Main Instrument Separation from Stereophonic Audio Signals Using a Source/Filter Model

J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard and B. David

Institut Télécom; Télécom ParisTech; CNRS LTCI
Presentation overview

- Introduction
- Proposed method
  - Signal model
  - System overview
- Results
- Conclusion
Introduction: MIR and Source Separation

- Music Information Retrieval (MIR) and Audio Source Separation: e.g. [Vincent, 06] or [Gillet and Richard, 08].

- Using the main melody to separate the corresponding instrument:
  - “Audio Melody Extraction” at MIREX since 2004.
  - Source separation inspired by [Ozerov et al., 07].

- Contributions:
  - Stereo extension of our system for mono.
  - Unvoiced part.
  - Smoothed filters.
Task definition 1/2

**Mixture** $X = V + M$

- **$V$**: "Main instrument", i.e. leading instrument.
  - Harmonic and mono-pitch,
  - Predominant energy,
  - Continuous melody line.

- **$M$**: accompaniment.
  - Multiple sources, multiple pitches, percussive sounds...
Task definition 2/2

- Stereo mixture: \( X = [X_R, X_L] \)
Model: stereophonic signal

- **Stereo mixture:** \( X = [X_R, X_L] \)
  - Instantaneous, “panning” effect.

- **Assumption on STFT distribution:**

\[
\begin{align*}
X_R, f_n &\sim N_c(0, \alpha_R^2 S_{V, f_n} + \sum_{j=1}^{J} \beta_{Rj}^2 S_{M, j, f_n}) \\
X_L, f_n &\sim N_c(0, \alpha_L^2 S_{V, f_n} + \sum_{j=1}^{J} \beta_{Lj}^2 S_{M, j, f_n})
\end{align*}
\]

- \( p(X) = p(X_R) p(X_L) \)
Model: unvoiced part in solo source/filter 1/2

\[ S_V \text{ product of 2 contributions:} \]

\[ S_V = S_\Phi \cdot S_{F_0} = S_\Phi \cdot [W_{F_0} H_{F_0}] \]

- \( W_{F_0} \) fixed dictionary of source spectra:
  - Voiced: spectral combs.
  - Unvoiced: “white noise”

- Assumption: unvoiced and voiced parts filtered by same spectral shapes.
Model: unvoiced part in solo source/filter 2/2
Model: smooth filters

\[ S_V = S_\Phi \cdot S_{F_0} = \left[ W_\Gamma H_\Gamma H_\Phi \right] \cdot S_{F_0} \]

- \( W_\Gamma \) fixed dictionary of smooth elements:
Model: accompaniment

J components:

\[
\begin{align*}
S_{M_r, fn} &= \sum_{j=1}^{J} w_{fj} \beta_{Rj}^2 h_{jn} = [W_M B_R H_M]_{fn} \\
S_{M_L, fn} &= \sum_{j=1}^{J} w_{fj} \beta_{Lj}^2 h_{jn} = [W_M B_L H_M]_{fn}
\end{align*}
\]

with

\[
\begin{align*}
B_R &= \text{diag} \left( [\beta_{R1}^2, \ldots, \beta_{Rj}^2, \ldots, \beta_{RJ}^2] \right) \\
B_L &= \text{diag} \left( [\beta_{L1}^2, \ldots, \beta_{Lj}^2, \ldots, \beta_{LJ}^2] \right)
\end{align*}
\]
Estimation of parameters

\[
\begin{align*}
S_{X_R} &= \alpha_R^2 (W \Gamma H \Gamma H_\Phi) \cdot (W_{F_0} H_{F_0}) + W_M B_R H_M \\
S_{X_L} &= \alpha_L^2 (W \Gamma H \Gamma H_\Phi) \cdot (W_{F_0} H_{F_0}) + W_M B_L H_M
\end{align*}
\]

- **ML Criterion:**
  - \( \ln p(X) = \ln p(X_R) + \ln p(X_L) \)

- **Estimation of parameters:**
  - Gradient method, analogous to classic NMF rules.
Proposed method: system overview

1\textsuperscript{st} estimation round

2\textsuperscript{nd} estimation round

3\textsuperscript{rd} estimation round

Melody

Mixture

X

\( H_{F_0} \)

\( \hat{H}_{F_0} \)

\( \Theta \)

\( \Theta \)

\( \hat{V}_{V-IMM} \)

\( \hat{M}_{V-IMM} \)

\( \hat{V}_{VU-IMM} \)

\( \hat{M}_{VU-IMM} \)
Results: database and evaluation criteria

- **Synthetic database:**
  - 13 synthetic instantaneous mixtures from separated tracks.

- **Evaluation criteria:**
  - *BSS Eval* criteria for SiSEC: **SDR, ISR, SIR, SAR**.
Results: effects of the contributions

- **Stereo vs. Mono**
  - ~2dB SDR gain
  - Resulting audio signal more “coherent”.

- **Smooth filters**
  - Practically no improvement,
  - More realistic, suitable for other applications.

- **Unvoiced model**
  - ~0.3dB SDR gain,
  - Catches drum signals...
  - … and misses some consonants.
Results: SiSEC08

- SiSEC'08, Professionally Produced Music Recordings
  - Development set: 2 stereo signals, 2 artists,
  - Test set: 2 stereo signals (from the same songs).

