Recent advances in Digital Music Processing and Indexing

Acoustics'08 warm-up

TELECOM ParisTech

Gaël RICHARD
Telecom ParisTech (ENST)
www.enst.fr/~grichard/

Content

- **Introduction and Applications**

- **Components of an Audio indexing system**
  - Architecture
  - Features extraction and selection
  - Classification : the example of automatic musical instrument recognition

- **Signal Decomposition and source separation**
  - An alternative to Bag of frames approaches
  - Non-Negative Matrix decomposition
  - Main melody estimation and extraction

- **Other audio indexing examples**
  - *Drum Processing :*on combining tempo, source separation and transcription
  - *Drum loop Retrieval:* on combining Speech recognition and transcription
  - *"Cross-modal" Retrieval:* On combining visual and audio correlations for audio-based music video retrieval (or video-based audio retrieval).

- **(The evaluation problem)**
- **Conclusion**

Introduction

- **Audio is now widely available in digital format and in huge databases**

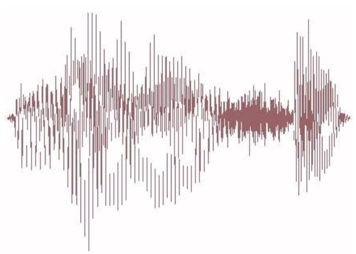
  - New means of interaction with audio information are possible and desirable

  - Strong needs for efficient audio indexing techniques to automatically obtain a detailed and **meaningful symbolic representation**

Introduction

- **For music signals ?**

Music signal



+ emotions, playing style, performers,…

Search by content .....?

Applications: for content providers

- Ease composition
  - Search/Retrieval of sound samples or drum loops in large databases of sounds (drum loops retrieval)
  - Content-aware Musical edition software
  - Automatic musical analysis for expressive music synthesis
  - Musical "Oracle" : predict which kind of public will like a given piece.
  - Hit predictor…
  - Search/Retrieval of video scenes from Audio (use of multimodal correlations)

Drum loops retrieval reference:

O. Gillet and G. Richard , « *Drum loops retrieval from spoken queries* », Journal of Intelligent Information Systems, 24:2/3, pp 159-177, Springer Science, 2005

Applications: for broadcasters

- Ease radio program set up
  - Navigation interfaces in large data collections
  - Automatic play lists (podcast,…)
  - Mixing/Remixing/DJing : Tempo, rhythm, texture synchronization (*Tempo*)

- Identify what is broadcasted
  - Plagiarism detection.
  - Broadcast monitoring.

Tempo extraction reference:

M. Alonso, G. Richard and B. David, "Accurate tempo estimation based on harmonic+noise decomposition", *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 82795, 14 pages, 2007.

Applications: for consumers

- Increase listening capabilities
  - Automatic analysis of user taste for music recommendation.
  - Automatic play list by emotion, genre.
  - Structured navigation in music: « skip the intro, « replay the chorus »,….)
  - Personalized listening (***Remixing***)

- Allow new usages
  - Search by content (*Query by humming*)
  - Smart Karaoke
  - Active listening.

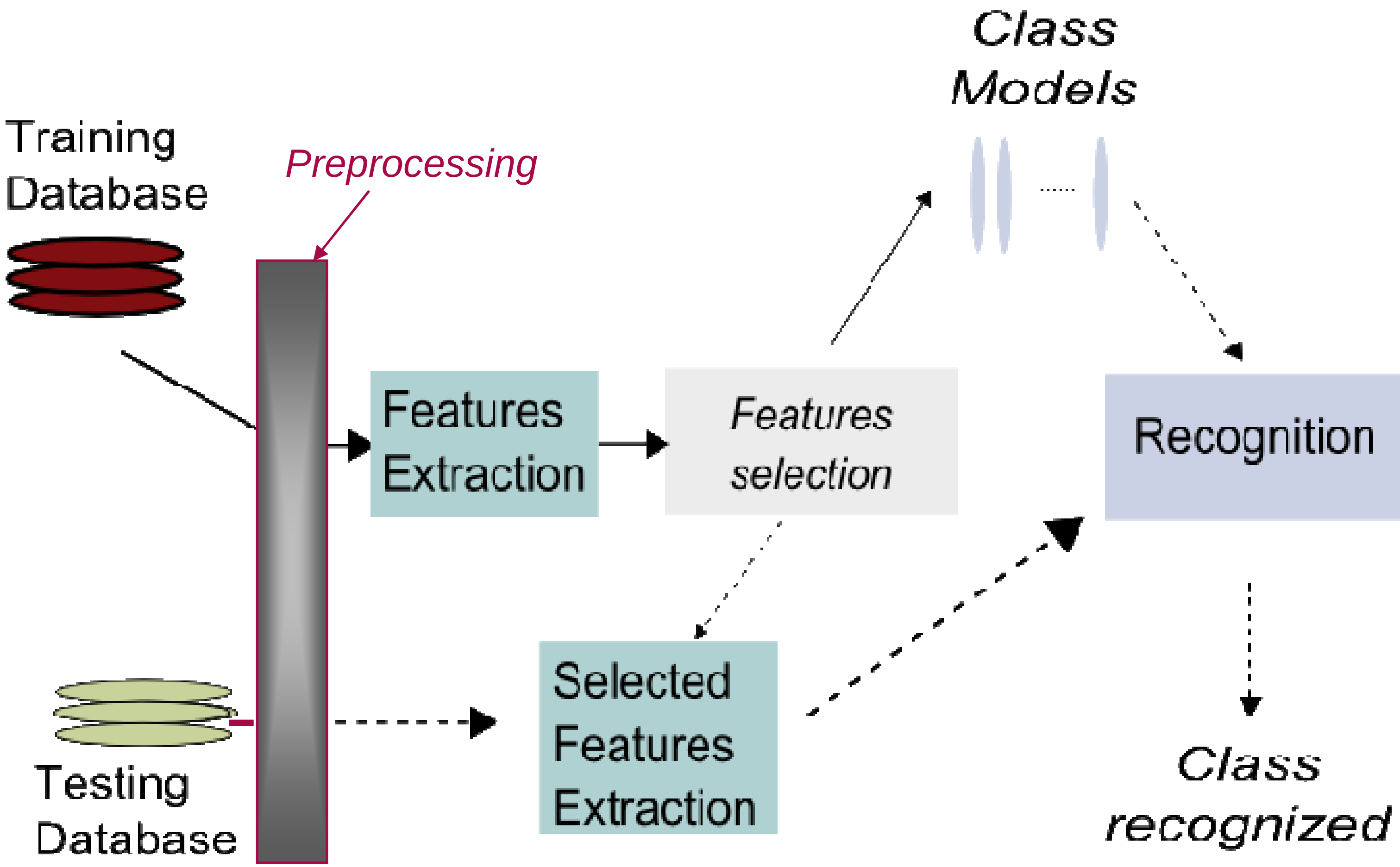
Drum Extraction and Remixing:

