Research in audio Processing

25th October, 2012

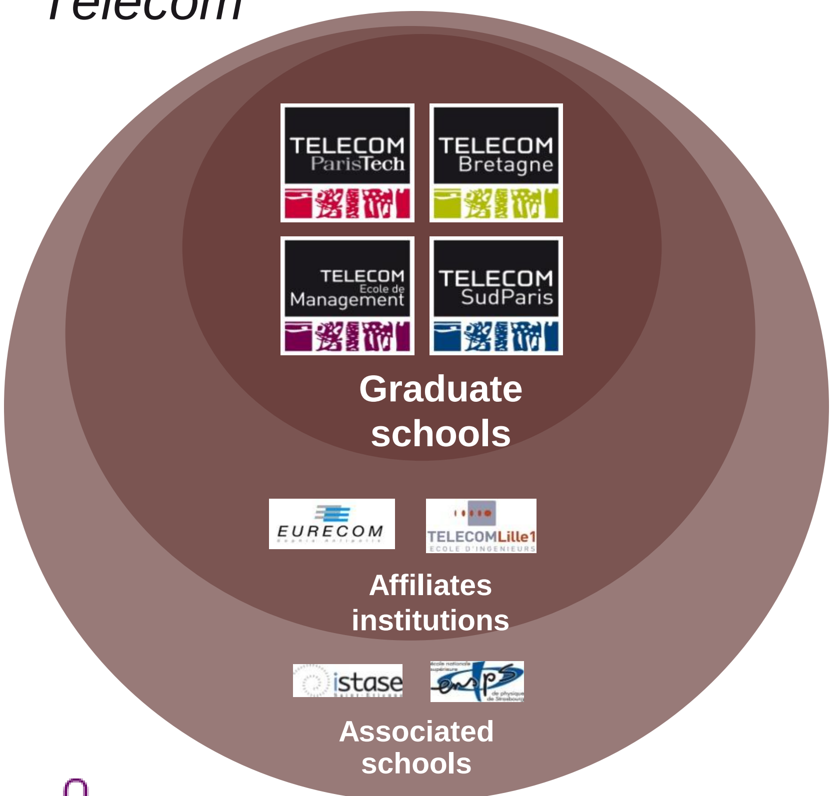
Gaël Richard

Content

- Presentation of Telecom ParisTech (TPT) and the AAO research group
- A view on Greedy pursuits algorithms for representing audio signals: with applications to Compression, Source separation and Audio Fingerprint

Télécom ParisTech

A school within the *Institut Mines-Télécom*



A school within *ParisTech*

ParisTech brings together twelve of the foremost French institutes of education and research

- The full range of sciences and technologies,

Telecom ParisTech, the leading graduate school in Information & Communication Technology (ICT)

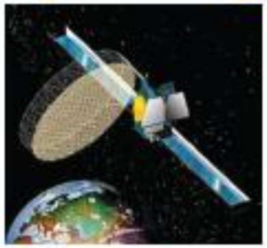
- **A public institution founded in 1878,** placed under the aegis of the minister for Industry
- **Invented the term *telecommunication* in 1904**
- **1st graduate engineering school in ICT in France,** 5th in the national rankings of Engineering schools
- **Hosts the 1st French ICT Incubator** which creates 2-3 start-ups/ month
- **A Budget of 66,4 M€,** including 30% self-financing
  - **4 Missions , 1 ambition** ***Innovating in a Digital World***

Education	Continuing education & life-long training	Research	Business start-up support
Public teaching at the highest level in the domain of ICT		From theory to industrial transfer	Development of entrepreneurial spirit .. to Incubation

Disciplines including **all the sciences and technologies of ICT**

Communication & Electronics - Computer Science & Networking – Signal and Image processing - Economics, Management & Social Sciences

Optimizing



information transport

Networks and mobility
High speed links and optical systems
Digital communications
Aeronautic and satellite systems
Algorithm / Architecture matching

Improving

information processing

Statistical automatic learning methods
Speech, images and audio processing
Multimedia content production and processing
Information system

TELECOM ParisTech

Bringing

services closer to users

Local access and proximity communications
Ambient intelligence
Services creation
Virtual communities : games, education, citizenship

Establishing

and restoring user confidence

Regulation
Cryptography, security, biometric identification
Private life, sociability, culture, ethics
New technologies and society : electronic trade, teleworking, CAL

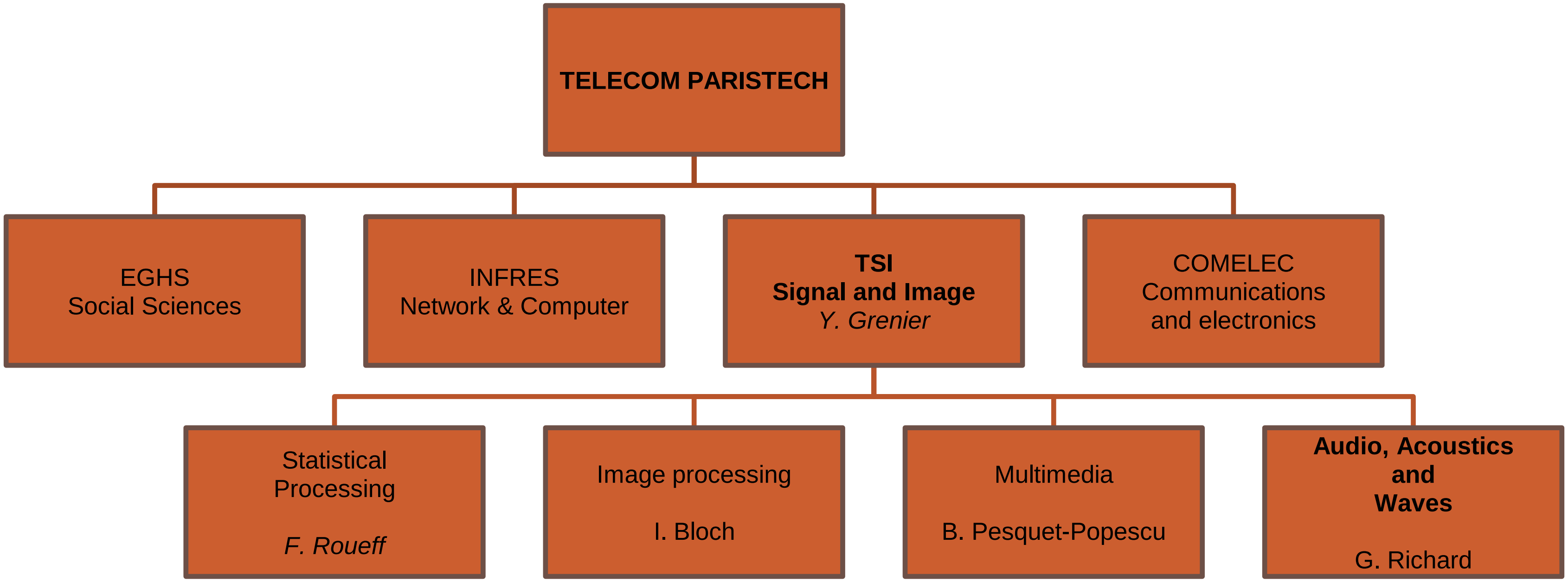
Safeguarding

and enriching our cultural heritage

Databases, indexing, consultation, data mining
Signal, images, music, text processing
Virtual reality, creation on-line
Art and information technologies

TELECOM ParisTech

Télécom ParisTech : Research



- **Signal and Image Processing department**
  - 4 Research and Education Groups
  - 35 Permanent Members
    - 20 Faculty Members
    - 10 Full time Research Members (CNRS)
    - 5 Technical & Administrative Support
  - 55 PhD candidates
  - 5~10 Post-Docs & Sabbatical

Audio, Acoustics and Waves group

« AAO »

- The AAO group (6 permanent staff):

G. Richard

B. David

R. Badeau

Y. Grenier

S. Essid

A. Gramfort

+ 3 post docs / 1 Engineer (T. Fillon, A. Drémeau, C. Damon)

+ 13+ PhDs (M. Maazaoui, M. Moussalam, B. Fuentes, R. Foucard, S. Fenet, A. Liutkus, N. Lopez, X. Jaureguiberry, A-C. Conneau, A. Masurelle, F. Rigaud, N. Seichepine, C. Fox, H. Bai)

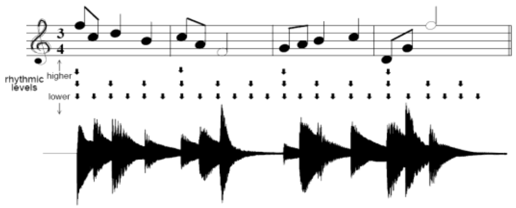
- <u>Aim of the group :</u> to develop digital signal processing methods with applications to audio, multimodal and biomedical signals.

  - from theoretical work on machine learning, signal models and sparse representations …
  - … to computational optimization of algorithms.

AAO « *Audio, Acoustics and Waves* »

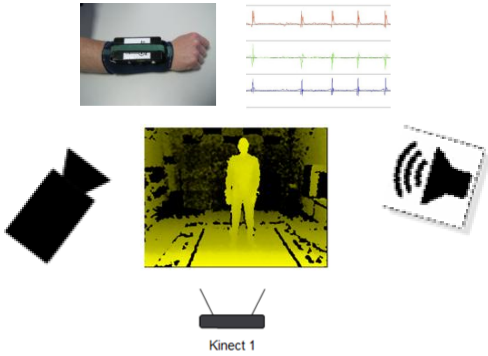
http://www.tsi.telecom-paristech.fr/aao/en/

Music signals processing



Multimedia streams analysis

heterogenous sensors arrays signal processing



Biological signals processing

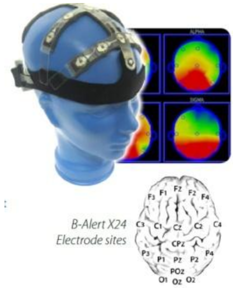


Image from http://www.bmedical.com.au/

AAO *« Audio, Acoustics and Waves »*

http://www.tsi.telecom-paristech.fr/aao/en/

Methods: Multimedia and audio signal representations and models

Deterministic and probabilistic models

- HR methods for sinusoidal estimation (adaptive tracking of the signal subspace)
- Non-Negative Matrix factorization (NMF)
- Kernel methods for classification, feature selection
- Sparse decompositions *(Matching Pursuit, …)*
- Source separation

Capture

Echo cancellation
Machine Audition
Acoustics
….

