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

Queen Mary University of London

Centre for Digital Music

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

centre for digital music

Queen Mary
University of London

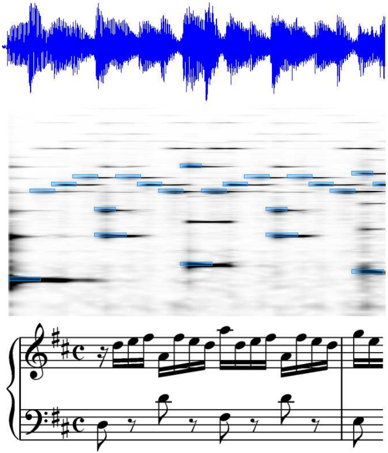
Outline

centre for digital music

Queen Mary University of London

Introduction (1)

Automatic music transcription:



centre for digital music

Queen Mary University of London

Introduction (2)

- 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

PLCA

- Probabilistic Latent Component Analysis (PLCA): probabilistic version of NMF, easy to generalize and interpret:

$$P(\omega, t) = P(t) \sum_z P(\omega|z)P(z|t) \tag{1}$$

$P(\omega, t)$ 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

SIPLCA (1)

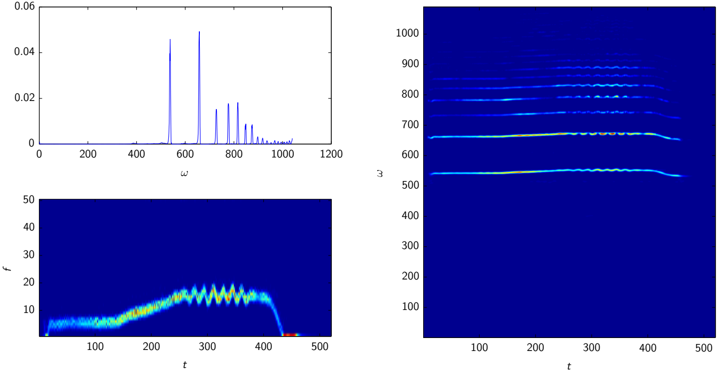
- Shift-Invariant Probabilistic Latent Component Analysis (SIPLCA) (Smaragdis09): extract shift-invariant structures in non-negative data

$$P(\omega, t) = \sum_z P(z) P(\omega|z) *_\omega P(f, t|z) \tag{2}$$

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

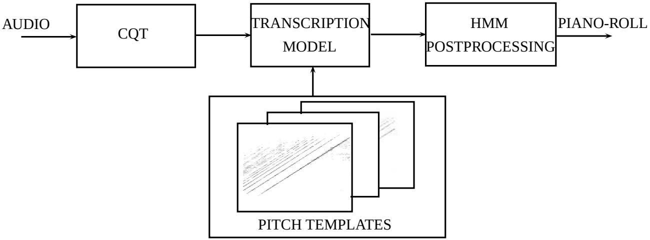
- Systems for multi-pitch estimation expanding on SIPLCA in Mysore09, Fuentes11

Figure: SIPLCA example for a violin glissando



SIPLCA-based Transcription System (1)

- 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

SIPLCA-based Transcription System (2)

- Transcription model:

$$V_{\omega,t} \approx P(\omega,t) = P(t)\sum_{p,s} P(\omega|s,p) *_{\omega} P(f|p,t)P(s|p,t)P(p|t) \quad (3)$$

$P(\omega,t)$: 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

centre for digital music

Queen Mary University of London

SIPLCA-based Transcription System (3)

Pitch template extraction: MAPS and RWC databases, using PLCA

Instrument	Lowest note	Highest note
Cello	26	81
Clarinet	50	89
Flute	60	96
Guitar	40	76
Harpsichord	28	88
Oboe	58	91
Organ	36	91
Piano	21	108
Violin	55	100

Table: MIDI note range of the extracted instrument templates.