O. Gillet, G. Richard. *Transcription and separation of drum signals from polyphonic music*. in IEEE Trans. on Audio, Speech and Language Proc., Volume 16, N°3, March 2008 Page(s):529 - 540.

Classification systems

- **Several problems, a similar approach**
  - Automatic musical genre recognition
  - Automatic music instruments recognition.
  - Sound samples classification.
  - Sound track labeling (speech, music, special effects etc…).
  - Automatically generated Play list
  - Hit predictor...

Traditional “Bag-of-Frames” approach



Features extraction

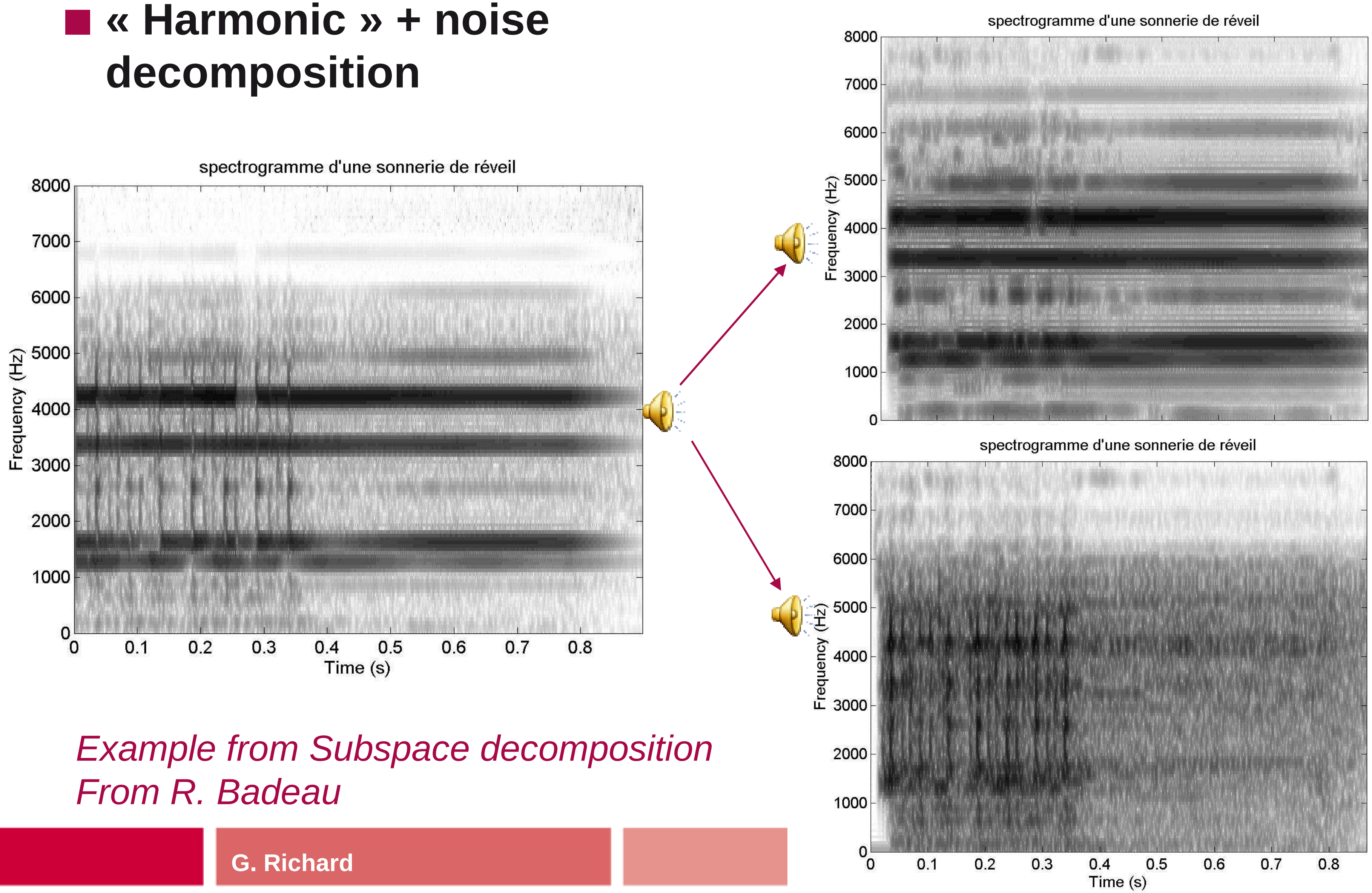
- **A need for a compact representation of the audio space using descriptors:**
  - **Temporal parameters :** *(Zero crossing rate, envelope, amplitude modulation (4 Hz, or 10-40 Hz), crest factor, signal impulsivity, tempo,…*
  - **Spectral or Cepstral parameters:** *(MFCC, LPCC, Warped LPCC, spectral centroïd, spectral slope, high order spectral moments, filter banks, harmonic to noise ratio,…)*
  - **Perceptual parameters:** (*Relative specific loudness, sharpness and spread,….)*