Analysis

Indexing, Transcription
Segmentation AV
Biomedical Signals
Fingerprinting, ….

Transmission

Watermarking
Compression

Restitution

Binaural reproduction
Remasterisation
(Remix / Upmix)

Some audio tools and technologies available…

http://www.tsi.telecom-paristech.fr/aao/en/

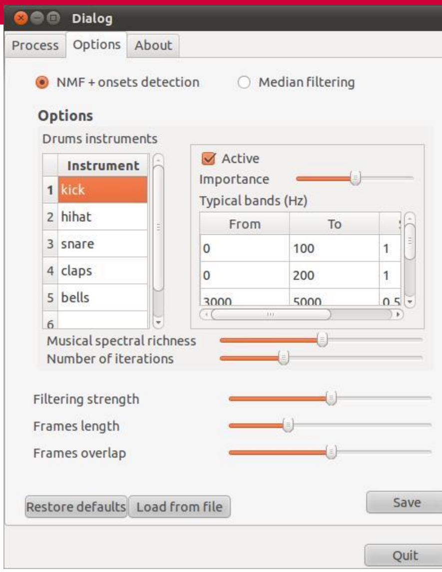
- **Databases**
  - **ENST-Drums (2006)**
  - **MAPS (2009)**
  - **3DLife ACM Multimedia Grand Challenge 2011 Dataset**
  - **Romeo-HRTF (2011)**
  - **QUASI (2012)**
  - …
- **Softwares**
  - **Yaafe :** An efficient toolbox for audio feature extraction (licence LGPL)
  - **Smarc :** Efficient sampling frequency conversion (licence LGPL)
  - **Desam Toolbox** : Matlab toolbox for audio signal processing (licence GPL)
  - **Audio separators**
  - …

  - Accessible at *http://www.tsi.telecom-paristech.fr/aao/en/*

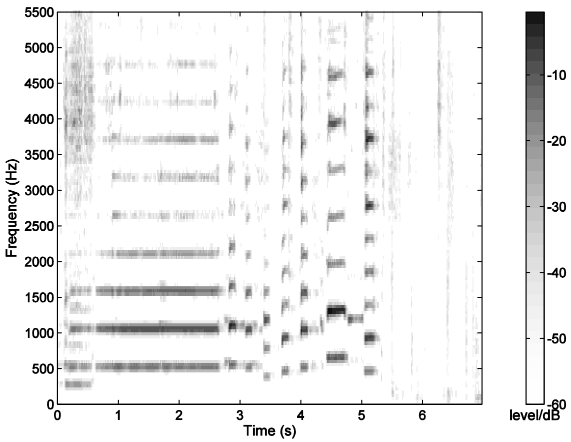
(some) Available tools for separation

http://www.tsi.telecom-paristech.fr/aao/en/

- **Drum extractor:**
  - Available at : http://perso.telecom-paristech.fr/~liutkus/



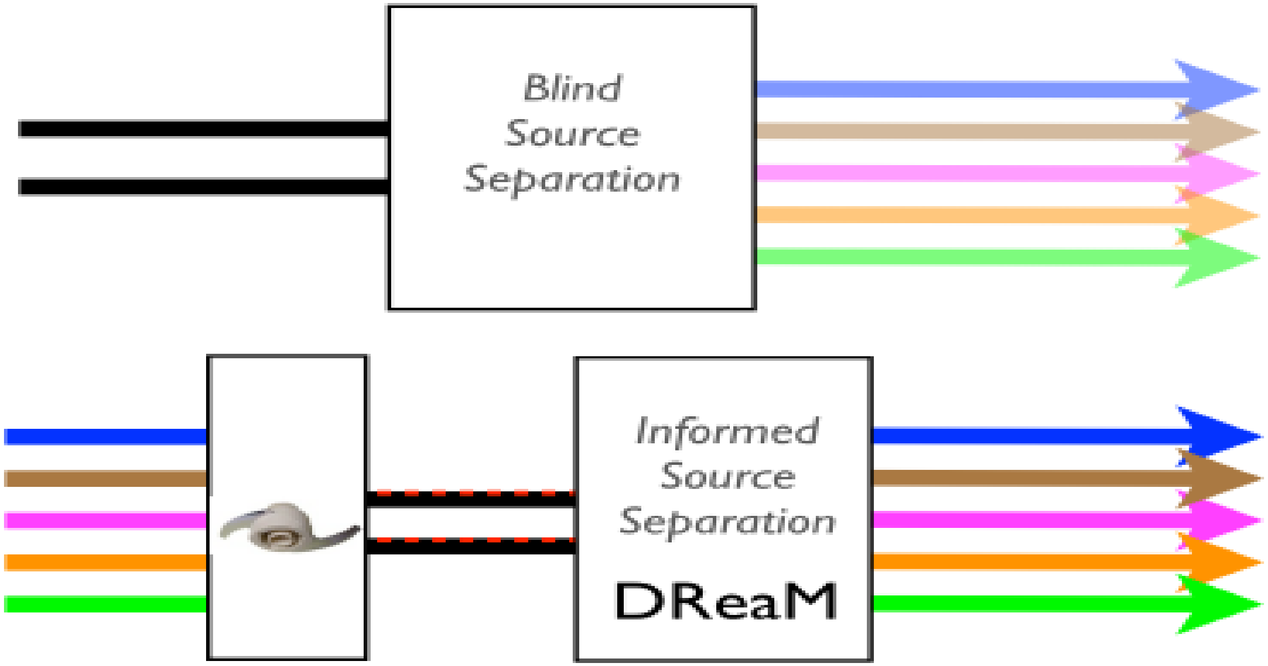
- **Leading voice extractor**
  - Available at: http://www.tsi.telecom-paristech.fr/aao/2011/06/07/402/



Collaborations…

- Involved in a variety of projects sponsored by industry or national and European bodies (ANR, EC, Oseo, …)

- One example the ANR-Dream project



Greedy pursuits algorithms for representing audio signals

with applications to Compression, Source separation and Audio Fingerprint

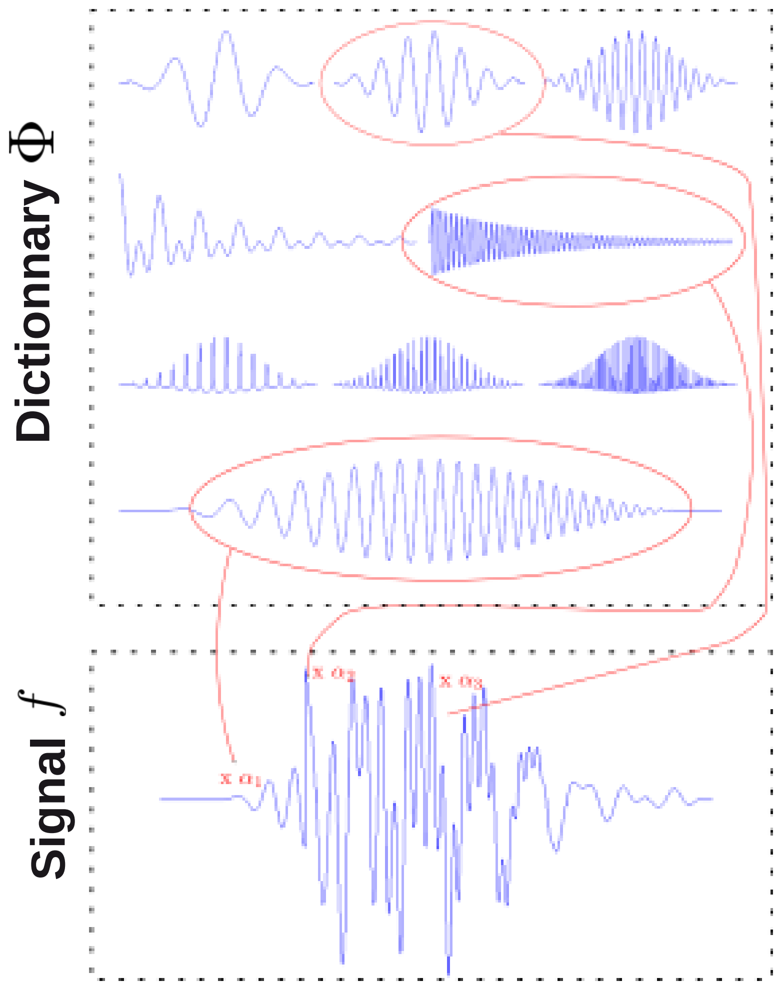
with Manuel Moussallam and Laurent Daudet

Content

- **Matching Pursuit (MP): a greedy approach for audio signal representation**

- **Three variations of MP**
  - Random MP: An interesting extension for compression
  - Redundant MP: An interesting extension for source separation
  - Coverage constrained MP: An interesting extension for Audio ID

Sparse representation of audio signals



- **Standard formulation**

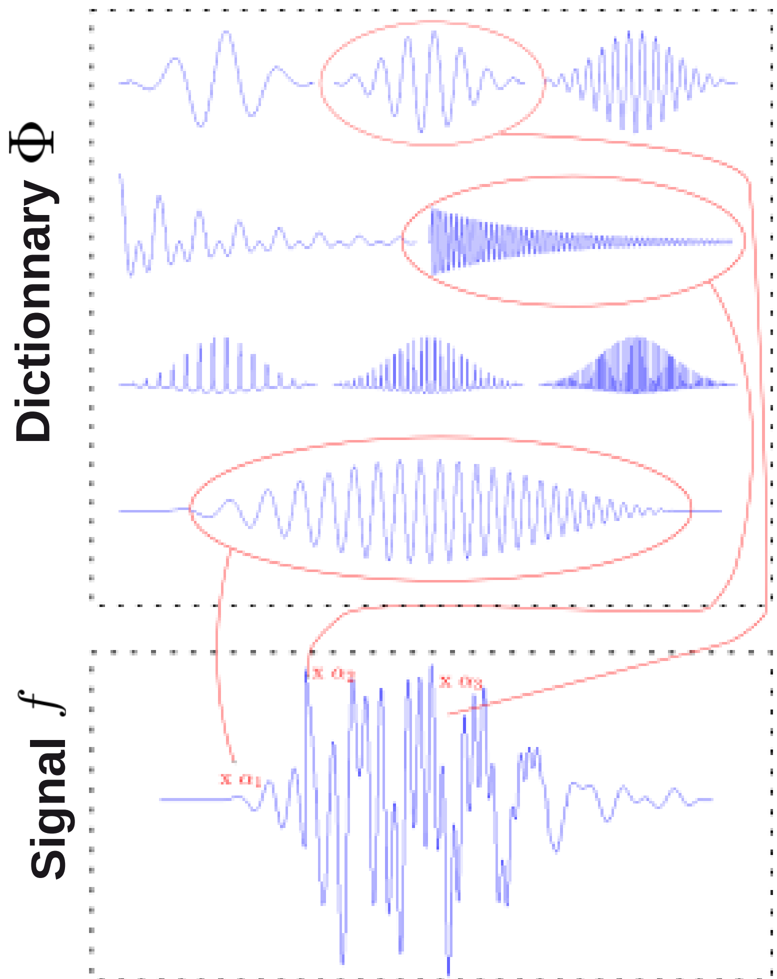
Let $f \in \mathbb{R}^N$, find the sparsest linear expansion of the signal f in a dictionary $\Phi = \{\phi_i\}_{i \in [0..M-1]}$

