INFORMED AUDIO SOURCE SEPARATION

Gaël Richard
Institut Telecom, Telecom ParisTech, CNRS LTCI, France

With help from A. Liutkus, A. Ozerov

AES 53rd Conference on Semantic Audio
Audio recordings

- What is an audio recording?
Audio recordings

- What is an audio recording?

  - It is composed of *audio objects* or *sources*…

  - … Which are mixed together into a *mixture* (i.e. the audio recording) which is possibly multichannel (stereo is the most common for music)
Audio recordings

● What is an audio recording?

- It is composed of *audio objects* or sources…
  - piano, drums, guitar, ….

- …. Which are mixed together into a *mixture* (i.e. the audio recording) which is possibly multichannel (stereo is the most common for music)

● In most cases only the mixture is available which limits *Active Listening* capabilities …
Applications

What could we do if we had the separated audio objects?

- Active listening
- Karaoke
- Remixing
- Music information retrieval
  - Cover song detection,
  - Music transcription (audio-to-midi, instrument recognition,...)

- ....
From Source separation to Informed Source Separation

How to recover the audio objects?

- Using blind source separation
  - Separation is only done using the audio mixture.
  - But…quality is often not sufficient for active listening applications.
  - Exemple of Blind leading voice extraction [Durrieu&al.2011]…

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Backgrounds</th>
<th>Leading voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singing voice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trumpet</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Source separation to Informed Source Separation

■ How to recover the audio objects?

• Or … relying on Informed Source Separation (ISS)
  – Side information is transmitted to the separation module
  – Separation is done using the mixture and the side information
From Source separation to Informed Source Separation

- How to recover the audio objects?

  - Or … relying on Informed Source Separation (ISS)
    - Side information is transmitted to the separation module
    - Separation is done using the mixture and the side information

    - *Side information can be:*
      - Information about the sources (e.g. MIDI scores, information extracted from cover versions, types of the sources, etc. …)
      - Directly extracted from the source signals in an encoding stage but with an additional constraint: this information needs to be small
Keynote content

Objective

- To provide an overview of major trends in Informed Source Separation (ISS)

Outline of the keynote

- Introduction on Informed Source Separation
- Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
- Overview of three trends in ISS:
  - Auxiliary data-informed source separation,
  - User-guided source separation,
  - Coding-based informed source separation
- Conclusion
Source separation by filtering techniques

- **General principle:**
  - The sources are recovered by filtering the mixtures

\[
\hat{s} = \mathcal{F} \left\{ x, \Theta \right\}
\]
A popular model for audio source separation: NMF

- NMF = Non-negative Matrix Factorization

Image from R. Hennequin

Original spectrogram

“Activations”

“Templates or Atoms”
A popular model for audio source separation: NMF

- NMF does not necessarily provide a semantically meaningful decomposition in absence of “constraints”

*Templates correspond to musical notes*

- Templates are built from half of each note and are less semantically meaningful
- Activations are less sparse

\[ WH \approx V \]
A popular model for audio source separation: NMF

- How the template matrix $W$ and activation matrix $H$ are obtained [Lee&al. 1999]?

- Minimization of $D(V||WH)$

- Problem separately convex in $W$ and $H$ (for Euclidean and Kullback-leibler divergence)

- Resolution leads to multiplicative update rules

\[
\begin{align*}
H &\leftarrow H \odot \frac{W^TV}{W^T(WH)} \\
W &\leftarrow W \odot \frac{VH^T}{(WH)H^T}
\end{align*}
\]
A popular model for audio source separation: NMF

- What types of constraints can be used?
  - Harmonicity of the templates [Raczinsky&al.2007]
    - To have a decomposition in “harmonic notes”
  - Spectral smoothness of the templates [Bertin&al.2010]
    - To obtain realistic timbral notes
  - Temporal continuity of activation [Virtanen2007]
    - To take into account that note activations are not erratic
  - Sparsity of the activations [Hoyer04][Smaragdis08]
    - To take into account that not too many notes are played in a given time
A popular model for audio source separation: NMF

- An example of model-based constraints for main melody separation:

- The model: \( A_{\text{audio}} = V_{\text{voice}} + M_{\text{music}} \)
  - The voice \( V_{\text{voice}} \) follows a source filter production model: \( V_{\text{voice}} = S_{\text{source}} \ast F_{\text{filter}} \)
  - Each component (Voice and Music) is represented by separate NMF
An example of model constrained NMF for singing voice extraction

- Exploitation of a source/filter production model

- Exploitation of redundancy of the accompanying music
  - Simple NMF model for background music \( \sum^m \) et \( A^m \)

In Informed audio Source Separation (ISS), “a priori” constraints may be replaced (or completed) by specific “information”

- Overview of three trends in ISS:
  - Auxiliary data-informed source separation,
  - User-guided source separation,
  - Coding-based informed source separation
Overview of three trends in ISS

Auxiliary data-informed source separation,
User-guided source separation,
Coding-based informed source separation
Auxiliary data-informed source separation

“Score-informed” source separation

Musical Score

Midi representation of each track (or source)

Use the MIDI information
to guide audio separation

Separated tracks of improved quality

Auxiliary data-informed source separation
“Score-informed” source separation

- An example in the framework of NMF \((V = W \cdot H)\)

**Matrix W**: synthetic harmonic templates are defined for each note

**Matrix H**: Idealized activations obtained from the MIDI score

Due to multiplicative update rules, zero entries at the initialization stay at zero.
Auxiliary data-informed source separation
“Score-informed” source separation

- An example in the framework of NMF (V = W . H)

**Matrix W:** obtained after convergence

**Matrix H:** obtained after convergence

Null entries at init. remain null
Auxiliary data-informed source separation
“Score-informed” source separation

Demonstration: “left hand” – “right hand” separation

Original recording (Chopin)

MIDI synthesis of the score

Auxiliary data-informed source separation

“Text-informed” speech separation

- Extension of the source-filter model of Durrieu & al.