Most of these features described in

- *G. Peeters. "A large set of audio features for sound description (similarity and classification) in the cuidado project." Technical report, IRCAM (2004)*
- *S. Essid. "Classification automatique des signaux audio-fréquences: reconnaissance des instruments de musique", Ph.D. thesis, Univ. Paris 6 (in French).*

Feature extraction

- « Harmonic » + noise decomposition



Example from Subspace decomposition
From R. Badeau

Classification methods for automatic musical instrument recognition

- **Different strategies were proposed**
  - Early studies based on K-Nearest Neighbors (K-NN) on isolated notes and later on solos
  - **Current trend:** to use more sophisticated modelling approaches such as Discriminant analysis, neural networks, Gaussian Mixtures Models (GMM) and Support Vector Machines
  - See for a recent review:

    P. Herrera, A. Klapuri, M. Davy. Chap.6 Automatic Classification of Pitched Musical Instrument Sounds. *in Signal Processing methods for Music transcription,* Edited by A. Klapuri and M. Davy, Springer (2006)

Automatic instrument recognition in polyphonic signals

- **From isolated notes to polyphonic signals : a problem of increased complexity**

  Isolated notes → *solos* → *polyphony*

- **In fact, in most cases the methods designed for the monophonic case will not work :**
  - Features extraction is non-linear
  - The “additivity” of the sources cannot be used

Automatic instrument recognition in polyphonic signals

- **Some directions:**
  - Ignore the features perturbed by other sources (missing features theory) *(see Eggink, 2003)*
  - Apply a source separation approach before recognising the different instruments *(see Gillet 2008 for percussive instruments)*
  - Use specific models such as pitch dependent instrument models *(Kitahara, 2005)*
  - Learn a model for each combination of instruments with an automatically induced hierarchical taxonomy *(Essid, 2006)*

An alternative to “bag-of-frames” approaches

- **Sparse atomic decomposition of polyphonic music signals**

  - **Principle:** Derive an “atomic” decomposition of the audio signal.

In other words, the signal is approximated as a linear combination of atoms $h_\lambda(t)$ from a fixed dictionary:

$$x(t) = \sum_{\lambda \in \Lambda_x} \alpha_\lambda h_\lambda(t)$$

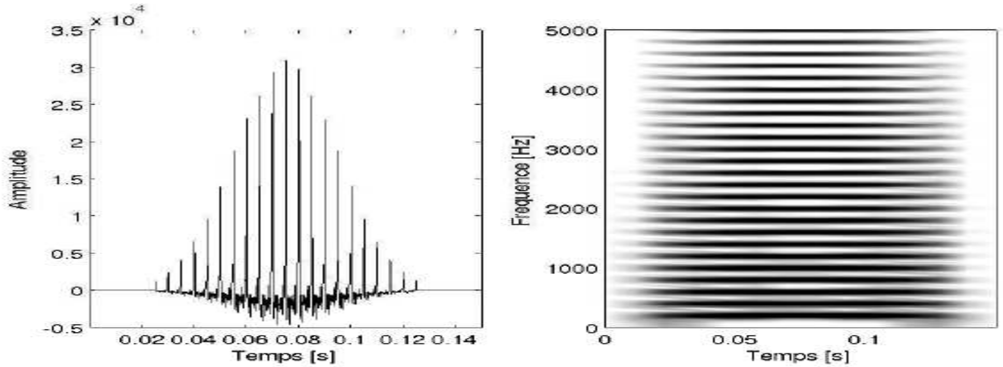
Sparse atomic decomposition of polyphonic music signals

■ **Atoms used: harmonic chirp Gabor atoms**

$$h_{s,u,f_0,c_0,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m\times f_0,m\times c_0}(t)$$

- a_m (resp ϕ_m) vector of partials amplitude (resp. phases)
- s scale parameter
- u time localization
- f_0 (resp c_0) fundamental frequency and chirp rate

(from P. Leveau, *“Décompositions parcimonieuses structurées: Application à la représentation objet de la musique”,* Ph.D. thesis, Univ. Paris 6, 2007)



