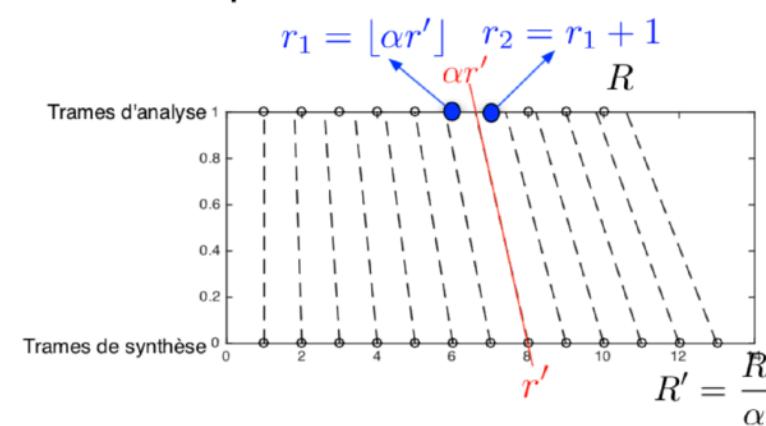
Vocodeur de phase

Dilatation/Contraction du temps

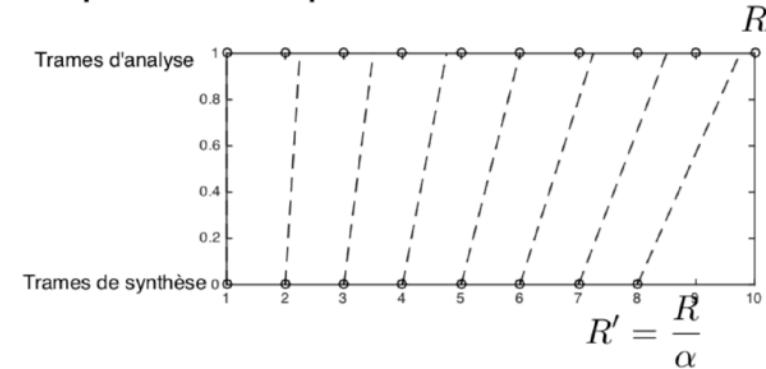
- **Manipulation du spectre d'amplitude**

- Le spectre d'amplitude à la trame r' , est obtenu par **interpolation linéaire** des spectres d'amplitude en r_1 et $r_2 = r_1 + 1$:
 - $A(k, r') = (1 - \Delta)A(k, r_1) + \Delta A(k, r_2)$
 - avec $\Delta = \alpha r' - r_1$

Dilation du temps $\alpha < 1$



Compression du temps $\alpha > 1$

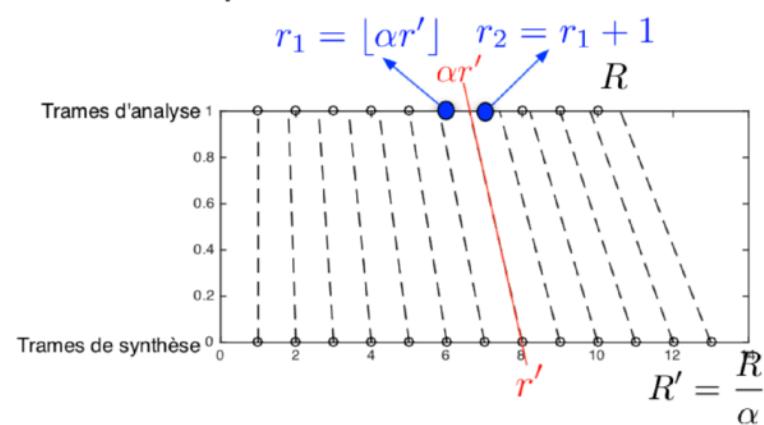


Vocodeur de phase

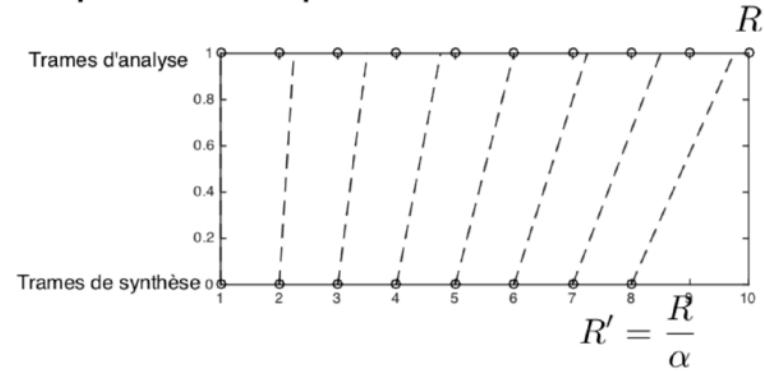
Dilatation/Contraction du temps

- **Manipulation du spectre de phase**
 - C'est plus compliqué !!!