That is $$\min \|\alpha\|_0 \text{ s.t. } f = \Phi\alpha$$

Or alternatively

$$\min K \text{ s.t. } f = \sum_{k=1}^{K} \alpha_k \phi_k$$

Sparse *approximation* of audio signals



- **Standard formulation**

Let $f \in \mathbb{R}^N$, find the sparsest linear expansion of the signal f in a dictionary $\Phi = \{\phi_i\}_{i \in [0..M-1]}$

$$\min \|\alpha\|_0 \text{ s.t } .\|f - \Phi\alpha\|_2 \leq \epsilon$$

Or alternatively

$$\min K \text{ s.t. } \|f - \sum_{k=1}^{K} \alpha_k \phi_k\|_2 \leq \epsilon$$

How to obtain the sparse approximation ?

- **Many existing approaches**
  - Convex optimisation : $||.||_0 \rightarrow ||.||_p$
  - Bayesian approaches (MAP)
  - Greedy methods (such as those based on Matching Pursuit)

A greedy approach: Matching pursuit

$$\min K \text{ s.t. } \|f - \sum_{k=1}^{K} \alpha_k \phi_k\|_2 \leq \epsilon$$

A simple process:

- The most prominent atom (*i.e. the most correlated with the signal)* is extracted.
- The selected atom is subtracted from the original signal.
- Iterate the procedure until a predefined criterion is met

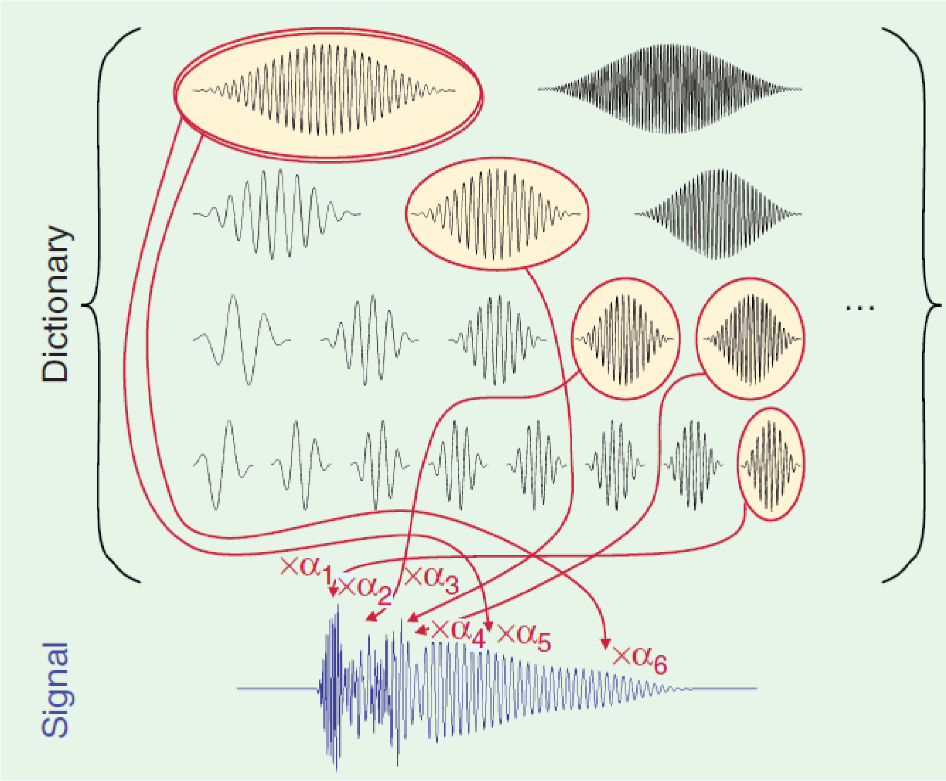
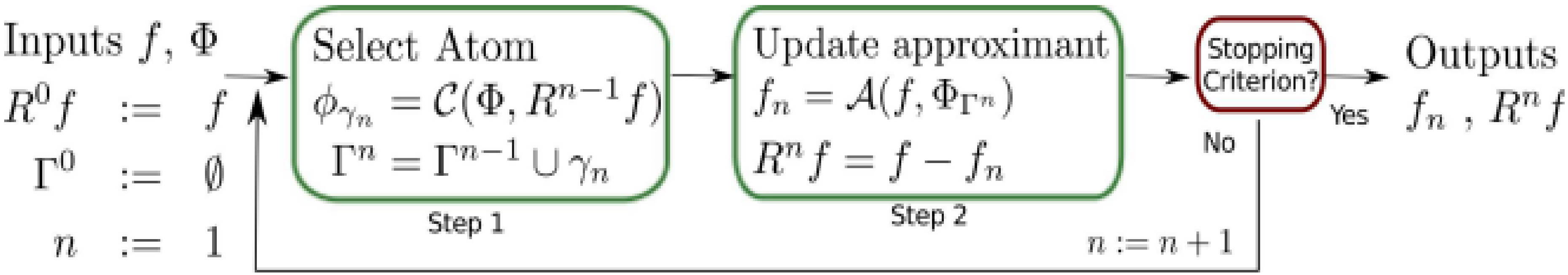


Figure from L. Daudet: *Audio Sparse Decompositions in Parallel,* IEEE Signal Processing Magazine, 2010

Matching pursuit

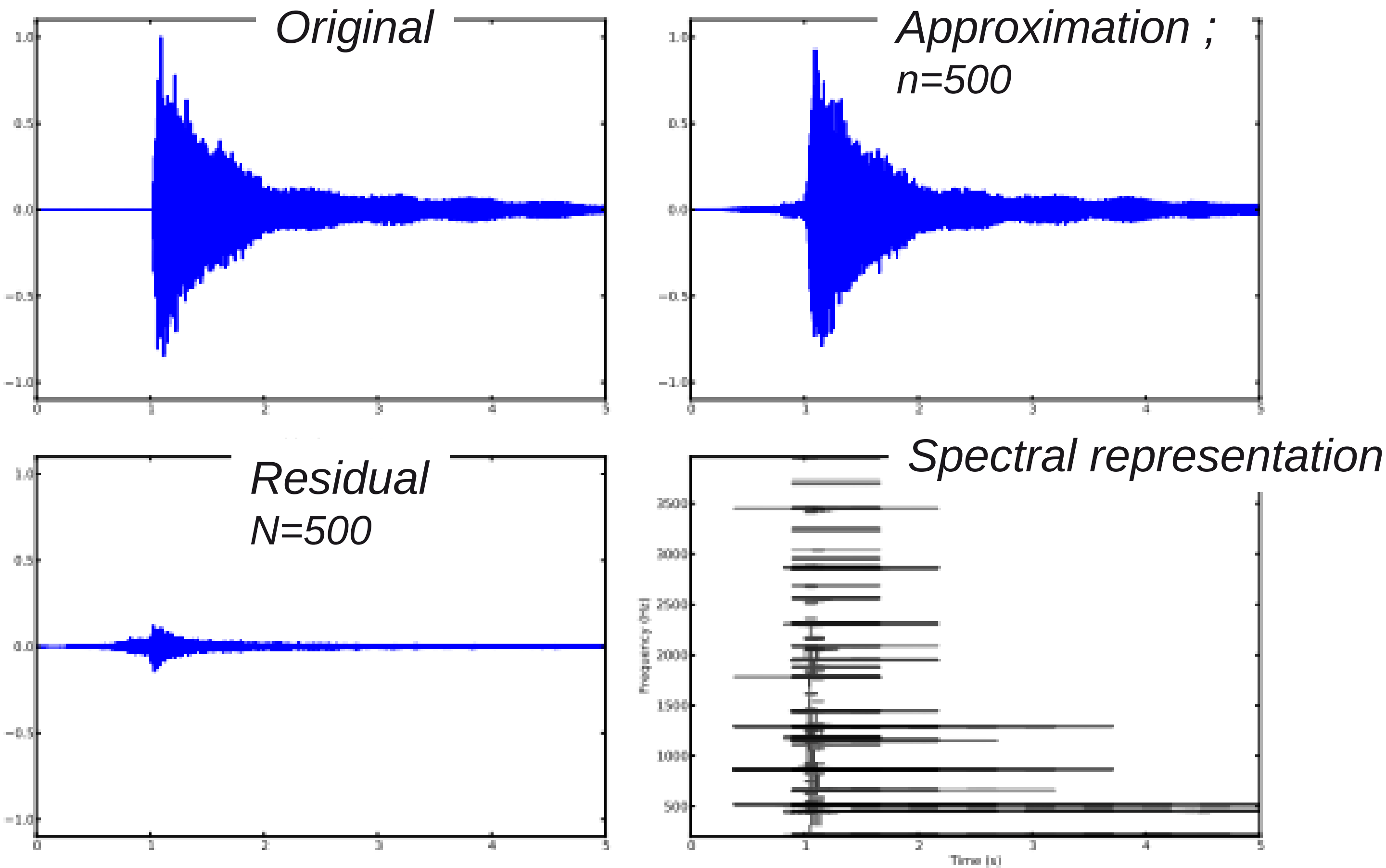


- $R^n f$: *Residual signal after n iterations*
- Γ^n : Set of selected atoms
- f^n : Approximated signal after *n* iterations

Parameters Φ	**Examples**
The Dictionary $\mathcal{C}$	Dictionary of Gabor atoms
The Selection Rule $\mathcal{A}$	$\arg\max_{\phi_\gamma \in \Phi} \lvert\langle R^n f, \phi_\gamma \rangle\rvert$
The Update Strategy	addition of new contribution
The Stopping Criterion	Signal to Noise Ratio

