Musical Source Separation Using Time-Frequency Source Priors

Presented by
Adam Cohen
Scope

- Paper Review for “Speech Processing in Noisy Environment” course.
- “Musical Source Separation Using Time-Frequency Source Priors” by Emmanuel Vincent
Agenda

- Problem Definition
- Introduction
- Solution Overview
  - Model Presentation
  - Model Estimation
  - Instrument Parameter Learning
  - Filtering
- Algorithm Evaluation
- Conclusion
Problem Definition

Known Priors

\[ \begin{align*}
\theta_1 & \rightarrow m_1 \\
\theta_1 & \rightarrow m_2 \\
\theta_1 & \rightarrow m_3 \\
\end{align*} \]
Introduction

- Applications
- Main Problems
- Existing Methods
  - CASA
  - Statistical Spatial models
  - Statistical Spectral models
Introduction - Basic Assumptions

- Known instrument and locations (priors)
- Two microphone (technical)
- Channel Model:

\[ x_i = \sum_{j=1}^{J} a_{i,j} * s_j + n_i \]

- \( J \) number of instruments
- \( n_i \) is Gaussian noise

- Analysis in the STFT domain
Solution Overview

- Source Modeling for spectrum estimation
- Model require parameter learning
  - Offline instrument parameter learning
  - Mixture modeling to estimate instrument state
- Source filtering for separation
Solution Overview

- Prior
- Wiener Filter
- Model
- Instrument Learning (offline)

\[ M \]

\[ \Theta \]

\[ S_{\text{img1,1}} \]
\[ S_{\text{img1,2}} \]
\[ S_{\text{img2,1}} \]
\[ S_{\text{img2,2}} \]

STFT \[ X_1 \]
\[ X_2 \]

STFT\(^{-1}\)
Model Presentation

● Target:
  - Estimate instrument (source) spectrum
  - Estimate environment parameters
  - Estimation is used for filtering

● Means
  - Three Layer Bayesian Model
  - Learned Parameters
  - MAP

● Input: $X_i$
Model Presentation - Overview

State
- Active Note

Source
- Instrument
- Spectrum

Mixture
- Environment

\[ E_{j,h,t} \]
\[ m_{j,t,f} \]

\[ P^{sta} = p(E_{j,h,t}) \quad P^{src} = p(m_{j,t,f} \mid E_{j,t}, M) \quad P^{mix} = p(o_t \mid m_{j,t}, \Theta) \]
Model Presentation - Preprocess

- **Input:**
  
  \[ O_{pow} = \log \left( \frac{\|x_1\|^2 + \|x_2\|^2}{g_f} + 1 \right) \]

  \[ O_{pha} = \langle x_1, x_2 \rangle \]

  \[ O_{coh} = \frac{\|\langle x_1, x_2 \rangle\|}{\|x_1\| \cdot \|x_2\|} \]

- **Filter Bank:** logarithmic bandwidth
Model Presentation - State

- Target: probability of active note
  \[ P^{sta} = p(E_{j,h,t}) \]

- Factorial Model:
  - State is independent Bernoulli Process
  \[ P^{sta} = \prod_{j=1}^{J} \prod_{t=0}^{T-1} (1 - Z)^{#A_{j,t}} Z^{#I_{j,t}} \]
  - Z is fixed
  - Higher Z implies high probability for low number of active note

- Segmental Model
Model Presentation - Source

- **Target:** Probability of spectra given State
- **Assumptions:**
  - Different note are uncorrelated
  - Note Spectrum is stationary
  - Active note are log-Gaussian independent

\[
m_{j,t,f} = \sum_{h=H_j}^{H_j} e_{j,h,t} \Phi_{j,h,f}
\]

\[
P_t^{src} = p(m_{j,t,f} \mid E_{j,t}, \Phi_{j,h}, \mu_{j,h}, \sigma_{j,h}^e) = \prod_{j=1}^{J} \prod_{h \in A_{j,t}} N(\log e_{j,h,t}; \mu_{j,h}, \sigma_{j,h}^e)
\]

- \(\Phi\) the spectrum of note (instrument learn)
- \(\mu, \sigma\) Model parameters (instrument learn)
Model Presentation - Mixture

- Target: Channel model – \( P_t^{mix} = p(o_t \mid m_{j,t}, n, a^{pow}) \)
- Mixing filters are modeled via \( a^{pow} \) & \( a^{pha} \)
  - \( a^{pha} \) phase response is known from the direction:
    \[
    a_{j}^{pha} = 2\pi \sin \theta_j \mod 2\pi
    \]
  - \( a^{pow} \) power response - environment parameter (estimate)
  - Experiments shows need to estimate \( a^{phase} \)
  - \( n \) noise spectrum - environment parameter (estimate)