Dilation du temps $\alpha < 1$



Compression du temps $\alpha > 1$



Vocodeur de phase

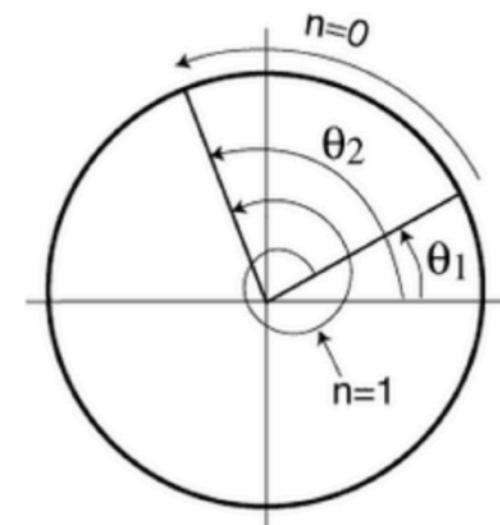
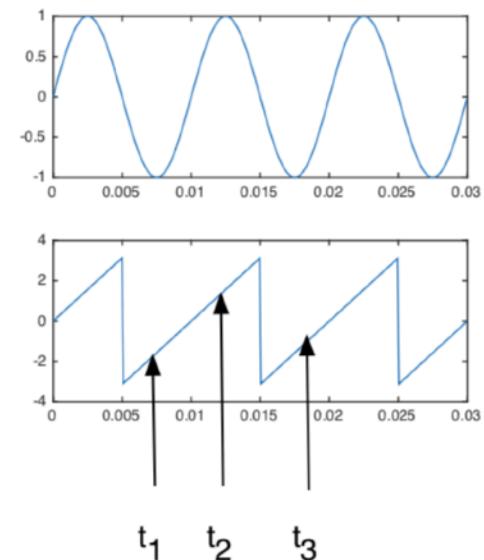
Dilatation/Contraction du temps

- **La phase et la fréquence instantanée**

- Considérons un signal formé d'une sinusoïde à la fréquence f_0 :
 - $x(t) = \sin(2\pi f_0 t) = \sin(\phi(t))$
- Entre les instants t_1 et t_2 , sa phase a ``tournée'' de $\phi(t_1)$ à $\phi(t_2)$
- Puisqu'il s'agit d'une sinusoïde pure, elle a tournée de
 - $\phi(t_2) = \phi(t_1) + 2\pi f_0(t_2 - t_1)$
- On peut donc estimer f_0 à partir de la différence de phase
 - $f_0 = \frac{\phi(t_2) - \phi(t_1)}{2\pi(t_2 - t_1)}$

- **Problème:**

- la phase est uniquement définie dans l'intervalle $[-\pi, \pi]$
 - $\phi(t) \in [-\pi, \pi]$
- donc en pratique le $\hat{\phi}(t_2)$ qu'on observe n'est pas $\phi(t_2)$ mais
 - $\hat{\phi}(t_2) = \phi(t_2) + n2\pi = \phi(t_1) + 2\pi f_0(t_2 - t_1) + n2\pi$
 - avec $n \in \mathbb{N}$ indéterminé
- pour estimer f_0 il faut donc déterminer n
 - $f_0 = \frac{\phi(t_2) + n2\pi - \phi(t_1)}{2\pi(t_2 - t_1)}$