Matching pursuit

- **Decomposing a bell sound in a multiscale MDCT-based dictionary**



Different types of dictionaries for different applications

- **Use "informed atoms"**
  - Specific instruments atoms for instrument recognition
  - Specific pitched atoms for multipitch estimation
  - Specific atoms of a given source for source separation
  - Specific atoms for audio inpainting or denoising

- **Use single or union of orthogonal bases (such as MDCT)**
  - Interesting for Compression

Three extensions of MP

- **Random Sequential Sub-dictionaries Matching Pursuit**
  - Application to audio compression
- **Matching pursuit using similarity**
  - Application to audio fingerprint
- **Matching pursuit using structure**
  - Application to singing voice separation

Weak Matching pursuit

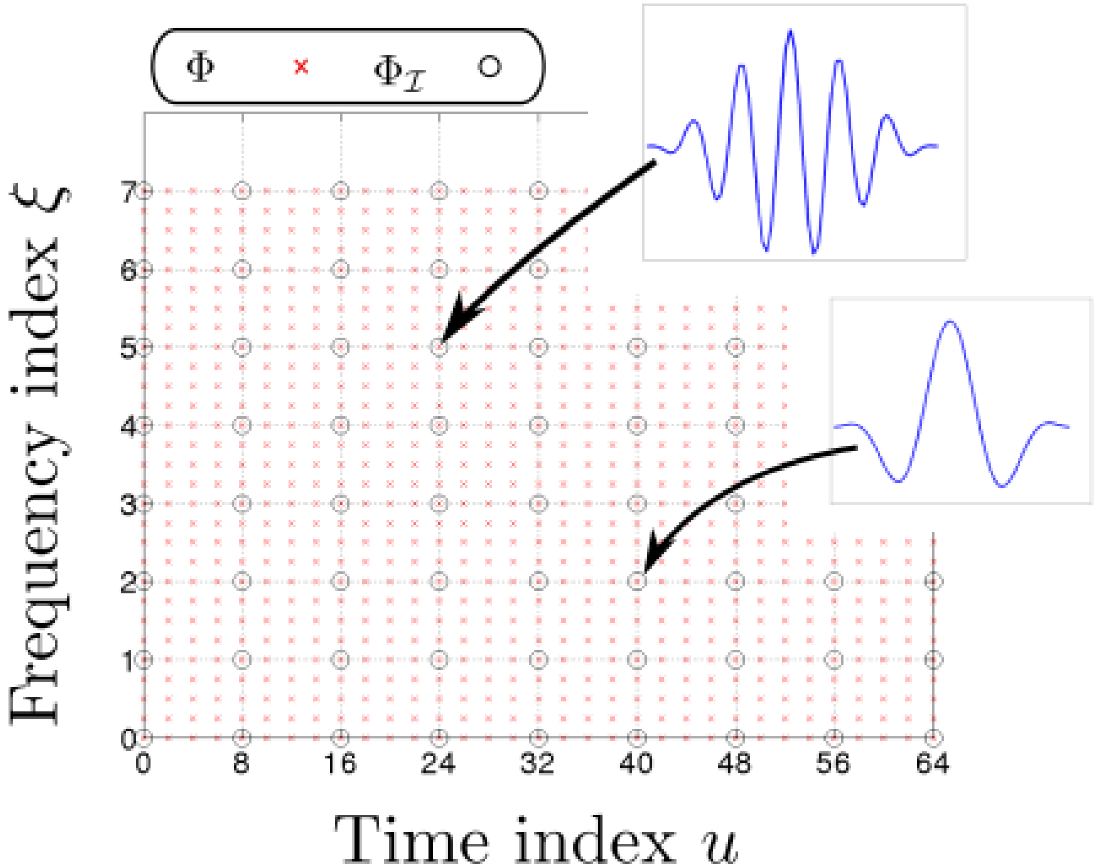
- **A search in the complete dictionary may be too complex**

- **A solution:**
  - use only a sub-dictionary (which only contains parts of the complete dictionary).

- **In practice**
  - This results in a gain of complexity but in a slower convergence
  - Selected atoms are less appropriate
  - Different possibilities for building the sub-dictionaries

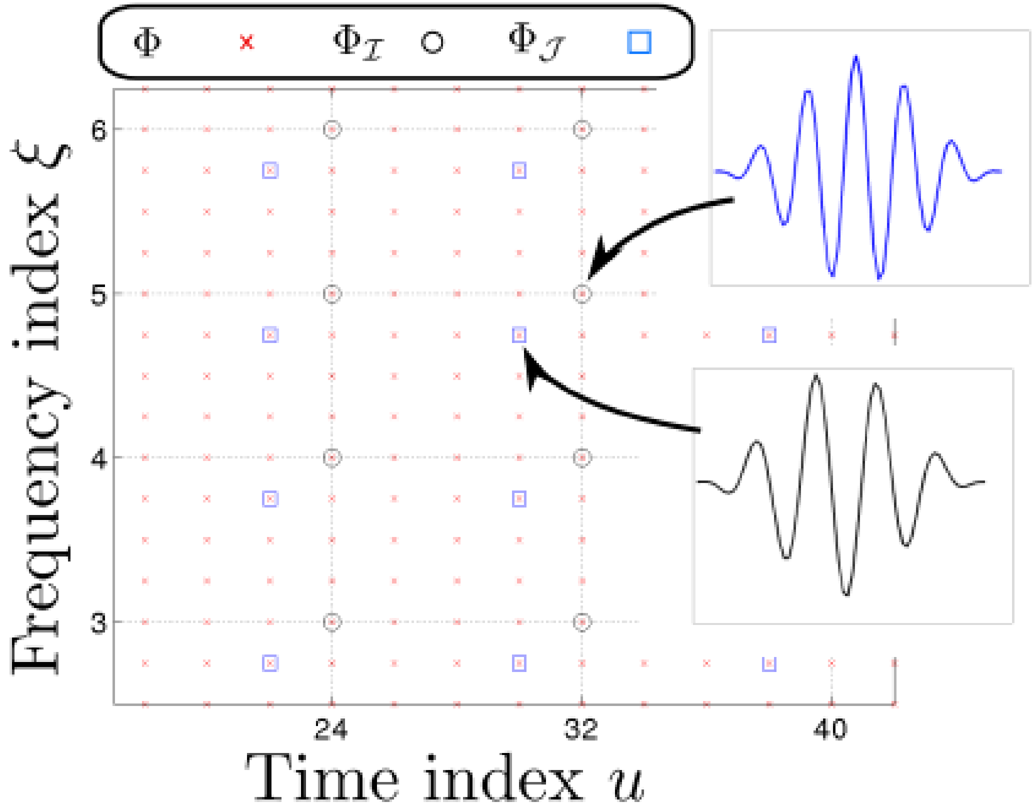
Weak Matching pursuit (2)

- **Example with a dictionary of time-frequency atoms (full and undersampled dictionaries)**



Weak Matching pursuit

- **Different choices for the sub-dictionaries**
  - A different choice leads to a different decomposition



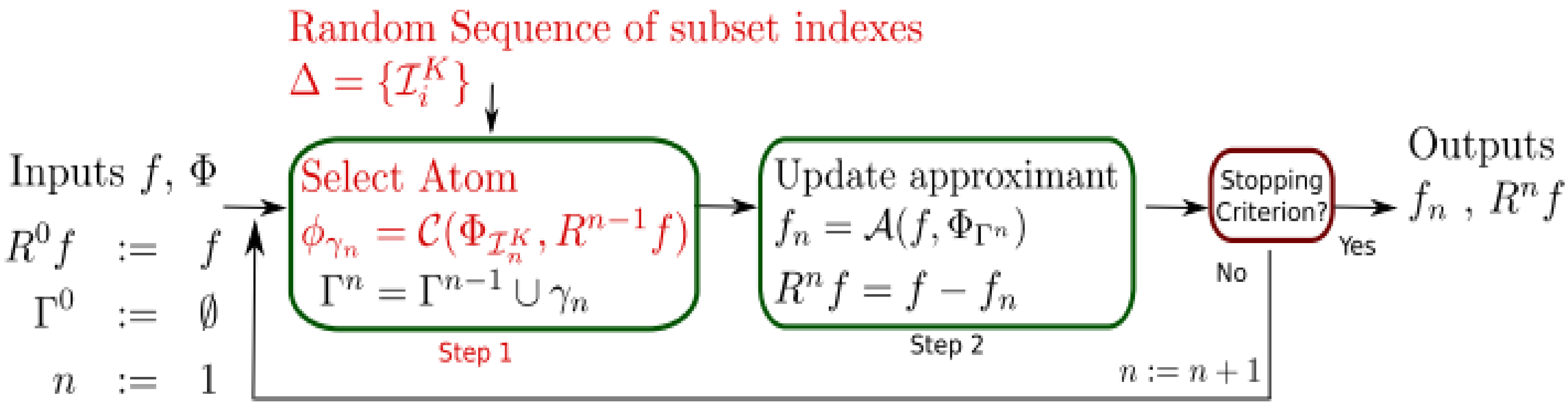
Sequences of sub-dictionaries

- **Usually:**
  - The dictionary is fixed for the whole decomposition
  - A few exceptions :
    - Probabilistic matching pursuit (a posteriori mean of multiple *runs* on different (but fixed) sub-dictionnaries for each decomposition)
    - "Adaptive" dictionaries (each atom is locally optimised after selection)

- **Our approach:**
  - Use a different dictionary at each iteration
  - The sub-dictionary changes according to a (pseudo) random sequence

Random Sequential Subdictionaries MP (RSS-MP)