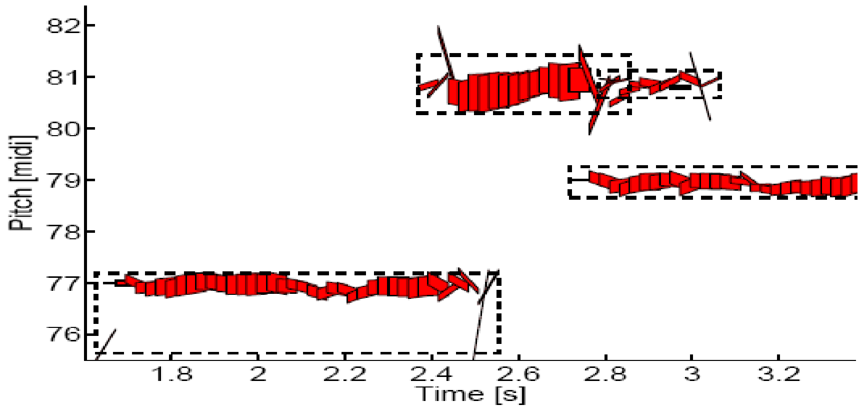
Sparse atomic decomposition of polyphonic music signals

- **The atomic decomposition is obtained by means of (for example) matching pursuit:**
  - First, the most prominent atom (*i.e.* the most correlated with the signal) is extracted and subtracted from the original signal.
  - Iterate the procedure until a predefined number of atoms have been extracted or until a pre-defined SNR of the representation is reached.

Sparse atomic decomposition of polyphonic music signals

- **For musical instrument recognition**
  - Use Instrument specific atoms
  - Obtain representatives atoms for each instrument and for each fundamental frequency
  - Use a decomposition algorithm such as matching pursuit
  - Instrument-specific harmonic “Molecule” can be obtained as a group of successive atoms

Démo



For more details, see

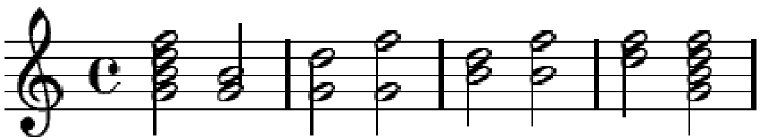
P. Leveau, E. Vincent, G. Richard, L. Daudet, “Instrument-specific harmonic atoms for mid-level music representation”, in IEEE Trans. on ASLP, Volume 16, N°1 Jan. 2008 Page(s):116 - 128.

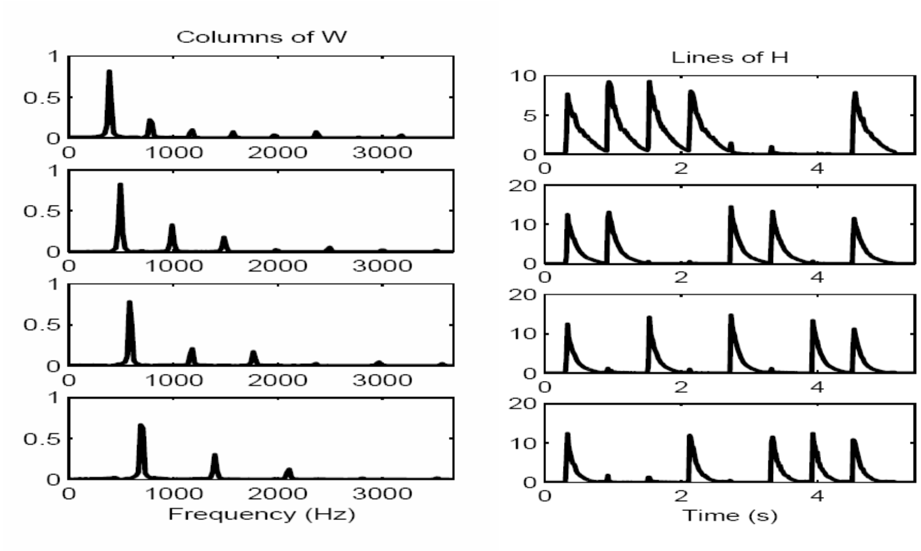
TELECOM ParisTech

Signal Decomposition and source separation for music transcription

- **Non Negative Matrix factorization**

Spectrogram ≈ W(basis) * H(activation)





See for example

N. Bertin, R. Badeau, G. Richard, "Blind Signal Decompositions for Automatic transcription of polyphonic music: NMF and K-SVD on the Benchmark", in Proc. of ICASSP'07

NMF

- **Towards improved transcription**
  - For example use of harmonicity constraints

Classic NMF

$$X_{tf} = \sum_{i=1}^{I} H_{it} W_{if} + R_{tf}$$

"Harmonic" NMF

$$X_{ft} = \sum_{p=p_{\mathrm{low}}}^{p_{\mathrm{high}}} \sum_{i=1}^{I_p} A_{pit} S_{pif} + R_{ft}$$

with

Basis spectra

$$S_{pif} = \sum_{k=1}^{K_p} E_{pik} P_{pkf}$$

Amplitude coefficients
(model the spectral envelope)

Narrowband spectra
(represent adjacent partials at harmonic frequencies)

Towards improved transcription

- An example of transcription

Original *Transcription*

More details in:

E. Vincent, N. Bertin and R. Badeau *Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription*,, in Proc of *ICASSP'08.*

See also for another approach:

V. Emiya, R. Badeau, B. David, "*Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches*, Proc. of EUSIPCO'08

and numerous piano transcription examples:
http://perso.telecom-paristech.fr/~emiya/EUSIPCO08/bench1.html

Main melody (Singing voice) extraction

- **Combine NMF (or GMM) and a production model**

$$X(f,t) = V(f,t) + M(f,t)$$

Original spectrum

Singing Voice

Background Music

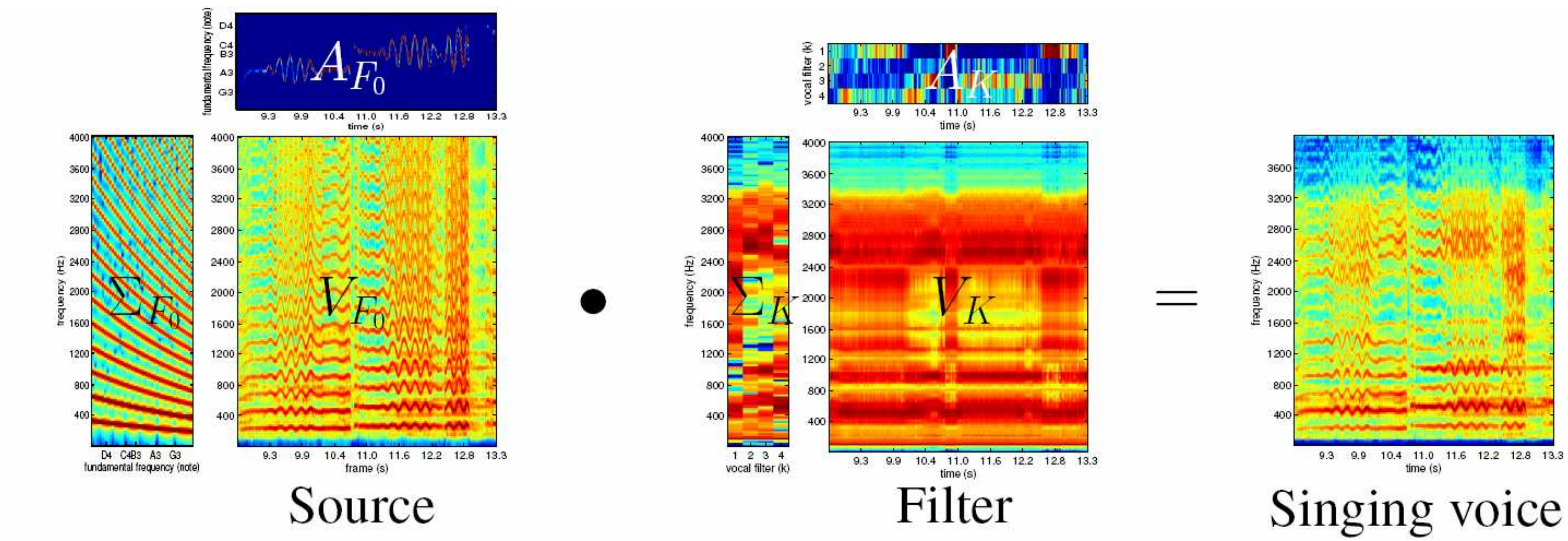
- **Background music model :** instantaneous mixtures of R Gaussian sources

$$M(f,t) \sim \mathcal{N}_c(0, \underbrace{\sum_{r=1}^{R} a_r^2(t)\sigma_r^2(f)}_{D_R(f,t)})$$