Vocodeur de phase

Dilatation/Contraction du temps

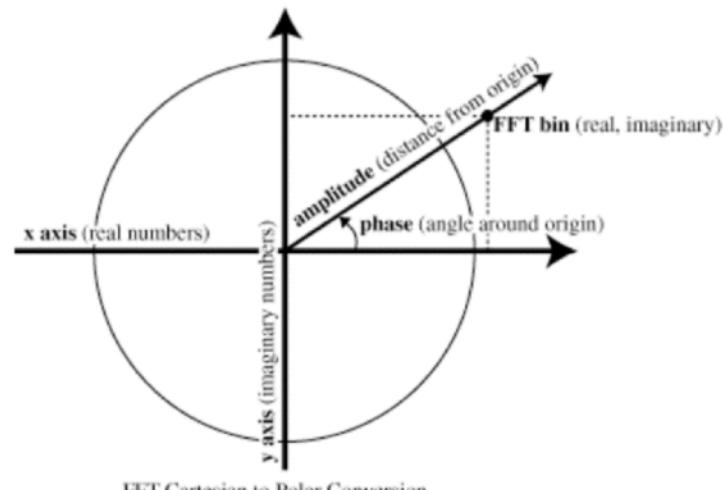
- **La phase dans le Transformée de Fourier à Court Term (TFCT)**

- Pour chaque trame n et fréquence k la TFTC est un nombre complexe

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

- Il peut se décomposer en amplitude (module) et phase:

- $$X(k, n) = A(k, n) e^{j\phi(k, n)}$$

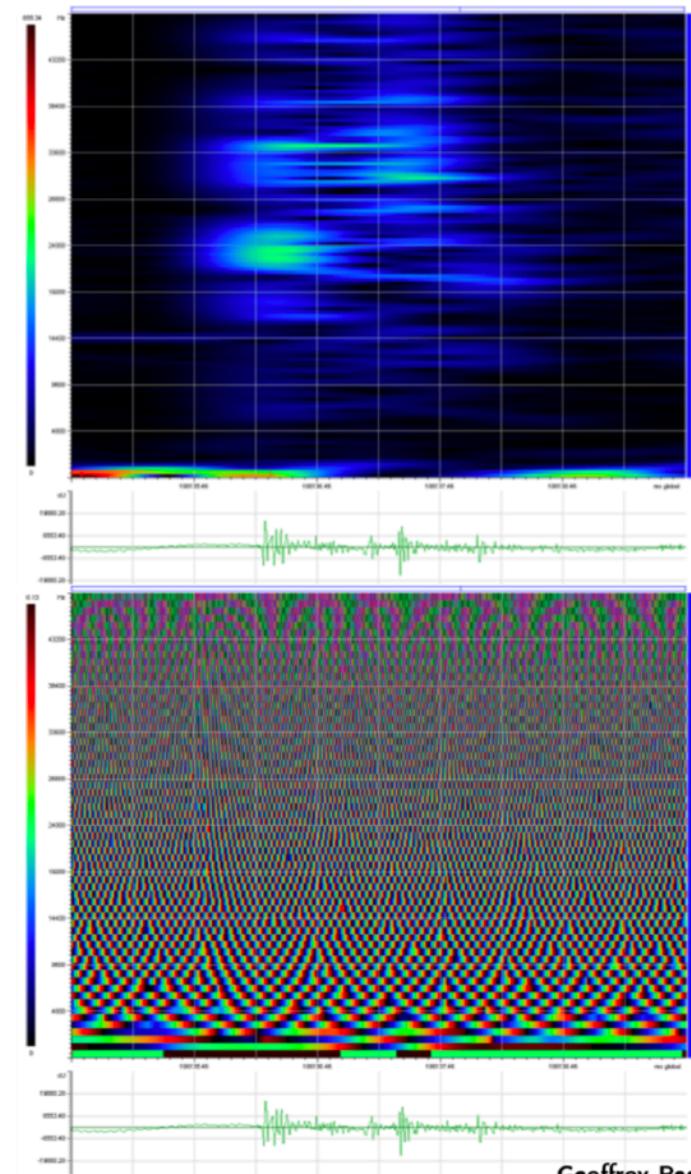


FFT Cartesian to Polar Conversion

Vocodeur de phase

Dilatation/Contraction du temps

- On a donc une valeur d'amplitude et de phase pour chaque (k, n)
- Spectrogramme
 - d'amplitude $A(k, n)$
 - de phase $\phi(k, n)$
- La phase indique la position de la co-sinusoïde,
- Pour une fréquence k donnée, la variation temporelle de phase indique la fréquence instantanée *observée à travers le filtre de la TFCT centré sur k*
 - Pour chaque fréquence k , et chaque couple de trames successives $(n - 1) \rightarrow n$, on peut donc calculer une fréquence instantanée

