Polyphonic Music Transcription using Shift-invariant Latent Variable Models

Emmanouil Benetos and Simon Dixon
{emmanouilb,simond}@eecs.qmul.ac.uk

Centre for Digital Music
Queen Mary University of London

November 2011
Centre for Digital Music (c4dm):

- Formed in 2003
- About 50 full-time members (academic staff, research staff, research students)
- Research areas:
  - Music informatics
  - Machine listening
  - Audio engineering
  - Interactional sound & music
  - Music cognition
Outline

1 Introduction

2 SIPLCA-based Transcription System

3 HMM-constrained SIPLCA for pitch detection

4 HMM-constrained SIPLCA for multi-pitch detection

5 Conclusions and Future Work
Automatic music transcription:
Applications:
- Music information retrieval
- Interactive music systems
- Computational musicology

Subtasks:
- Pitch estimation
- Onset/offset detection
- Instrument identification
- Rhythmic parsing
- Identification of dynamics/expression
- Typesetting

Core problem: Multi-pitch estimation

Still remains an open problem
Probabilistic Latent Component Analysis (PLCA): probabilistic version of NMF, easy to generalize and interpret:

\[ P(\omega, t) = P(t) \sum_{z} P(\omega|z)P(z|t) \]  

(1)

\( P(\omega, t) \) is the input spectrogram, \( P(t) \) the frame energy, \( P(\omega|z) \) the spectral template for each component, and \( P(z|t) \) the component gain.

Unknown parameters estimated via EM algorithm.
Shift-Invariant Probabilistic Latent Component Analysis (SIPLCA) (Smaragdis09): extract shift-invariant structures in non-negative data

\[
P(\omega, t) = \sum_z P(z)P(\omega | z) *_\omega P(f, t | z)
\]  \hspace{1cm} (2)

\(P(\omega, t)\) is the log-frequency spectrogram, \(P(z)\) the source prior, and \(P(f, z | t)\) pitch impulse distribution.

Systems for multi-pitch estimation expanding on SIPLCA in Mysore09, Fuentes11
Figure: SIPLCA example for a violin glissando

\[ \begin{align*}
\omega & \quad 0 & 200 & 400 & 600 & 800 & 1000 & 1200 \\
0 & & & & & & \\
0.02 & & & & & & \\
0.04 & & & & & & \\
0.06 & & & & & & \\
\end{align*} \]

\[ \begin{align*}
f & \quad 0 & 20 & 40 & 50 \\
10 & & & & & & \\
20 & & & & & & \\
30 & & & & & & \\
40 & & & & & & \\
50 & & & & & & \\
\end{align*} \]
E. Benetos and S. Dixon, “Multiple-instrument polyphonic music transcription using a convolutive probabilistic model”, in SMC 2011. (+MIREX11 participation)

Motivation: framework with multiple templates per pitch, instrument
Contribution of each instrument source is pitch- and time-dependent
Supports frequency modulations and tuning changes
Transcription model:

\[ V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{p,s} P(\omega|s, p) * \omega P(f|p, t)P(s|p, t)P(p|t) \] (3)

- \( P(\omega, t) \): input log-frequency spectrogram
- \( P(t) \): frame energy
- \( P(\omega|s, p) \): spectral templates for instrument \( s \) and pitch \( p \)
- \( P(f|p, t) \): pitch impulse distribution
- \( P(s|p, t) \): time- and pitch-dependent instrument contribution
- \( P(p|t) \): piano-roll transcription.

- Shifting for each template spans one semitone
- Unknown parameters estimated using EM algorithm
Pitch template extraction: MAPS and RWC databases, using PLCA

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Lowest note</th>
<th>Highest note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cello</td>
<td>26</td>
<td>81</td>
</tr>
<tr>
<td>Clarinet</td>
<td>50</td>
<td>89</td>
</tr>
<tr>
<td>Flute</td>
<td>60</td>
<td>96</td>
</tr>
<tr>
<td>Guitar</td>
<td>40</td>
<td>76</td>
</tr>
<tr>
<td>Harpsichord</td>
<td>28</td>
<td>88</td>
</tr>
<tr>
<td>Oboe</td>
<td>58</td>
<td>91</td>
</tr>
<tr>
<td>Organ</td>
<td>36</td>
<td>91</td>
</tr>
<tr>
<td>Piano</td>
<td>21</td>
<td>108</td>
</tr>
<tr>
<td>Violin</td>
<td>55</td>
<td>100</td>
</tr>
</tbody>
</table>

Table: MIDI note range of the extracted instrument templates.
Sparsity encouraged on:
- \( P(p|t) \) - few notes active for each time frame
- \( P(s|p, t) \) - each note produced by few instruments

Postprocessing on \( P(p, t) = P(t)P(p|t) \) using on/off HMMs for each pitch (note smoothing/tracking)
- State priors and transitions computed from RWC database MIDI files
Figure: The time-pitch representation $P(f, t)$ for RWC-MDB-C-2001 No. 12 (string quartet). Original recording: 🎧 Synthesized transcription: 🎧
More examples: http://www.eecs.qmul.ac.uk/~emmanouilb/transcription.html
HMM-constrained SIPLCA for pitch detection (1)

**Motivation:** A musical note could be expressed by a sequence of sound states (e.g. attack, transient, sustain, decay)

![CQT spectrogram of a C1 piano note](image)

**Figure:** The CQT spectrogram of a C1 piano note.
Related Work:

- Non-negative Hidden Markov Model (N-HMM) in Mysore10
- NMF with Markov-chained bases in Nakano10

Proposed Model:

- **Goal**: represent produced notes as a sequence of sound state spectral templates, also shifted across log-frequency
- Attempt to address current drawbacks of spectrogram factorization-based methods for pitch estimation
Model: HMM-constrained SIPLCA, where each component corresponds to a sound state

Formulation:

\[ V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{q_t} P_t(q_t|\bar{\omega}) P(\omega|q_t) * \omega \ P_t(f|q_t) \]  \hspace{1cm} (4)

- \( P(\omega, t) \): log-frequency spectrogram approximation (input spectrogram: \( V_{\omega,t} \))
- \( P(t) \): spectrogram energy
- \( P_t(q_t|\bar{\omega}) \): time-varying sound state contribution
- \( P(\omega|q_t) \): sound state spectral template
- \( P_t(f|q_t) \): pitch impulse distribution
HMM-constrained SIPLCA for pitch detection (4)

Temporal constraints for sound states using HMMs:

\[ P(\bar{\omega}) = \sum_{\bar{q}} \sum_{\bar{f}} P(q_1) \prod_{t} P(q_{t+1}|q_t) \prod_{t} P_t(\omega_t|q_t) \]  

(5)

\( P(q_1) \): state prior distribution
\( P(q_{t+1}|q_t) \): transition probability
\( P_t(\omega_t|q_t) \): time-dependent observation probability

Observation probability:

\[ P_t(\omega_t|q_t) = 1 - \frac{\| P(\omega, t|q_t) - V_{\omega,t} \|_2}{\sum_{q_t} \| P(\omega, t|q_t) - V_{\omega,t} \|_2} \]  

(6)

Thus, the sound state spectrogram that better approximates the input spectrogram using the \( l^2 \) norm has a greater observation probability.
HMM-constrained SIPLCA for pitch detection (5)

- **Parameter Estimation**: using the EM algorithm
- **Update equations** are a combination of the SIPLCA update rules and the HMM forward-backward procedure
- **Expectation step**:

  \[
  P_t(f, q_t | \bar{\omega}) = \frac{P(\omega - f | q_t)P_t(f | q_t)}{\sum_f P(\omega - f | q_t)P_t(f | q_t)} \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_t} \alpha_t(q_t)\beta_t(q_t)}
  \]  

  where \( \alpha_t(q_t) \) and \( \beta_t(q_t) \) are the HMM forward and backward variables.

  \[
  P_t(q_t, q_{t+1} | \bar{\omega}) = \frac{\alpha_t(q_t)P(q_{t+1} | q_t)\beta_{t+1}(q_{t+1})P_t(\omega_{t+1} | q_{t+1})}{\sum_{q_t, q_{t+1}} \alpha_t(q_t)P(q_{t+1} | q_t)\beta_{t+1}(q_{t+1})P_t(\omega_{t+1} | q_{t+1})}
  \]
HMM-constrained SIPLCA for pitch detection (6)

- **Maximization step:**

\[
P(\omega|q) = \frac{\sum_{f,t} V_{\omega+f,t} P_t(f, q_t|\omega + f)}{\sum_{\omega,f,t} V_{\omega+f,t} P_t(f, q_t|\omega + f)}
\]

(9)

\[
P_t(f|q_t) = \frac{\sum_{\omega} V_{\omega,t} P_t(f, q_t|\omega)}{\sum_{f,\omega} V_{\omega,t} P_t(f, q_t|\omega)}
\]

(10)

\[
P(q_{t+1}|q_t) = \frac{\sum_t P_t(q_t, q_{t+1}|\bar{\omega})}{\sum_{q_{t+1}} \sum_t P_t(q_t, q_{t+1}|\bar{\omega})}
\]

(11)

\[
P(q_1) = P_1(q_1|\bar{\omega})
\]

(12)
HMM-constrained SIPLCA for pitch detection (7)

Figure: Decomposition of a piano sound (C4) using the proposed model.
Figure: Sound state contribution for a piano melody.
HMM-constrained SIPLCA for multi-pitch detection (1)

- Extending the single-pitch, single-source model for multi-pitch detection of multiple instrument sources
- Utilizing multiple pitch-wise HMMs or factorial HMMs
HMM-constrained SIPLCA for multi-pitch detection (2)

- **Model**: multiple-HMM-constrained SIPLCA
- **Formulation**:
  \[
  V_{\omega,t} \approx P(\omega, t) = \\
  P(t) \sum_{s, p} P_t(p) P_t(s | p) \sum_{q^{(p)}} P_t(q_t^{(p)} | p, \bar{\omega}) P(\omega | s, p, q_t^{(p)}) * \omega P_t(f | p)
  \]
  \[(13)\]
  - **Temporal constraints** through pitch-wise HMMs:
  \[
  P(\bar{\omega}) = \sum_{\bar{q}^{(p)}} \sum_{\bar{s}} \sum_{\bar{p}} \sum_{\bar{f}} P(q_1^{(p)}) \prod_t P(q_{t+1}^{(p)} | q_t^{(p)}) \prod_t P(\omega_t | q_t^{(p)})
  \]
  \[(14)\]
HMM-constrained SIPLCA for multi-pitch detection (3)

Figure: (a) Pitch spectrogram \( P(f, t) \) of an excerpt of “RWC-MDB-J-2001 No. 7” (guitar). (b) The pitch ground truth of the same recording. The abscissa corresponds to 10ms. Original recording: 🎸 Synthesized transcription: 🎹
Conclusions and Future Work

Conclusions:

- Proposed models for automatic music transcription based on SIPLCA
- Transcription results are very promising (e.g. MIREX11)
- HMM-constrained models can capture the temporal evolution of musical sounds

Future Directions:

- Instrument identification using the current framework
- Model note durations in HMM-constrained models
- Joint multi-pitch detection and note tracking steps