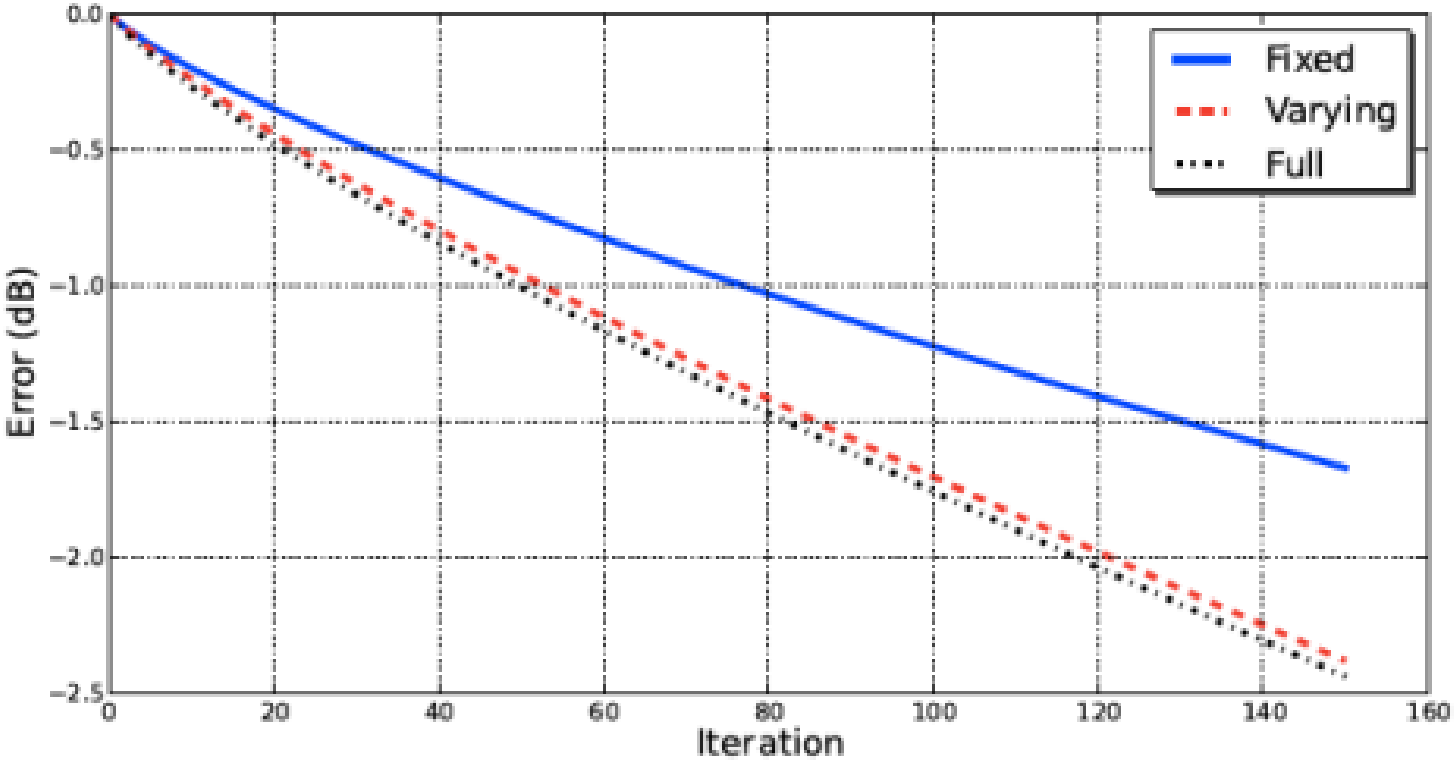
- **Only the first step is changed compared to the classical MP:**



- **Recall:**
- $R^n f$: *Residual signal after n iterations*
- Γ^n : Set of selected atoms
- f^n : Approximated signal after *n* iterations

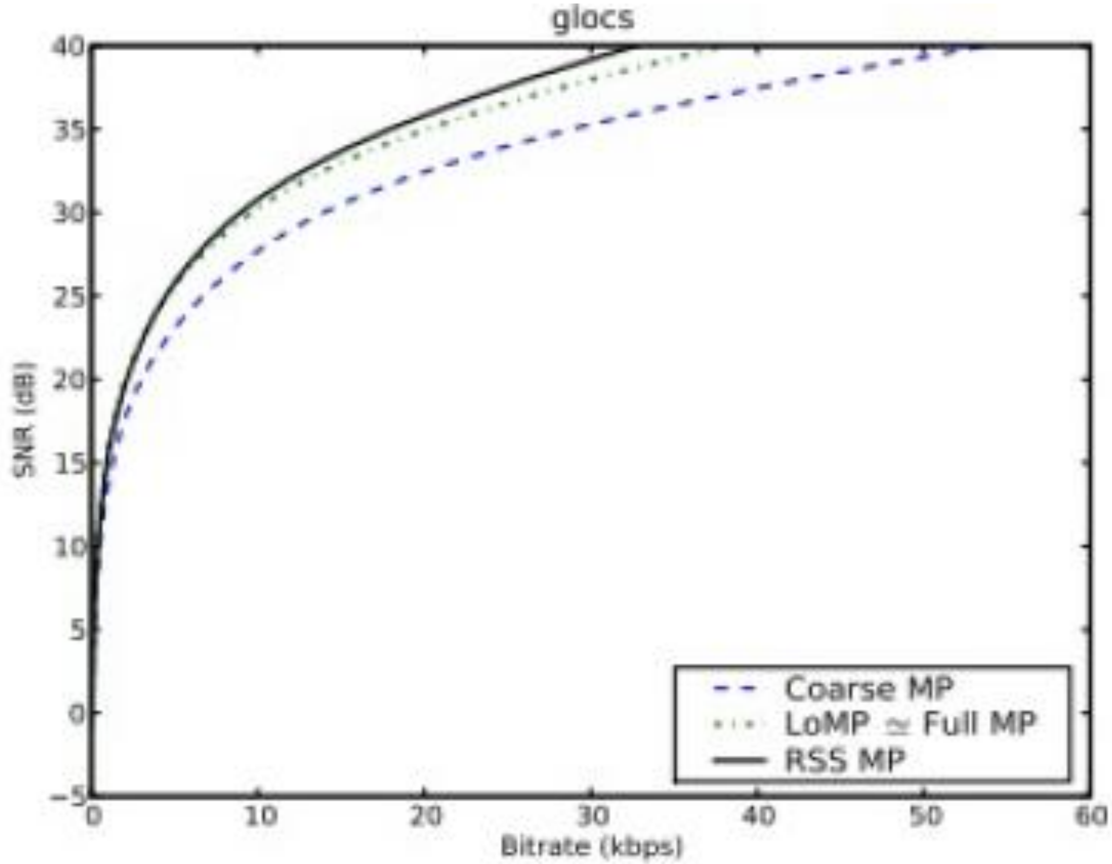
Random Sequential Subdictionaries MP (RSS-MP)

- **Performance :** close fo full dictionary case
- **Cost :** close to under-sampled dictionary case



Random Sequential Subdictionaries MP (RSS-MP)

- Clear advantage for compression (the sequence of sub-dictionaries is not transmitted)



Three extensions of MP

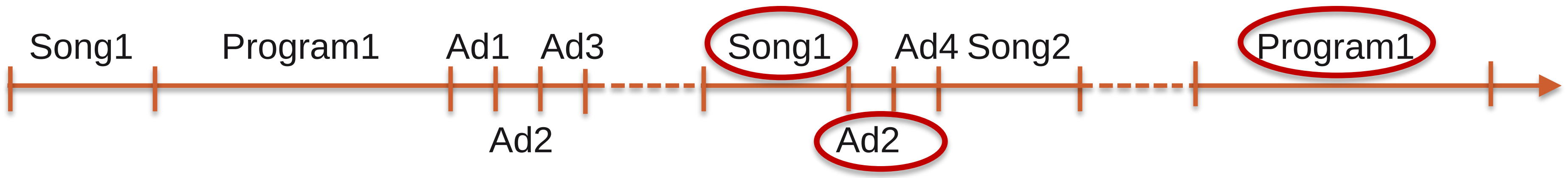
- **Random Sequential Sub-dictionaries Matching Pursuit**
  - Application to audio compression

- **Matching pursuit for audio fingerprint (repeating audio objects detection)**

- **Matching pursuit using structure**
  - Application to singing voice separation

The broadcast use case: detecting of repeating audio objects

- Broadcast streams are quite repetitive

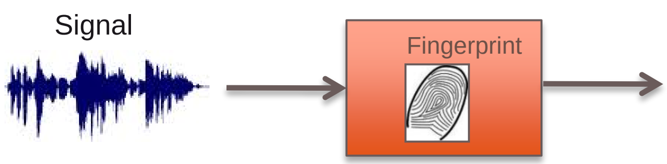


- Repeated objects might be distorted (different volume, equalization, background noise…)

- Detecting these repetitions opens the door to numerous applications (compression, automatic annotation, segmentation…)

Fingerprint systems

- **Most fingerprint systems rely on the following transform of the signal**



$$\left.\begin{array}{c}(key_1, t_1)\\(key_2, t_2)\\\ldots\\(key_n, t_n)\end{array}\right]$$