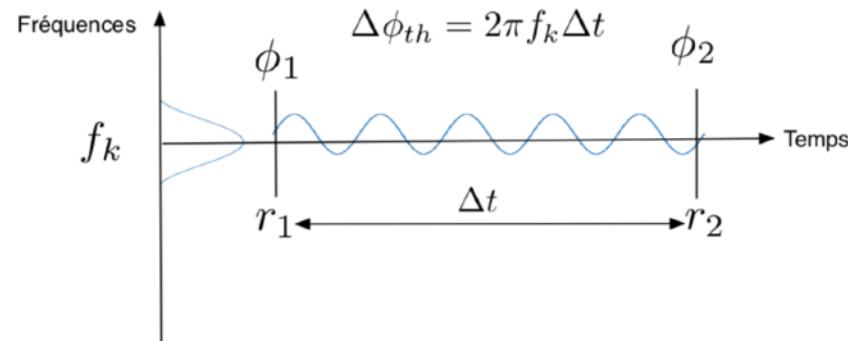


Vocodeur de phase

Dilatation/Contraction du temps

- **Manipulation du spectre de phase**

- A la trame r' le spectre de phase dans le filtre k de la TFCT est obtenu en propageant la phase à partir de la fréquence contenu dans ce filtre



- **1) Solution fausse:**

- On suppose qu'à travers le filtre k de la TFCT on observe une sinusoïde à la fréquence f_k

- si c'était le cas, on propagerait l'évolution de la phase au cours du temps en utilisant la prédition théorique de la phase à la fréquence f_k :

$$\Delta\phi_{th} = 2\pi f_k \Delta t$$

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

- avec comme phase **initiale**:

$$\bullet \phi(k, r' = 1) = \phi(k, r = 1)$$

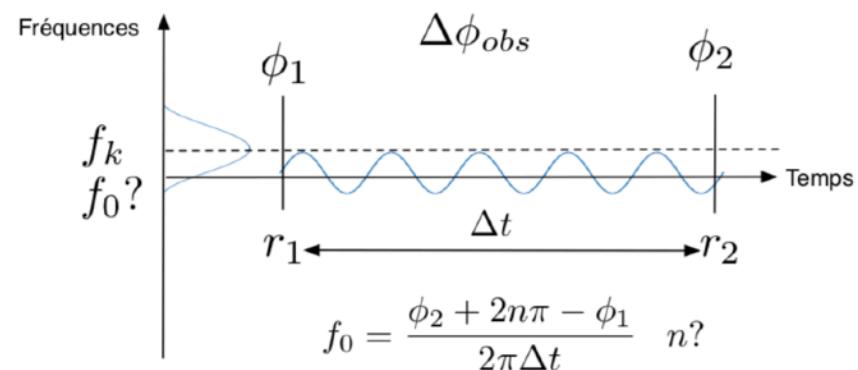
- En réalité, dans le filtre k on peut observer

- une sinusoïde dans le filtre mais à une fréquence $f_0 \simeq f_k$
 - une sinusoïde dans un filtre voisin ($k - 1, k + 1$)
 - ceci est du à la largeur du lobe, les lobes secondaires

Vocodeur de phase

Dilatation/Contraction du temps

- **Manipulation du spectre de phase**
 - A la trame r' le spectre de phase dans le filtre k de la TFCT est obtenu en propageant la phase à partir de la fréquence contenu dans ce filtre
- **2) Solution correcte:**
 - En pratique, à travers le filtre k de la TFCT, on peut observer des sinusoïdes à des fréquences proches mais différentes de f_k
 - ceci est du à largeur du lobe principale, aux lobes secondaires
 - Il faut donc d'abord **estimer cette fréquence** f_0 et ensuite appliquer la propagation de phase correspondante
 - $\phi(k, r') = \phi(k, r' - 1) + 2\pi f_0 \Delta t$



Vocodeur de phase

Dilatation/Contraction du temps

- **Comment estimer cette fréquence f_0 ?**

- En utilisant la fréquence instantanée:

- $$f_0(n) = \frac{\phi_2 + n2\pi - \phi_1}{2\pi\Delta t}$$

- **Oui mais comment déterminer n ?**

- on cherche n tel que $f_0 \simeq f_k$

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

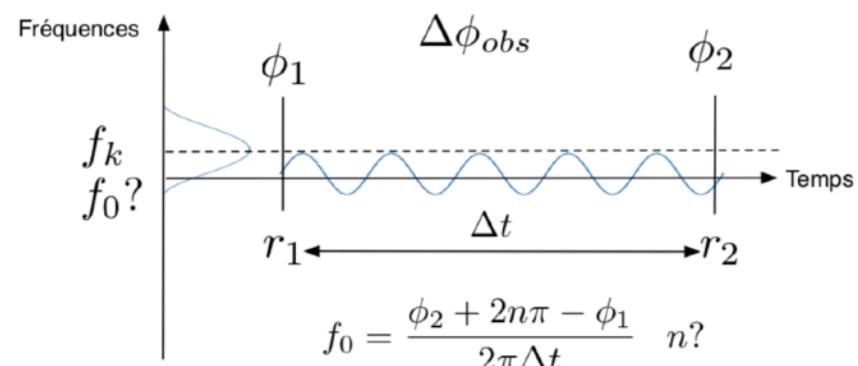
$$\min_n \left| \frac{\phi_2 + n2\pi - \phi_1}{2\pi\Delta t} - f_k \right|$$

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

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

- ce qui revient à

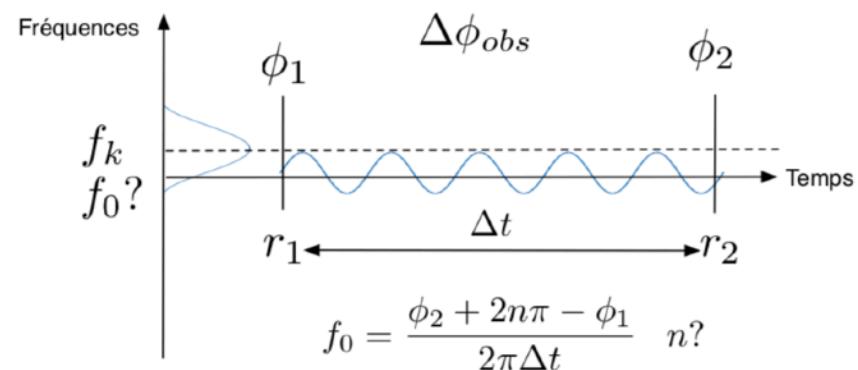
- trouver la détermination principale (la valeur dans l'intervalle $[-\pi, \pi]$) de
 - $n = \lfloor (\phi_2 - \phi_1 - \Delta\phi_{th}) / (2\pi) \rfloor$
- il s'agit de la différence de phase non-expliquée par le modèle théorique $\Delta\phi_{th}$



Vocodeur de phase

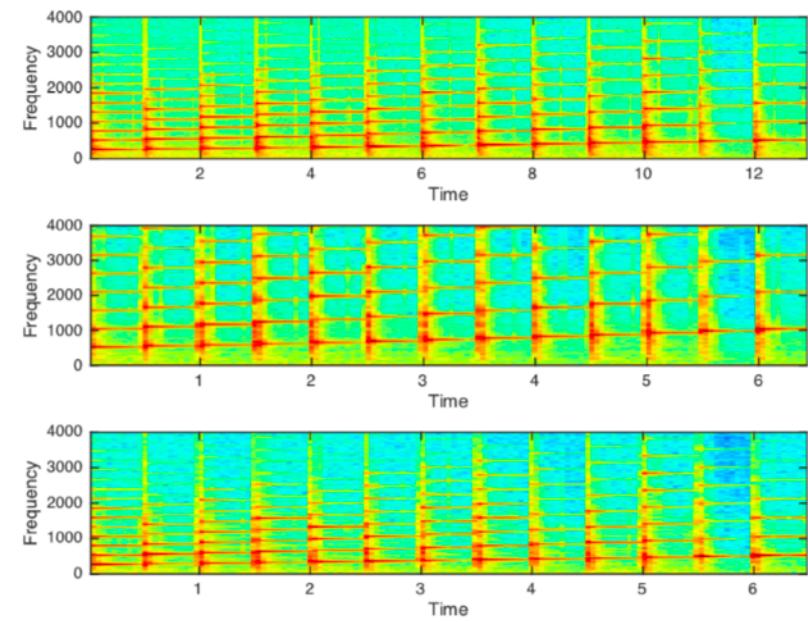
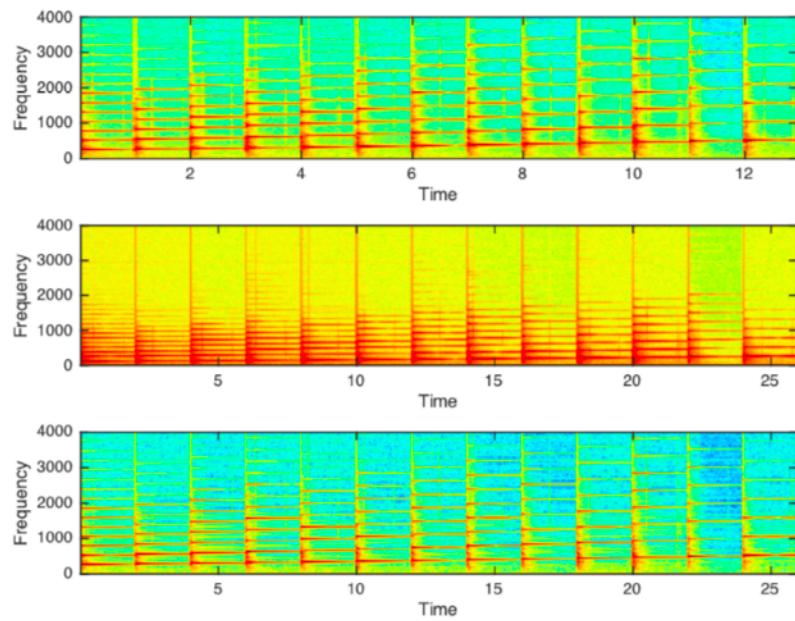
Dilatation/Contraction du temps