- 1 result submitted: Tamy.
  - Female singer + guitar,

- Results:
  - Success of algorithms based on melody tracking,
  - Music sound difficult because of guitar signal.
Conclusion

- Improvements over previous de-soloing system:
  - Explicit stereo model,
  - Smoothness of filter part,
  - Taking into account unvoiced parts.


- Perspectives:
  - Use in applications such as lyrics recognition,
  - Better unvoiced model,
  - Take into account more mutual information between the channels [Ozerov and Févotte, 09].
Conclusion

- Improvements over previous de-soloing system:
  - Explicit stereo model,
  - Smoothness of filter part,
  - Taking into account unvoiced parts.


- Perspectives:
  - Use in applications such as lyrics recognition,
  - Better unvoiced model,
  - Take into account more mutual information between the channels [Ozerov and Févotte, 09].

  Some examples + questions?
Additional material
Main instrument/Accompaniment model

[Durrieu et al., ICASSP'09]

\[ S_X = S_\phi \cdot S_{F_0} + S_M \]

\[ S_X = (W_\phi H_\phi) \cdot (W_{F_0} H_{F_0}) + (W_M H_M) \]

Filter  
Source  
Accompaniment

![Graphical representation of the model](image)
Spectral shapes for the estimated filters

\[ W_{\Phi,k}(dB) \]

\[ W_{\Phi,k}(dB) \]

\[ W_{\Phi,k}(dB) \]

Freq. (Hz) \( \times 10^4 \)

Freq. (Hz) \( \times 10^4 \)

Freq. (Hz) \( \times 10^4 \)
“Convergence”: evolution of Itakura Saito criterion over the iterations.

Evolution of the ML criterion w.r.t. iteration number: blue lines = first estimation, red lines = second estimation.
Audio source separation

Wiener filter:

\[ \hat{V}_R = \frac{\alpha_R^2 S_V}{X_R} \]

\[ \alpha_R^2 S_V + \sum_{j=1}^{J} \beta_{Rj}^2 S_{M,j} \]

\[ \hat{V}_R = \frac{\alpha_R^2 (W_{\Gamma} H_{\Gamma} H_{\Phi}) \cdot (W_{F_0} H_{F_0})}{\alpha_R^2 (W_{\Gamma} H_{\Gamma} H_{\Phi}) \cdot (W_{F_0} H_{F_0}) + W_M B_R H_M} X_R \]
Results: details, MTG MASS

- Average on our database: for each criterion, results given as solo/accompaniment:

<table>
<thead>
<tr>
<th>Method</th>
<th>SDR</th>
<th>ISR</th>
<th>SIR</th>
<th>SAR</th>
<th>gSDR</th>
<th>gSIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>5.8/6.9</td>
<td>9.0/21.8</td>
<td>16.8/9.6</td>
<td>5.8/11.5</td>
<td>6.9/5.8</td>
<td>17.8/8.5</td>
</tr>
<tr>
<td>V-IMM0</td>
<td>7.9/8.9</td>
<td>12.1/23.0</td>
<td>19.2/12.6</td>
<td>8.2/12.5</td>
<td>8.9/7.9</td>
<td>20.2/11.6</td>
</tr>
<tr>
<td>V-IMM1</td>
<td>7.9/8.9</td>
<td>12.5/22.1</td>
<td>18.4/12.8</td>
<td>8.3/11.6</td>
<td>8.9/7.9</td>
<td>19.4/11.8</td>
</tr>
<tr>
<td>VU-IMM0</td>
<td>8.2/9.3</td>
<td>12.4/23.3</td>
<td>19.9/12.9</td>
<td>8.7/12.7</td>
<td>9.3/8.2</td>
<td>20.9/11.8</td>
</tr>
<tr>
<td>VU-IMM1</td>
<td>8.2/9.3</td>
<td>13.0/21.8</td>
<td>18.6/13.2</td>
<td>8.8/12.0</td>
<td>9.3/8.2</td>
<td>19.6/12.2</td>
</tr>
</tbody>
</table>
### SiSEC'08

<table>
<thead>
<tr>
<th>System</th>
<th>Singer SDR</th>
<th>Guitar SDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancela2</td>
<td>9.7</td>
<td>8.6</td>
</tr>
<tr>
<td>VU-IMM</td>
<td>7.8</td>
<td>9.4</td>
</tr>
<tr>
<td>Cancela1</td>
<td>8.7</td>
<td>8.0</td>
</tr>
<tr>
<td>V-IMM</td>
<td>6.9</td>
<td>8.6</td>
</tr>
<tr>
<td>Cobos</td>
<td>6.4</td>
<td>8.0</td>
</tr>
<tr>
<td>Ozerov</td>
<td>5.1</td>
<td>6.7</td>
</tr>
<tr>
<td>Ozerov/Févotte</td>
<td>3.6</td>
<td>5.3</td>
</tr>
<tr>
<td>Vinyes Raso</td>
<td>4.9</td>
<td>4.2</td>
</tr>
<tr>
<td><strong>Ideal Binary Mask</strong></td>
<td><strong>10.1</strong></td>
<td><strong>11.8</strong></td>
</tr>
</tbody>
</table>