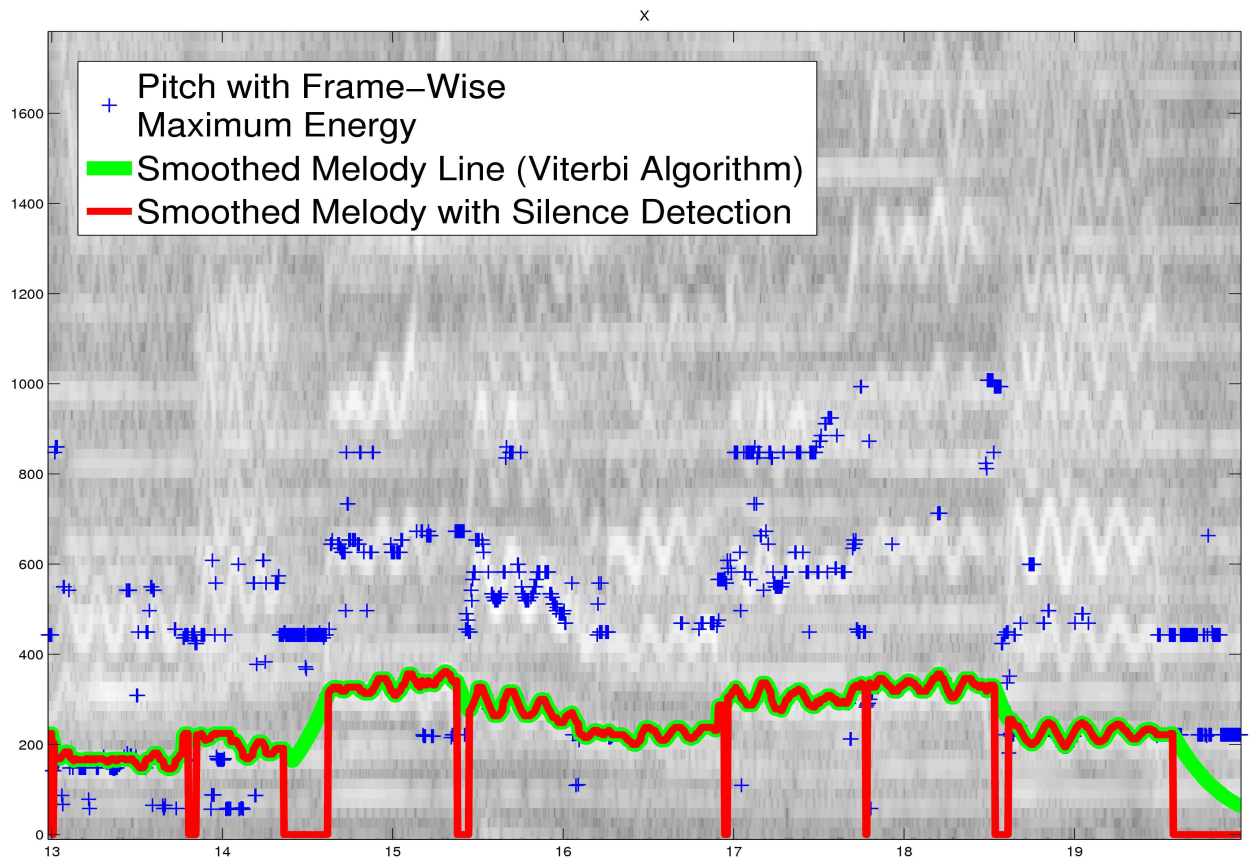
Main melody (Singing voice) extraction

- **Singing voice Model**

$$V(f,t) \sim \mathcal{N}_c\Big(0, \underbrace{\sum_k a_k^2(t)\sigma_k^2(f)}_{V_K(f,t)} \times \underbrace{\sum_{f_0} a_{f_0}^2(t)\sigma_{f_0}^2(f)}_{V_{F_0}(f,t)}\Big)$$



Main melody extraction



Main melody (Singing voice) extraction

- **Source Separation (by Wiener Filtering)**

$$\hat{V}(f,t) = \frac{V_K(f,t) \times V_{F_0}(f,t)}{D(f,t)} X(f,t) \qquad \hat{M}(f,t) = \frac{D_R(f,t)}{D(f,t)} X(f,t)$$

- **Two examples**

Signals	*Original (mp3)*	*Main voice*	*Music*
Opera (fem. Voice)			
Jazz (take 5)			

<u>More examples at</u>

http://perso.telecom-paristech.fr/~durrieu/en/results_en.html

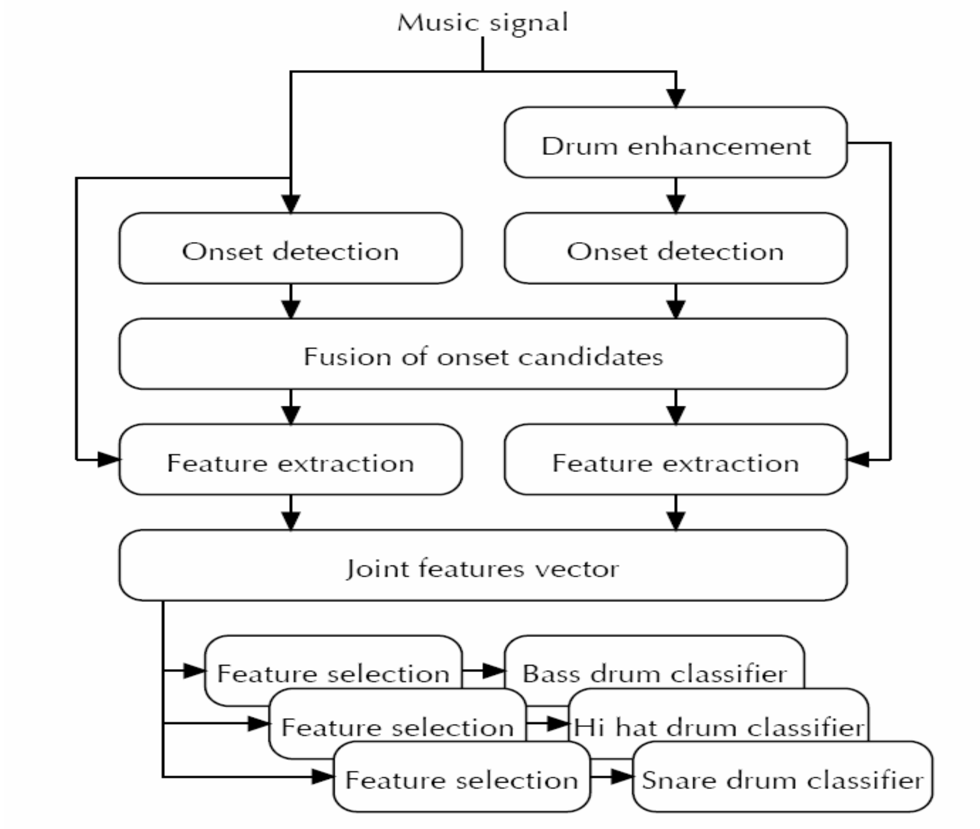
<u>More details in</u>

J-L. Durrieu, G. Richard, B. David, “Singer Melody Extraction in Polyphonic signals using source spearation methods”, in Proc of ICASSP’08.