- Mono Model
- Stereo Model
Model Estimation

- **Problem:**
  \[ \hat{s}_{i,j} = \arg \max P(s_{i,j} \mid x_i, \Theta, M) \]
  - Integral can’t be solved

- **Approximation:**
  \[ \hat{s}_{i,j} \approx \arg \max P(s_{i,j} \mid x_i, \hat{\Theta}, \hat{m}) \]
  - Algorithm target
  - Model target
  \[ \hat{\Theta}, \hat{m}_{jt} = \arg \max P(\Theta, m_{jt} \mid o_t, M) \]

- **Weighted Bayes Law:**
  \[ P^{tot} = P(o_t, m_{jt}, E_{jt}, \Theta, M) \propto \left( \prod_{t=0}^{T-1} P^\text{mix}_t \right)^{\omega_{\text{mix}}} \left( \prod_{t=0}^{T-1} P^\text{src}_t \right)^{\omega_{\text{src}}} P^{\text{sta}} \]
Model Estimation - Blocks

Known Instruments and directions

Preprocess

MAP Estimator

Newton Method

Inference Step

State Search

\( X_1 \)

\( X_2 \)

\( M \)

\( X^{\text{pha}} \)

\( X^{\text{pow}} \)

\( X^{\text{coh}} \)

\( e_{jh} \)

\( p \)

\( \Theta \)

\( m_i \)
Model Estimation – Inference Step

- For each given state estimate best spectrum
- Estimate source spectrum by note power \( e \)
  - MAP is solved by Iterative Newton method:
    \[
    \log e_{jht} \leftarrow \log e_{jht} + \Delta \log e_{jht}
    \]
  - \( \Delta \log(e) \) is calculated by derivative of \( P^{\text{tot}} \)
- Spectrum is built using \( \Phi \)
- Similar Method for Environment parameters
Model Estimation – State Search

- Target: Search $E_{jht}$ for optimization
- State Space is too large (more than $10^8$)
- Reducing the state search by heuristic
  - Search near previous state
  - Estimate state using another method
  - Using beam search methods
  - Finding inactive state
Instrument Parameters Learning

- Target: learn $M=(\Phi, \mu^e, \sigma^e)$ for instrument
  - ML
  - Approximation:
    \[ \hat{M} = \arg \max P(M \mid o_t, \Theta, E_{jt}) \]
    \[ \hat{M} \approx \arg \max P(M \mid o_t, \Theta, \hat{m}_{jt}) \]
    \[ \hat{m}_{jt} = \arg \max P(m_{jt} \mid o_t, \Theta, E_{jt}, M) \]
  - Solve by expectation maximization (EM) algorithm

- Offline learning from database
  - Single note learn
Filtering

- **Motivation**
  - Spectrum estimation is unnatural
  - Solving $\hat{s}_{i,j} \approx \arg\max P(s_{i,j} \mid x_i, \Theta, \hat{m})$

- **Wiener Filter is used:**

$$s_{img}(i, j) = \frac{a^{pow} \hat{m}_j}{\sum_{j=1}^{J} a^{pow} \cdot \hat{m}_j + n_f} x_i$$

- Every instrument have two outputs
Algorithm Evaluation

- Instrument parameters learn from music database
- Two instruments in two test Scenarios:
- Criteria:
  - SDR: Signal to Distortion Ration
  - SIR: Signal to Interference Ration
  - SAR: Signal to Artifact Ration
  - Human judgment
- First time for musical source separation?
# Algorithm Evaluation – Result Mix 1

<table>
<thead>
<tr>
<th>Mixture</th>
<th>Clarinet (dB)</th>
<th></th>
<th>Violin (dB)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
</tr>
<tr>
<td>ICA</td>
<td>-1</td>
<td>4</td>
<td>4</td>
<td>-9</td>
</tr>
<tr>
<td>Spatial Masking</td>
<td>4</td>
<td>15</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Mono factorial</td>
<td>14</td>
<td>25</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Mono segmental</td>
<td>18</td>
<td>37</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>Stereo factorial</td>
<td>15</td>
<td>28</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>Stereo segmental</td>
<td>16</td>
<td>37</td>
<td>16</td>
<td>13</td>
</tr>
</tbody>
</table>
## Algorithm Evaluation – Result Mix 2

<table>
<thead>
<tr>
<th>Mixture</th>
<th>Cello (dB)</th>
<th></th>
<th>Violin (dB)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SDR</td>
</tr>
<tr>
<td>ICA</td>
<td>-3</td>
<td>7</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>Spatial Masking</td>
<td>4</td>
<td>14</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Mono factorial</td>
<td>-17</td>
<td>12</td>
<td>-15</td>
<td>1</td>
</tr>
<tr>
<td>Stereo factorial</td>
<td>7</td>
<td>26</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Stereo segmental</td>
<td>13</td>
<td>39</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>
Conclusion