  - Observed signal is described as “Speech + background”
    \[ X = S + B \]

  - The speech S is modeled as an Excitation-Filter-Channel signal:
    \[ \hat{V}_S = \hat{V}_S^e \otimes \hat{V}_S^f \otimes \hat{V}_S^c \]

  - Spectrogram of S
  - Spectrogram of Excitation (or source)
  - Spectrogram of “Filter” (e.g. formants)
  - Spectrogram of channel (e.g. microphone, reverberation, …)
**Auxiliary data-informed source separation**

“Text-informed” speech separation

- How the text is used?

\[ \text{Text} \rightarrow \text{Speech synthesizer or Human speaking} \rightarrow \text{Speech example} \rightarrow \text{NMPCF model parameter estimation} \rightarrow \text{Wiener filtering} \rightarrow \text{Estimated speech and background} \]

\[ \text{NMPCF = Non Negative Matrix partial co-factorization} \]

Auxiliary data-informed source separation
“Text-informed” speech separation

- Each component of the speech model is represented by a NMF

\[
\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \left( \mathbf{W}^e \mathbf{H}^e_S \right) \odot \left( \mathbf{W}^\phi \mathbf{H}^\phi_S \right) \odot \left( \mathbf{w}^e_s \mathbf{i}_N^T \right) + \mathbf{W}_B \mathbf{H}_B
\]

- In this representation the text (which gives phonetic information) will directly give information on the matrix linked to what is said, which is: \( \hat{\mathbf{V}}^\phi_S \)
Auxiliary data-informed source separation

"Text-informed" speech separation: demonstration

Mixture = Speech + Music
Example produced by the user

Proposed example-guided source separation

Estimated speech
Estimated background
True sources

User
Example
Mixture
Overview of three trends in ISS

Auxiliary data-informed source separation,
User-guided source separation,
Coding-based informed source separation
User-guided source separation

- In this scenario, the user provides some partial information about the sources to be separated.

- Two illustrative examples:
  - Iterative source selection using a Graphical User Interface (GUI)
  - Hummed-query for main melody extraction
  - Both examples are based on Probabilistic Latent Component Analysis models (which are probabilistic models similar to NMF)
User-guided source separation
User-selection using a GUI

- The user paints the parts corresponding to the melody in the GUI
- Algorithm is re-run but with many zero values in the initial decomposition for the melody part
- Several iterations are possible

User-guided source separation
User-selection using a GUI

Demo

User-guided source separation
Hummed melody input

- The user hums the melody of the instrument track that he wish to separate
- The melody produced is used as information for separating the melody in the mixture

From https://ccrma.stanford.edu/~gautham/Site/Humming.html
User-guided source separation
Hummed melody input

- Demonstration: Video [Smaragdis & al. 2009]

Auxiliary data-informed source separation,
User-guided source separation,
Coding-based informed source separation
**Coding-based informed source separation**

- **Here, the information is obtained directly from the sources** (but the information needs to be well compressed to be useful)
- **Sources (or Audio objects) are known at a so-called encoding stage**

Note that informed source separation in this case shares many similarities with Spatial Audio Object Coding approaches (see [Ozerov&al.11] for a discussion)

---

Coding-based informed source separation

What type of information is in the “side information”

- Could be the sources but then no point of source separation and huge bandwidth
- Usually it is a partial information about the sources (obtained from the knowledge of the sources):
  - Time frequency activations of the two predominant sources [Parvaix & al.]
  - A compressed version of the source spectrograms (for example JPEG) [Liutkus & al.]

Coding-based informed source separation

What performances can be obtained?

Demo of CISS

- Original mix (7 sources)
- Demix signals (using 7 kbit/s per source for side info)

For comparison: AAC for a mono signal is around 32 – 64 kbit/s
Conclusion / Perspectives

■ Conclusion:
  • Audio source separation is an extremely challenging task, especially when considering real-world stereophonic full-tracks.
  • Blind separation techniques do exist, but their performance may be greatly improved by using any available information apart from the mere mixture.
  • The so-called Informed Source Separation was discussed with examples from three major trends, namely:
    – Auxiliary data-informed source separation,
    – User-guided source separation,
    – Coding-based informed source separation

■ Some perspectives
  • The type of information depends on the type of source separator and the application but how to limit the side-information to the minimum?
  • How to exploit several informed source separators (e.g. separator fusion) in an optimal way?
  • How to better exploit a multitrack cover version to perform source separation on the original recording?
  • ….
Additional References