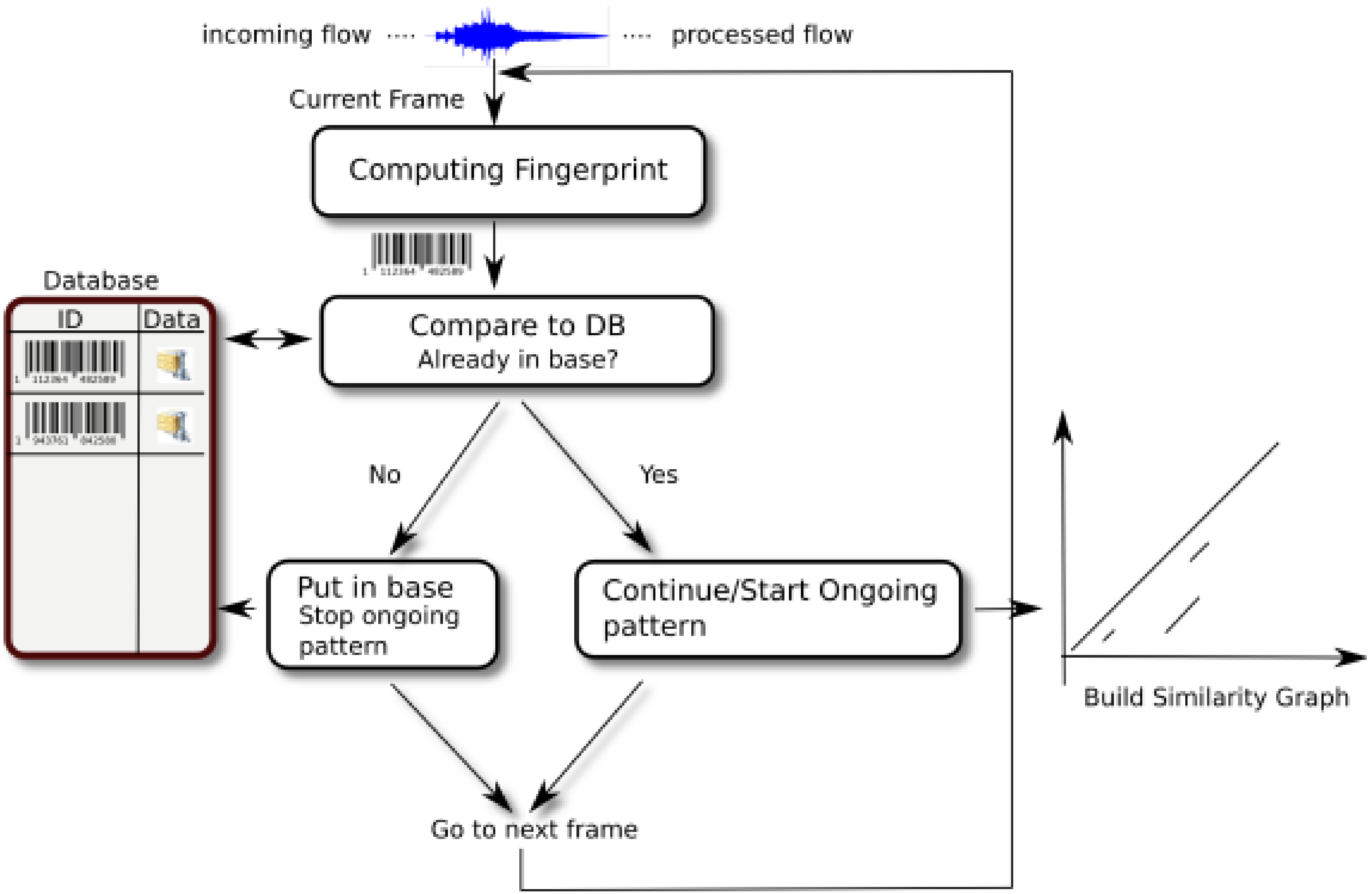
Fingerprint of the Signal

- **Idea: using Matching pursuit approach with a time-frequency plane coverage constraint**

$$\mathcal{C}_{\mathcal{M}}(R^n x, \Phi) = \arg\max_{\phi_i \in \Phi} \left(|\langle R^n x, \phi_i\rangle| \mathcal{M}(\phi_i|\Gamma^n)\right)$$

$$\mathcal{M}(\phi_i|\Gamma^n) = 1 - \max_{\gamma \in \Gamma^n} |\langle \phi_i, \phi_\gamma\rangle|$$

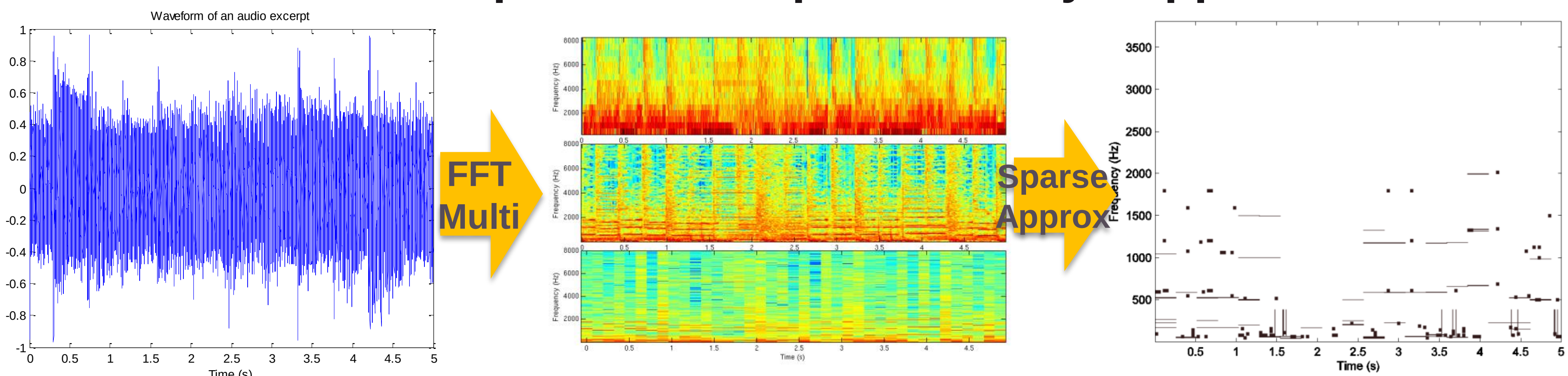
Repeating object detection : scheme



MP based fingerprint

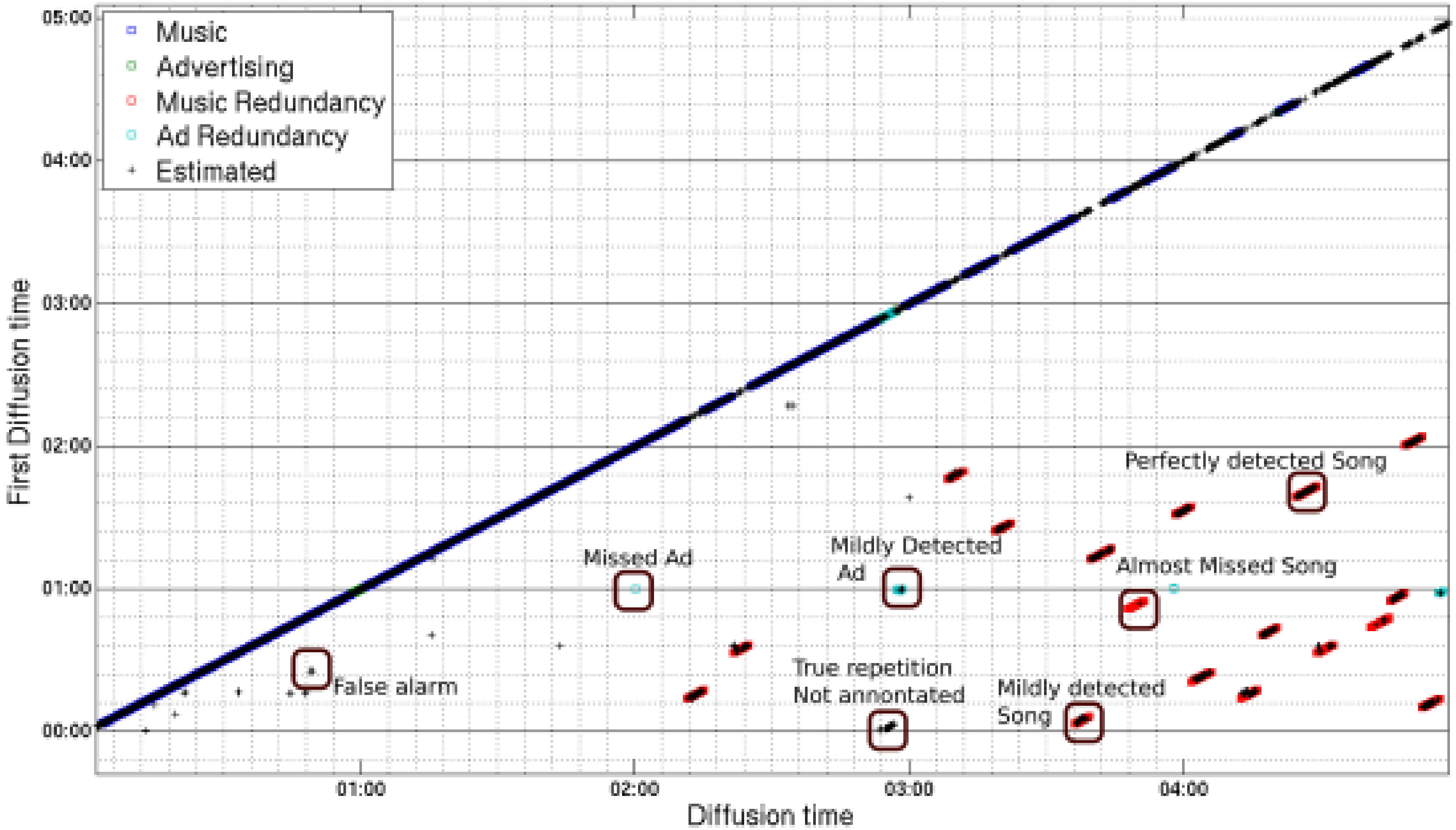
- **Sparse Approximation of the signal on a Multiscale Gabor Dictionary (STFT)**

- **Atoms selected with MP using a constraint on TF coverage: shallow decomposition → Sparse Binary Support**



- **One key = one atom (scale and frequency)**

The output


Music
Advertising
Music Redundancy
Ad Redundancy
Estimated
First Diffusion time
Diffusion time
05:00
04:00
03:00
02:00
01:00
00:00
01:00
02:00
03:00
04:00
Perfectly detected Song
Missed Ad
Mildly Detected Ad
Almost Missed Song
False alarm
True repetition Not annontated
Mildly detected Song

Evaluation

- **Preliminary synthetic evaluation**
  - Corpus = concatenation of 30-seconds experts – 240 in total, 100 of which are exact repetitions of previous ones
  - Analysis performed by the system on 5s segments – one decision per segment

Algorithm	CQT (reference)	MP
Precision (%)	95.1	94.5
Recall (%)	97.8	91.5
F-Measure (%)	96.5	93.0
CPU/segment (s)	0.20	0.40
Database (MB)	9.3	2.4

Recall = Nb of good detected repetions / Total nb of repetitions
Precision= Nb of good detected repetions / Total nb of detections

TELECOM ParisTech

Real World Evaluation (Quaero 2012)

- **2 real world corpora:**
  - 3 days of the same radio (72 h)

Algorithm	Télécom - CQT	Télécom - MP
Recall	1.00	0.95
Precision	0.99	0.99

  - The same day for 3 different radios (72 h)

Algorithm	Télécom - CQT	Télécom - MP
Recall	0.97	0.78
Precision	0.99	1.00

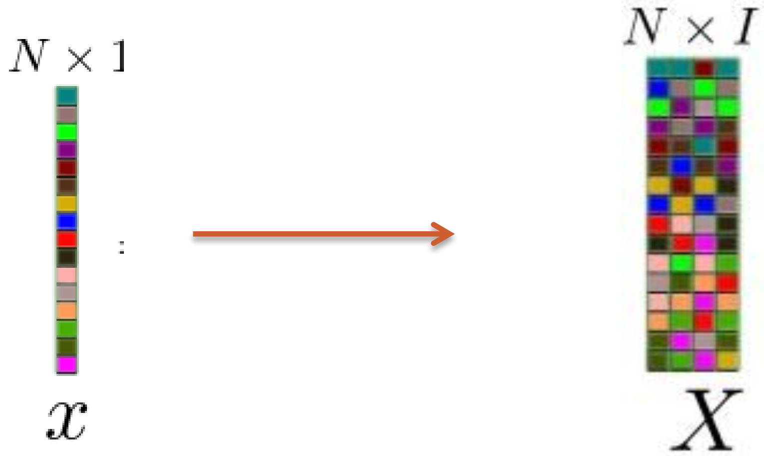
Three extensions of MP

- **Random Sequential Sub-dictionaries Matching Pursuit**
  - Application to audio compression
- **Matching pursuit for audio fingerprint (repeating audio objects detection)**
- **Matching pursuit using structure**
  - Application to singing voice separation