Some other examples

- **Drum Processing**
  - On Combining tempo, source separation and transcription for drum processing
- **Drum loop Retrieval**
  - On Combining Speech recognition and transcription for drum loop retrieval
- **“Cross-modal” Retrieval**
  - On combining visual and audio correlations for audio-based music video retrieval (or video-based audio retrieval).

Combining tempo, source separation and transcription for drum processing

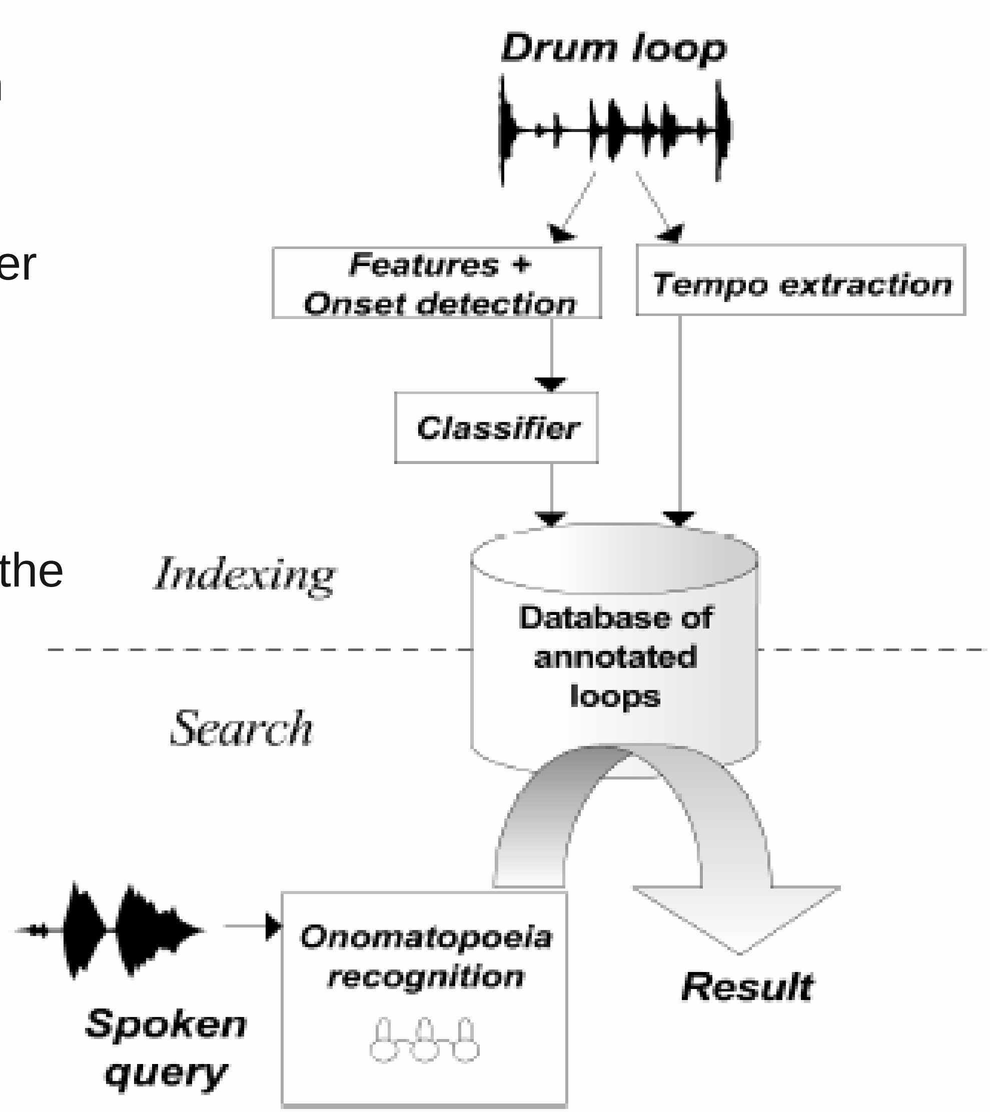


Reference:

O. Gillet, G. Richard. *Transcription and separation of drum signals from polyphonic music*. in IEEE Trans. on ASLP, Volume 16, N°3, March 2008 Page(s):529 - 540.