- **Manipulation du spectre de phase**
 - A la trame r' le spectre de phase dans le filtre k de la TFCT est obtenu en propageant la phase à partir de la fréquence contenu dans ce filtre
- **2) Solution correcte:**
 - Finalement la phase est incrémentée de
$$\begin{aligned}\phi(k, r') &= \phi(k, r' - 1) + 2\pi f_0 \Delta t \\ &= \phi(k, r' - 1) + \phi_2 + n2\pi - \phi_1\end{aligned}$$
 - avec comme phase **initiale**:
 - $\phi(k, r' = 1) = \phi(k, r = 1)$



Vocodeur de phase

Dilatation/Contraction du temps

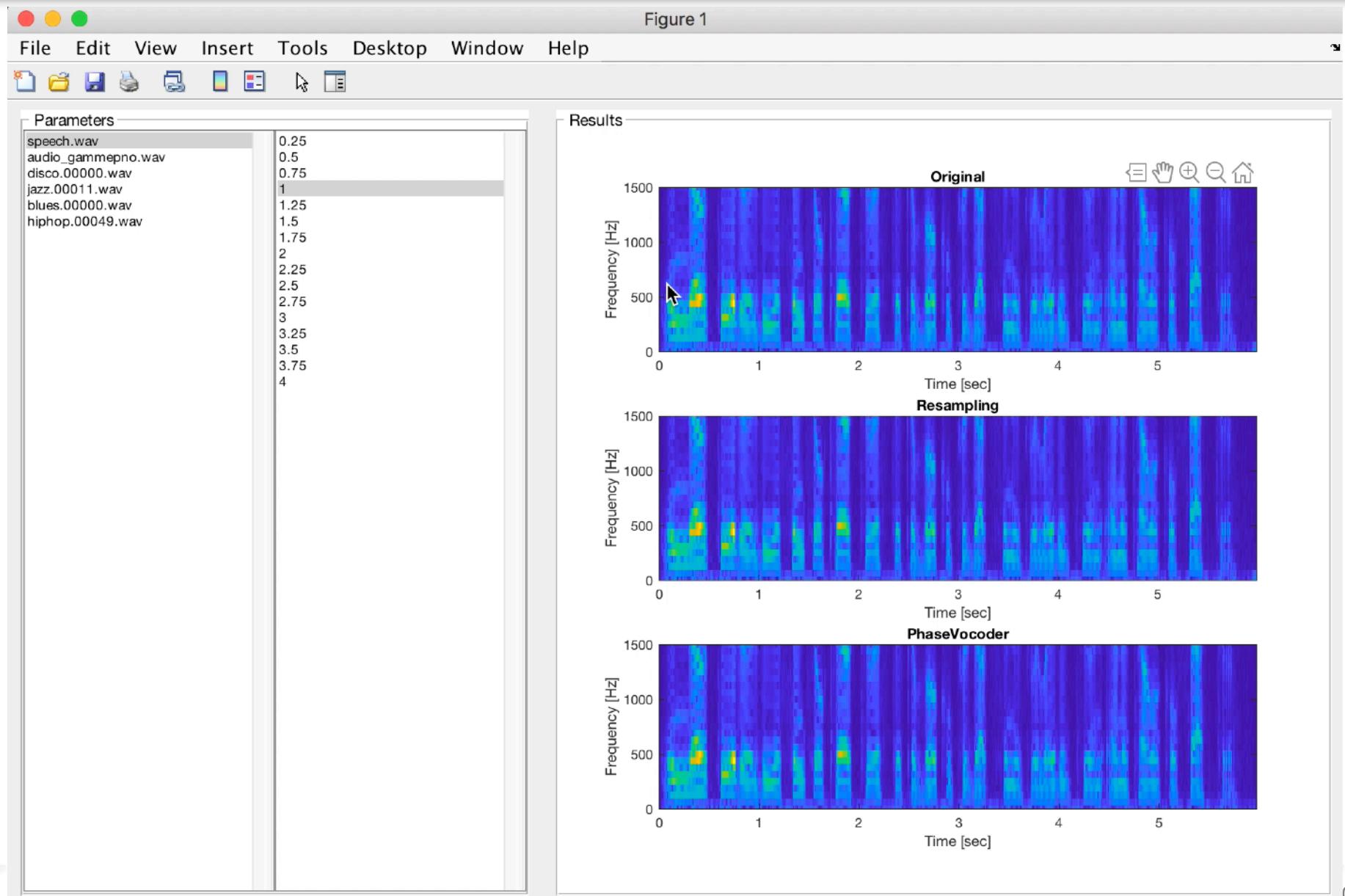


[haut] : signal original, [milieu] $a < 1$ par ré-échantillonnage, [bas] : $a < 1$ par vocodeur de phase

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

Vocodeur de phase

Dilatation/Contraction du temps



Vocodeur de phase