Singing voice separation

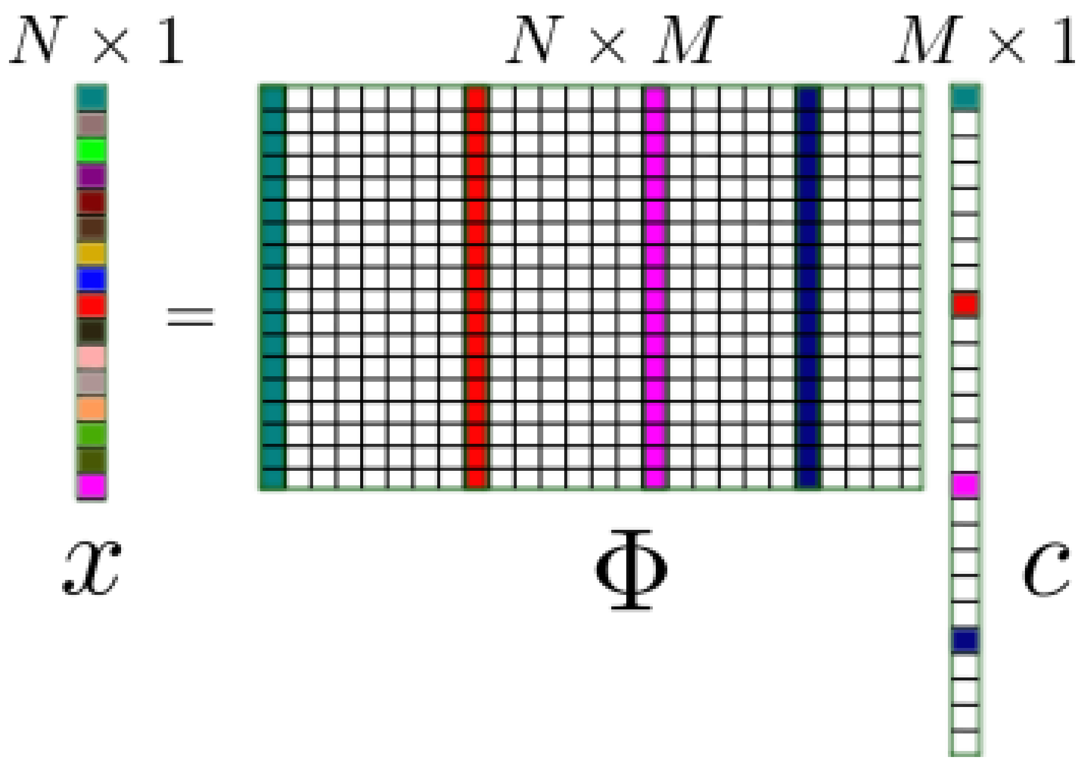
- **The idea:**
  - The singing voice is variable with time
  - The foreground music is somewhat repetitive due to the music structure (chorus – verse).

  - We suppose that we know where are the repetitions : the signal is sliced in I (known) repeating segments



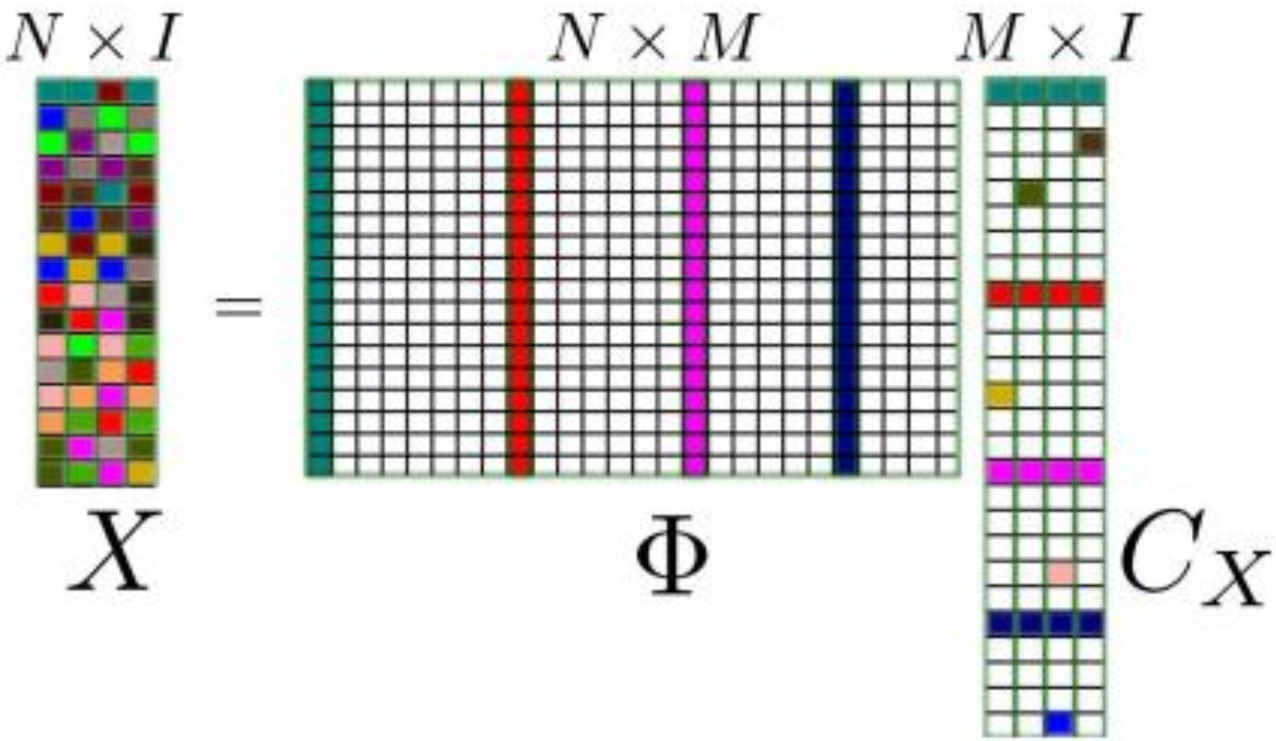
Singing voice separation

- Using Sparsity on the audio signal



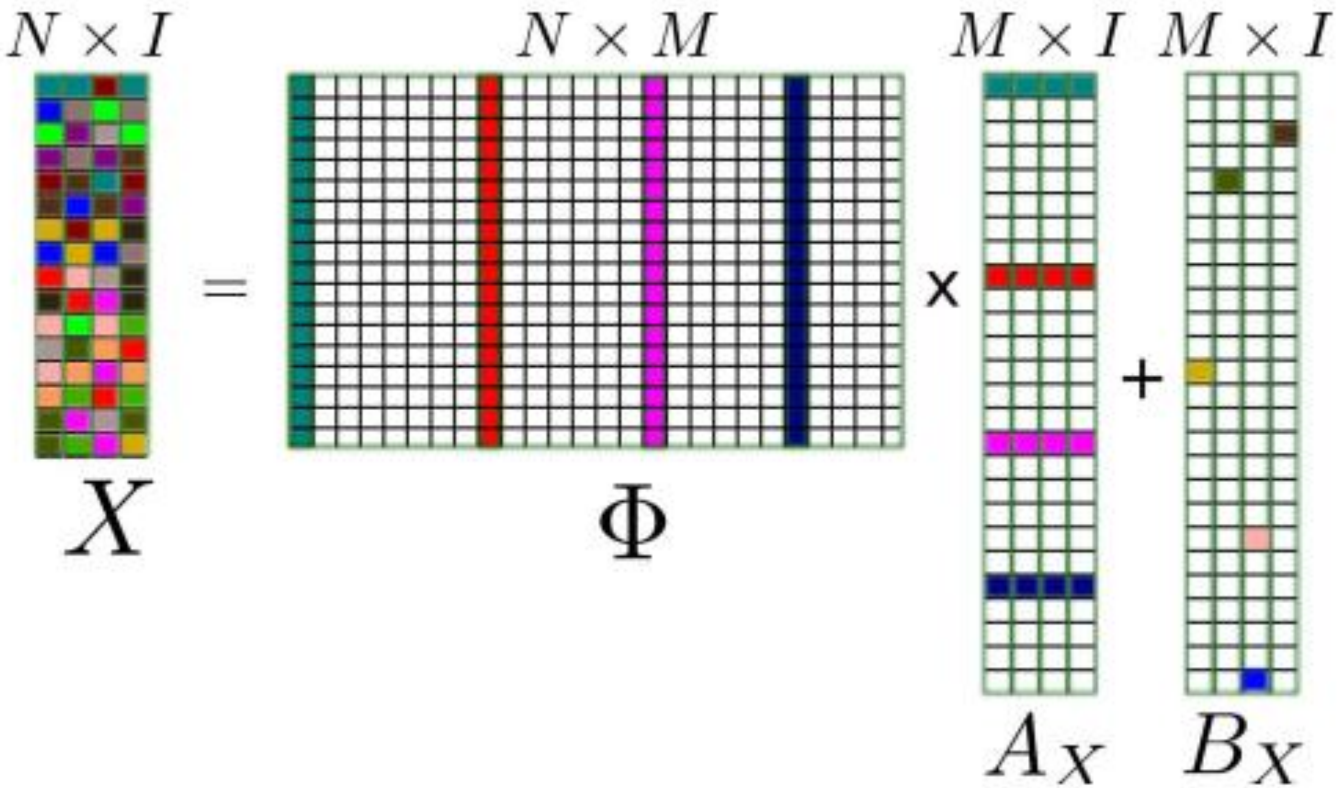
Singing voice separation

- Using Sparsity on the stacked audio signal



Singing voice separation

- Using **Structured** Sparsity on the stacked audio signal



- Separated the singing voice $\mathbf{\Phi}.\mathbf{B_X}$ and the background $\mathbf{\Phi}.\mathbf{A_X}$

Modified MP algorithm

Input: $\mathbf{X}$, $\mathbf{\Phi}$

$\mathbf{R}^0 := \mathbf{X}$, $n = 0$

repeat

  Step 1 : Select atom $\phi_k \leftarrow \mathcal{C}(\mathbf{\Phi}, \mathbf{R}^n f)$

  Step 1 bis : Decide if ϕ_k is background or not

  if ϕ_k is background **then**

    $\forall i, \mathbf{A_X}[i,k] = \langle \phi_k, R_i^n \rangle$

  else

    Find which channels $J \subset I$, ϕ_k belongs to.