- The paper doesn’t suggest new approach
  - A new application for exciting methods
- The paper achieve good result on source separation
  - Doesn’t use all space information
  - Test scenarios aren’t complete
- Possible improvements
  - Note estimation
  - Online instrument modifications
  - More history/future use
- New approach is needed for complete solution
- Single sensor maximum performance
References

- E. Vincent “Musical Source Separation Using Time-Frequency Source Priors”
- T. Kinoshita, S. Saki and H. Tanaka “Musical Sound Source identification based on frequency component adaptation”
- L. Benaroya, R. Gribonval, and F. Bimbot, “Non negative sparse representation for wiener based source separation with a single sensor”
- E. Vincent and X. Rodet “underdetermined source separation with structured source priors”
Model Presentation - State

- **Segmental Model**
  - Using prior temporal persistence
  - Markov chain model
  - No chords

### Mathematical Expression

\[
P_{\text{sta}} = \prod_{j=1}^{J} \prod_{r=0}^{R_j-1} D_{\text{not}}(d_r) \prod_{r=0}^{R_j-1} D_{\text{seg}}(t_{r+1} - t_r)(1-Z)^{#A_{j,0}} Z^{#I_{j,0}} \times \prod_{r=1}^{R_j-1} (#I_{j,t_r-1})^{-1}
\]

### Where

\[
D_{\text{not}}(d) = \begin{cases} 
\sum_{d' \geq d^n} N(\log d'; \mu^n, \sigma^n), & \text{if } d \geq d^n \\
0, & \text{otherwise}
\end{cases}
\]
Model Presentation – Mixture Mono

- Mono Model
  - Ignore phase information
    \[ o_t^{\text{pow}} = \log\left( \sum a_j^{\text{pow}} m_j + n_f \right) + \epsilon_t^{\text{pow}} \]
  - \( \epsilon \) is the residual error: Gaussian
  - Fix std \( \sigma^{\epsilon_{\text{pow}}} \)

\[
P_t^{\text{mix}} (\epsilon_t^{\text{pow}}) \propto N(0, \sigma^{\epsilon_{\text{pow}}})
\]
Model Presentation – Mixture Stereo

- Consider phase difference & azimuth cue
  - Different source are uncorrelated
  - Require estimate of noise phase

\[ a_{tf}^{\text{pha}} = \angle \left( \sum_{j=1}^{J} a_j^{\text{pow}} m_{jtf} \exp \left( i \theta_j^{\text{pha}} \right) + n_f \exp \left( i \theta_f^{\text{pha}} \right) \right) + \epsilon_{tf}^{\text{pha}} \mod 2\pi \]

\[ \sigma_{tf}^{\text{pha}} = \sigma^{\text{pha}} (1 - \rho_f^{\text{coh}})^{\lambda^{\text{pha}}} \]

- Low coherence less phase information

\[ P_{t_{\text{mix}}} = \prod_{f=1}^{F-1} \mathcal{N} \left( \epsilon_{tf}^{\text{pow}} ; 0, \sigma_{tf}^{\text{pow}} \right) \frac{\mathcal{N} \left( \epsilon_{tf}^{\text{pha}} ; 0, \sigma_{tf}^{\text{pha}} \right)}{\int_{-\pi}^{\pi} \mathcal{N} \left( \epsilon ; 0, \sigma_{tf}^{\text{pha}} \right) d\epsilon} \]
Model Presentation – Inference Step

- Derivative:

\[
\frac{\partial \log P_{\text{tot}}}{\partial \log e_{jht}} = \frac{w_{\text{mix}}}{\sigma_{\text{pow}}^2} \sum_{f=0}^{F-1} \epsilon_{tf}^{\text{pow}} \pi_{jhtf} - \frac{w_{\text{src}}}{\sigma_{jht}^2} (\log e_{jht} - \mu_{jht}^e)
\]

- For the mono model

\[
\Delta \log e_{jht} = \frac{\frac{w_{\text{mix}}}{\sigma_{\text{pow}}^2} \epsilon_{jht}^{\text{pow}} \pi_{jht} - \frac{w_{\text{src}}}{\sigma_{jht}^2} (\log e_{jht} - \mu_{jht}^e)}{\frac{w_{\text{mix}}}{\sigma_{\text{pow}}^2} \pi_{jht} + \frac{w_{\text{src}}}{\sigma_{jht}^2}}
\]

Where

\[
\pi_{jhtf} = \frac{a_f^{\text{pow}} e_{jht} \Phi_{jhtf}}{\sum_{j=1}^{J} a_f^{\text{pow}} m_{jtf} + n_f}
\]

- When note is masked $\pi = 0$: $\sigma$ is irrelevant