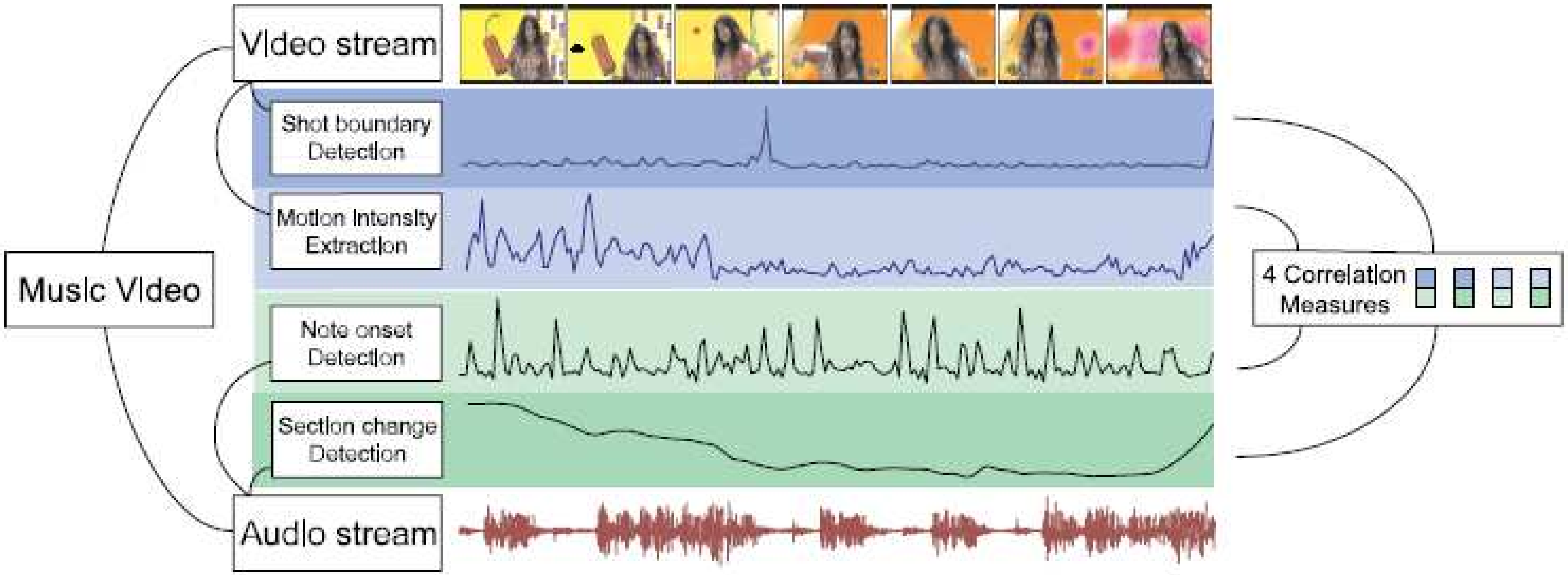
Combining Speech recognition and transcription for drum loop retrieval

- **Drum loop retrieval from spoken queries**
  - An automatic drum loop transcriber
  - A simple HMM based speech recognition engine
  - A retrieval engine that search for the best drum loops based on both textual transcription



Combining visual and audio correlations for music video retrieval

- **Music similarity search (or genre classification)**
- **For video-based audio retrieval**



Demo

- Reference

O. Gillet and G. Richard, « On the Correlation of Automatic Audio and Visual Segmentations of Music Videos » IEEE Trans. on CSVT, 2007

The evaluation problem

The evaluation problem in Music information indexing and retrieval

- **The domain does not have the historic background of other related domains such as Speech recognition for example.**

- **There is a lack of common and public databases and of common protocols for the evaluation**

- **Difficulty of obtaining signal annotation under the form of « metadata » (the use of MIDI signals can only partly resolve the problem since MIDI signals are far less complex to process than real audio signals)**

- **A big step forward in evaluation: the MIREX experience**
  - http://www.music-ir.org/mirex/2008/index.php/Main_Page

Evaluation: some results…(from MIREX http://www.music-ir.org/mirex/2007/abs/MIREX2007_overall_results.pdf)

Multi F0 Note Tracking		
Rank	**Participant**	**Avg. F-Measure**
1	Ryynänen & Klapuri (2)	0.614
2	Vincent, Bertin & Badeau (4)	0.527
3	Poliner & Ellis (2)	0.485

Audio Cover Song Identification		
Rank	**Participant**	**Avg. Precision**
1	Serrà & Gómez	0.521
2	Ellis & Cotton	0.330
3	Bello, J.	0.267

Audio Mood Classification		
Rank	**Participant**	**Avg. Raw Accuracy**
1	Tzanetakis, G.	61.50%
2	Laurier, C.	60.50%
3	Lidy, Rauber, Pertusa & Iñesta	59.67%

Audio Genre Classification		
Rank	**Participant**	**Avg. Raw Accuracy**
1	IMIRSEL (svm)	68.29%
2	Lidy, Rauber, Pertusa & Iñesta	66.71%
3	Mandel & Ellis	66.60%

MIREX 2006 Audio Tempo Extraction Summary Results		
Contestant	At least 1 tempo correct	Both tempi correct
klapuri	94.29%	61.43%
davies	92.86%	45.71%
alonso 2	89.29%	43.57%

Evaluation: some conclusions

- **A overall *percentage of correct recognition* has some meaning if:**
  - The database is known and well described
  - The evaluation protocol and metrics are well described

- **Comparative evaluations are very important**

- **But….evaluations (such as Mirex) are only an instantaneous picture of current algorithms on a given task on a given database with a given protocol…**

Conclusions

- **There is a great interest in exploiting "separation" and "transcription" jointly (and this for both individual tasks)**

- **Audio indexing technology**
  - ….is rapidly progressing
  - ….is supported by numerous applications
  - … still has some limitations especially for the complex case of polyphonic music.
  - … still suffers from the lack of common corpora and protocols for rigourous evaluation

- ….Thank you for you attention

- Contact info and publications:

  www.enst.fr/~grichard/

  Gael.Richard@enst.fr