    $\forall j \in J, \mathbf{B_X}[j,k] = \left\langle \phi_k, R_j^n \right\rangle$

  end if

  Step 2 : Update residual :

  $\mathbf{R}^n = \mathbf{X} - \mathbf{\Phi}.(\mathbf{A_X} + \mathbf{B_X})$

  $n \leftarrow n + 1$

until a stopping condition is met

Output: $\mathbf{R}^n$, $\mathbf{A_X}$ and $\mathbf{B_X}$

Different selection rules

- Decision based on $r_i^n(\phi) = |\langle R_i^n, \phi \rangle|^2$

- **Energetic criterion**

$$\mathcal{C}_S(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} \sum_{i=0}^{I-1} r_i^n(\phi)$$

- **Minimum risk**

$$\mathcal{C}_M(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} \min_i r_i^n(\phi)$$

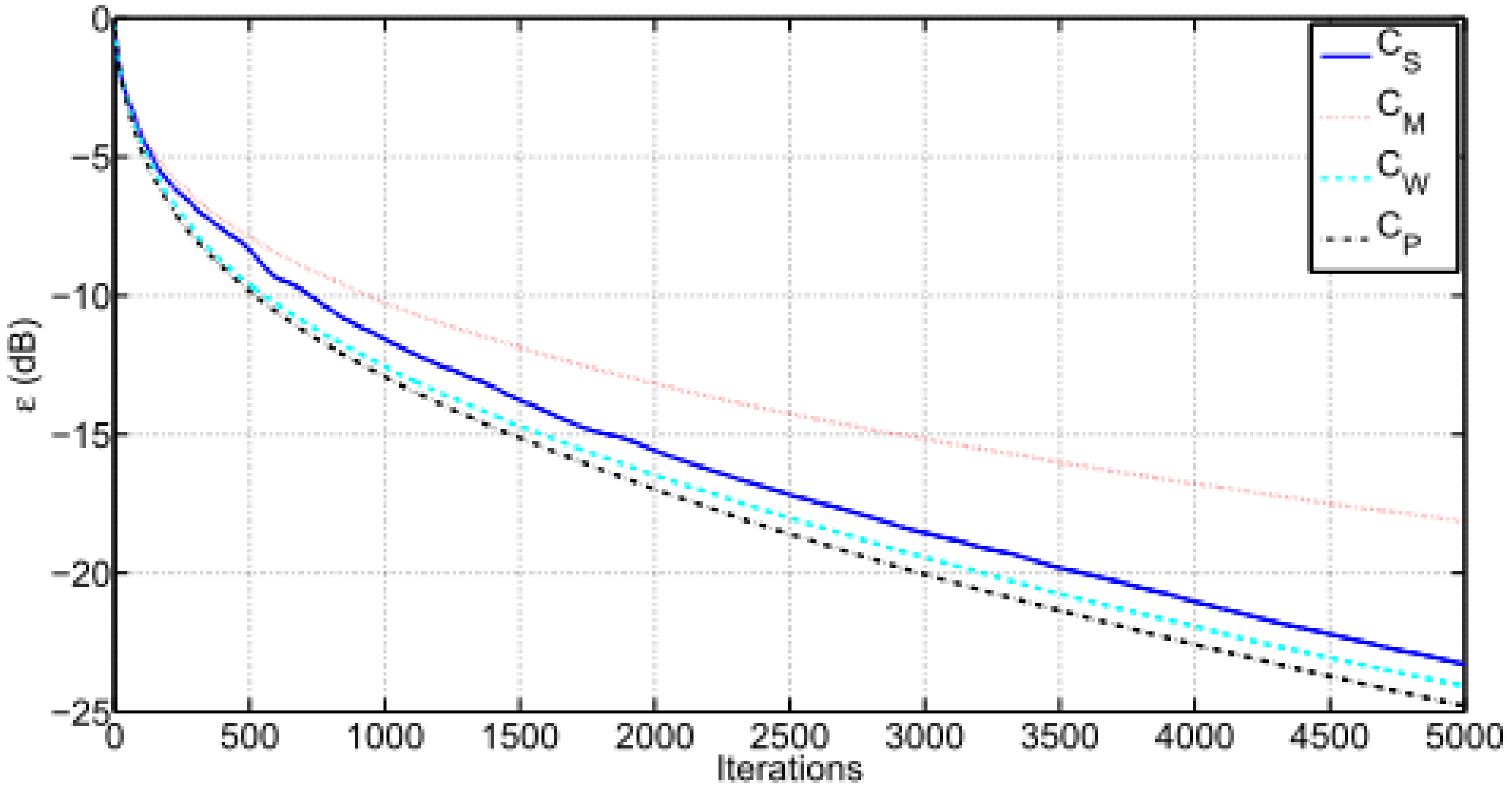
- **Favour background**

$$\mathcal{C}_W(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} w(\phi, \mathbf{R}^n). \sum_{i=0}^{I-1} r_i^n(\phi)$$

- **Penalized background**

$$\mathcal{C}_P(\mathbf{\Phi}, \mathbf{R}^n) = \arg\max_{\phi \in \mathbf{\Phi}} \sum_{i=0}^{I-1} r_i^n(\phi) + \sum_{i \neq j} |r_i^n(\phi) - r_j^n(\phi)|$$

Some results



- Strategy $\mathcal{C}_P$ gives best results both in terms of reconstruction error and source separation

Sound examples

- **Sound examples with a total of 10 000 atoms (e.g. a very low number of atoms)**
  - Original signal
  - Approximate (reconstruction)
  - Background estimate
  - Singing Voice estimate

Conclusion

- Greedy approaches allow to build specific representations for dedicated applications

- Sparsity, Structured sparsity, random sequences or coverage constraints are some of the potential extensions of the classical MP approach

- Open issues:
  - Build multi-objective representations
  - Build meaningful hierarchical and dynamic representations

Some References

References directly used in this presentation

- *M. Moussallam, L. Daudet, G. Richard, "Matching pursuits with random sequential subdictionaries", Signal Processing, 2012,*
  S. Fenet, M. Moussallam, Y. Grenier, G. Richard and L. Daudet, A Framework for Fingerprint-Based Detection of Repeating Objects in Multimedia Streams,
  M. Moussallam, G. Richard and L. Daudet, Audio Source Separation Informed by Redundancy with Greedy Multiscale Decompositions, in Proc. of Eusipco 2012.
- L. Daudet: *Audio Sparse Decompositions in Parallel,* IEEE Signal Processing Magazine, 2010

- ***Other references linked to this presentation:***

- *M. Mueller, D. Ellis, A. Klapuri, G. Richard « Signal Processing for Music Analysis, IEEE Trans. on Selected topics of Signal Processing, Oct. 2011*
- E. Ravelli, G. Richard, L. Daudet, Audio signal representations for indexing in the transform domain, IEEE Transactions on Audio, Speech and Language Processing, Vol. 18, No 3, March 2010, pp 434- 446
- P. Leveau, E. Vincent, G. Richard, L. Daudet, *Instrument-Specific Harmonic Atoms for Mid-Level Music Representation***,** IEEE Transactions on Audio, Speech and Language Processing, Volume 16, N°1 Jan. 2008 Page(s):116 - 128.

And a few web links….

http://perso.telecom-paristech.fr/~grichard/

http://www.tsi.telecom-paristech.fr/aao/en/

WIAMIS'13

- **Important dates**
  - Proposal for Special Session
    4th January 2013
  - Paper Submission
    8th March 2013
  - Acceptance Notification
    3rd May 2013
  - Camera-ready Papers
    24th May 2013

Organizing Committee

General Chairs
Gaël Richard
Telecom ParisTech, France
Slim Essid
Telecom ParisTech, France

Program Chairs
George Tzanetakis
University of Victoria, Canada
Noel O'Connor
Dublin City University, Ireland

Special Session Chair
Fernando Pereira
Instituto Superior Técnico - Instituto de Telecomunicações, Portugal

Local Organization Chair
Angélique Drémeau
Telecom ParisTech, France

Asian Liaison
Homer H. Chen
National Taiwan University, Taiwan

North American Liaison
Gerald Friedland
International Computer Science Institute, Berkeley, USA

South American Liaison
Pending

African Liaison
Meriem Jaïdane
Ecole Nationale d'Ingénieurs de Tunis, Tunisia

European Liaison
Petros Daras
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Industry Liaison
Pending

WIA^{2}MIS 3rd-5th July 2013

14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services

Telecom ParisTech, Paris, France

The International Workshop on Image and Audio Analysis for Multimedia Interactive Services is one of the main international events for the presentation and discussion of the latest technological advances in interactive multimedia services. The objective of the workshop is to bring together researchers and developers from academia and industry working in the areas of image, video and audio applications, with a special focus on analysis. After a series of successful meetings starting in 1997 in Louvain, WIAMIS 2013 will be held in Telecom ParisTech, Paris, France. As highlighted in the conference name and in the topics of interest below, this 14th edition will make more room for audio analysis and audio-driven multimedia analysis research.

See conference website for paper submission details
http ://wiamis2013.telecom-paristech.fr/

Topics of interest include, but are not limited to :
- Multimedia content analysis and understanding ;
- Content-based browsing, indexing and retrieval of images, video and audio ;
- Advanced descriptors and similarity metrics for multimedia ;
- Audio and music analysis, and machine listening ;
- Audio-driven multimedia content analysis ;
- 2D/3D feature extraction ;
- Motion analysis and tracking ;
- Multi-modal analysis for event recognition ;
- Human activity/action/gesture recognition ;
- Video/audio-based human behavior analysis ;
- Emotion-based content classification and organization ;
- Segmentation and reconstruction of objects in 2D/3D image sequences ;
- 3D data processing and visualization ;
- Content summarization and personalization strategies ;
- Semantic web and social networks ;
- Advanced interfaces for content analysis and relevance feedback ;
- Content-based copy detection ;
- Analysis and tools for content adaptation ;
- Analysis for coding efficiency and increased error resilience ;
- Multimedia analysis hardware and middleware ;
- End-to-end quality of service support ;
- Multimedia analysis for new and emerging applications ;
- Advanced multimedia applications.