Dilatation/Contraction du temps

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

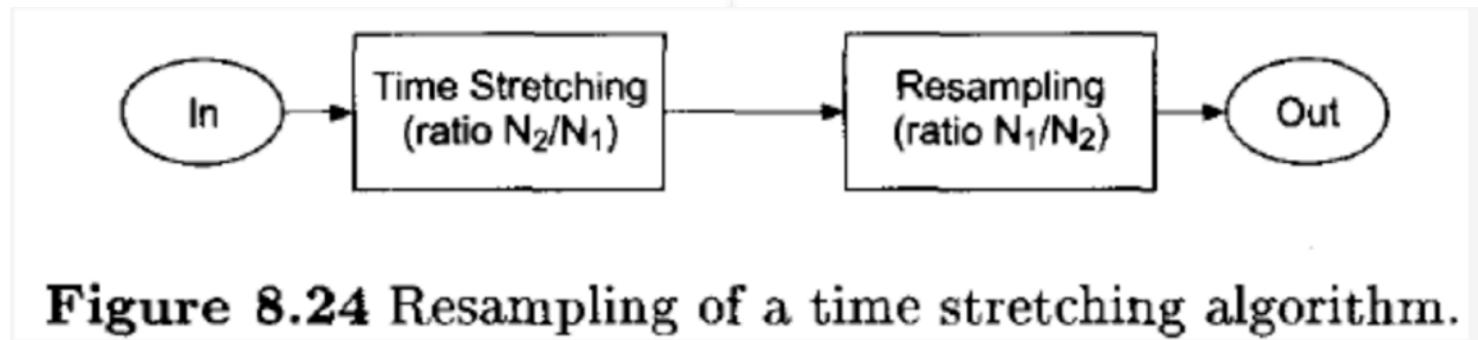


Figure 8.24 Resampling of a time stretching algorithm.

What you need to know

- Transformée de Fourier à Court Terme
 - formule, explication
 - influence des paramètres types et longueur de fenêtre
 - interprétation passe-bande, passe-bas
- Reconstruction du signal à partir de la TFCT
 - sans modifications: addition/ recouvrement (OLA): principe et formule
 - avec modifications: algorithme de Griffin & Lim: principe et formule
- Applications du traitement par TFCT
 - débruitage par soustraction spectrale: principe et formule
 - vocodeur de phase: principe et formule
 - fréquence instantanée
 - détermination du paramètre de unwrapping