Research in audio Processing

25th October, 2012
Gaël Richard

« Licence de droits d'usage"
http://formation.enst.fr/licences/pedago_sans.html
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
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

<table>
<thead>
<tr>
<th>Education</th>
<th>Continuing education &amp; life-long training</th>
<th>Research</th>
<th>Business start-up support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public teaching at the highest level in the domain of ICT</td>
<td>From theory to industrial transfer</td>
<td>Development of entrepreneurial spirit .. to Incubation</td>
<td></td>
</tr>
</tbody>
</table>
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

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 online
- Art and information technologies
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
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)

Aim of the group: 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.
Music signals processing

Multimedia streams analysis

heterogenous sensors arrays signal processing

Biological signals processing

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/aaol/en/

- **Databases**
  - ENST-Drums (2006)
  - MAPS (2009)
  - 3DLife ACM Multimedia Grand Challenge 2011 Dataset
  - Romeo-HRTF (2011)
  - QUASI (2012)
  - …

- **Softwares**
  - **Yaafé**: 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**
  - …

(some) Available tools for separation

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

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

**Leading voice extractor**
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

\[
\min ||\alpha||_0 \quad \text{s.t.} \quad f = \Phi \alpha
\]

Or alternatively

\[
\min K \quad \text{s.t.} \quad f = \sum_{k=1}^{K} \alpha_k \phi_k
\]
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 \( \Phi = \{ \phi_i \}_{i=0}^{M-1} \)

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

Or alternatively

\[
\min K \quad \text{s.t.} \quad \| f - \sum_{k=1}^{K} \alpha_k \phi_k \|_2 \leq \epsilon
\]
How to obtain the sparse approximation?

- Many existing approaches
  
  - Convex optimisation: $\| \cdot \|_0 \rightarrow \| \cdot \|_p$
  
  - 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.
Matching pursuit

Inputs $f$, $\Phi$

$R^0 f := f$

$\Gamma^0 := \emptyset$

$n := 1$

Select Atom

$\phi_{\gamma_n} = C(\Phi, R^{n-1} f)$

$\Gamma^n = \Gamma^{n-1} \cup \gamma_n$

Step 1

Update approximant

$f_n = A(f, \Phi \Gamma^n)$

$R^n f = f - f_n$

Step 2

$n := n + 1$

Stopping Criterion?

Yes

$R^n f$, $R^n f$

No

Examples

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

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\Phi$</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Dictionary</td>
<td>$C$</td>
<td>Dictionary of Gabor atoms</td>
</tr>
<tr>
<td>The Selection Rule</td>
<td>$A$</td>
<td>$\arg \max_{\phi, \gamma \in \Phi}</td>
</tr>
<tr>
<td>The Update Strategy</td>
<td></td>
<td>addition of new contribution</td>
</tr>
<tr>
<td>The Stopping Criterion</td>
<td></td>
<td>Signal to Noise Ratio</td>
</tr>
</tbody>
</table>
Matching pursuit

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

- Original
- Approximation; \( n=500 \)
- Residual \( N=500 \)
- Spectral representation
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-dictionaries 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
  - $\Gamma^n$ : Set of selected atoms
  - $f^n$ : Approximated signal after $n$ iterations
Random Sequential Subdictionaries MP (RSS-MP)

- **Performance**: close to 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…)

Diagram:
- Song1
- Program1
- Ad1
- Ad3
- Ad2
- Song1
- Ad4
- Song2
- Ad2
- Program1
Most fingerprint systems rely on the following transform of the signal:

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

\[
C_{\mathcal{M}}(R^n x, \Phi) = \arg \max_{\phi_i \in \Phi} (\|R^n x, \phi_i\| \mathcal{M}(\phi_i | \Gamma^n))
\]

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CQT (reference)</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (%)</td>
<td>95.1</td>
<td>94.5</td>
</tr>
<tr>
<td>Recall (%)</td>
<td>97.8</td>
<td>91.5</td>
</tr>
<tr>
<td>F-Measure (%)</td>
<td>96.5</td>
<td>93.0</td>
</tr>
<tr>
<td>CPU/segment (s)</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>Database (MB)</td>
<td>9.3</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Recall = Nb of good detected repetitions / Total nb of repetitions
Precision = Nb of good detected repetitions / Total nb of detections
2 real world corpora:

- 3 days of the same radio (72 h)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Télécom - CQT</th>
<th>Télécom - MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>Precision</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Télécom - CQT</th>
<th>Télécom - MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.97</td>
<td>0.78</td>
</tr>
<tr>
<td>Precision</td>
<td>0.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>
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

\[
\begin{align*}
X & \quad \Phi \\
N \times I & \quad N \times M & \quad M \times I \\
A_X & \quad B_X
\end{align*}
\]

- Separated the singing voice \( \Phi \cdot B_X \) and the background \( \Phi \cdot A_X \)
Modified MP algorithm

Input: $X$, $\Phi$

1. $R^0 := X$, $n = 0$
2. repeat
3. **Step 1**: Select atom $\phi_k \leftarrow C(\Phi, R^n)$
4. **Step 1 bis**: Decide if $\phi_k$ is background or not
5. if $\phi_k$ is background then
6. $\forall i, A_X[i, k] = \langle \phi_k, R^n_i \rangle$
7. else
8. Find which channels $J \subset I$, $\phi_k$ belongs to.
9. $\forall j \in J, B_X[j, k] = \langle \phi_k, R^n_j \rangle$
10. end if
11. **Step 2**: Update residual:
12. $R^n = X - \Phi.(A_X + B_X)$
13. $n \leftarrow n + 1$
14. until a stopping condition is met
Output: $R^n$, $A_X$ and $B_X$
Different selection rules

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

- **Energetic criterion**
  \[
  C_S(\Phi, R^n) = \arg \max_{\phi \in \Phi} \sum_{i=0}^{I-1} r_i^n(\phi)
  \]

- **Minimum risk**
  \[
  C_M(\Phi, R^n) = \arg \max_{\phi \in \Phi} \min_i r_i^n(\phi)
  \]

- **Favour background**
  \[
  C_W(\Phi, R^n) = \arg \max_{\phi \in \Phi} w(\phi, R^n). \sum_{i=0}^{I-1} r_i^n(\phi)
  \]

- **Penalized background**
  \[
  C_P(\Phi, R^n) = \arg \max_{\phi \in \Phi} \sum_{i=0}^{I-1} r_i^n(\phi) + \sum_{i \neq j} \left| r_i^n(\phi) - r_j^n(\phi) \right|
  \]
Some results

- Strategy $C_P$ 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

- S. Fenet, M. Moussallam, Y. Grenier, G. Richard and L. Daudet, A Framework for Fingerprint-Based Detection of Repeating Objects in Multimedia Streams,

Other references linked to this presentation:


And a few web links….

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