SIPLCA-based Transcription System (4)

- 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

centre for digital music

Queen Mary University of London

SIPLCA-based Transcription System (5)

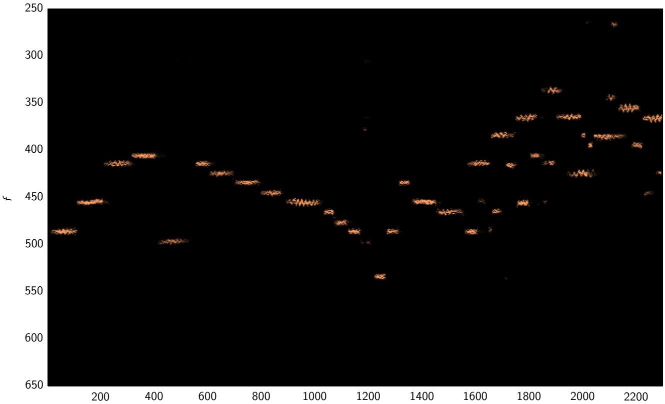


Figure: The time-pitch representation $P(f,t)$ for RWC-MDB-C-2001 No. 12 (string quartet). Original recording: 🔊 Synthesized transcription: 🔊
More examples: http://www.eecs.qmul.ac.uk/∼emmanouilb/transcription.html

centre for digital music

Queen Mary University of London

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)

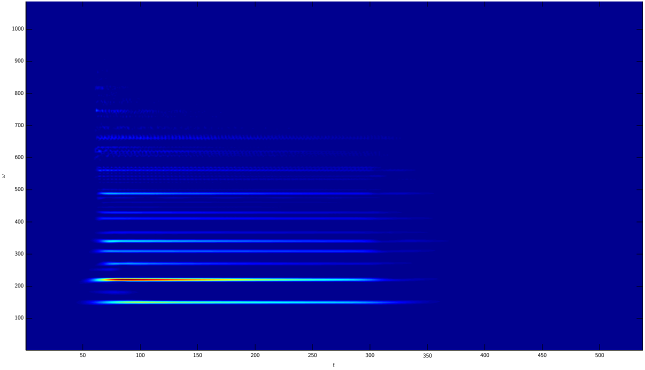


Figure: The CQT spectrogram of a C1 piano note.

HMM-constrained SIPLCA for pitch detection (2)

Related Work:

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

Proposed Model:

- E. Benetos and S. Dixon, "A temporally-constrained convolutive probabilistic model for pitch detection", in WASPAA 2011.
- 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

HMM-constrained SIPLCA for pitch detection (3)

- Model: HMM-constrained SIPLCA, where each component corresponds to a sound state
- Formulation:

$$V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{q_t} P_t(q_t|\bar{\omega}) P(\omega|q_t) *_\omega P_t(f|q_t) \qquad (4)$$

$P(\omega, t)$: 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:

$$P(\bar{\omega}) = \sum_{\bar{q}} \sum_{\bar{f}} P(q_1) \prod_t P(q_{t+1}|q_t) \prod_t P_t(\omega_t|q_t) \quad (5)$$

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

- Observation probability:

$$P_t(\omega_t|q_t) = 1 - \frac{||P(\omega, t|q_t) - V_{\omega,t}||_2}{\sum_{q_t} ||P(\omega, t|q_t) - V_{\omega,t}||_2} \quad (6)$$

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:

$$P_t(f, q_t|\bar{\omega}) = \frac{P(\omega - f|q_t)P_t(f|q_t)}{\sum_f P(\omega - f|q_t)P_t(f|q_t)} \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_t} \alpha_t(q_t)\beta_t(q_t)} \qquad (7)$$

$$P_t(q_t, q_{t+1}|\bar{\omega}) = \frac{\alpha_t(q_t)P(q_{t+1}|q_t)\beta_{t+1}(q_{t+1})P_t(\omega_{t+1}|q_{t+1})}{\sum_{q_t,q_{t+1}} \alpha_t(q_t)P(q_{t+1}|q_t)\beta_{t+1}(q_{t+1})P_t(\omega_{t+1}|q_{t+1})} \qquad (8)$$

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

HMM-constrained SIPLCA for pitch detection (6)

- Maximization step:

$$P(\omega|q) = \frac{\sum_{f,t} V_{\omega+f,t} P_t(f, q_t|\omega+f)}{\sum_{\omega,f,t} V_{\omega+f,t} P_t(f, q_t|\omega+f)} \quad (9)$$

$$P_t(f|q_t) = \frac{\sum_{\omega} V_{\omega,t} P_t(f, q_t|\omega)}{\sum_{f,\omega} V_{\omega,t} P_t(f, q_t|\omega)} \quad (10)$$

$$P(q_{t+1}|q_t) = \frac{\sum_t P_t(q_t, q_{t+1}|\bar{\omega})}{\sum_{q_{t+1}} \sum_t P_t(q_t, q_{t+1}|\bar{\omega})} \quad (11)$$

$$P(q_1) = P_1(q_1|\bar{\omega}) \quad (12)$$

HMM-constrained SIPLCA for pitch detection (7)

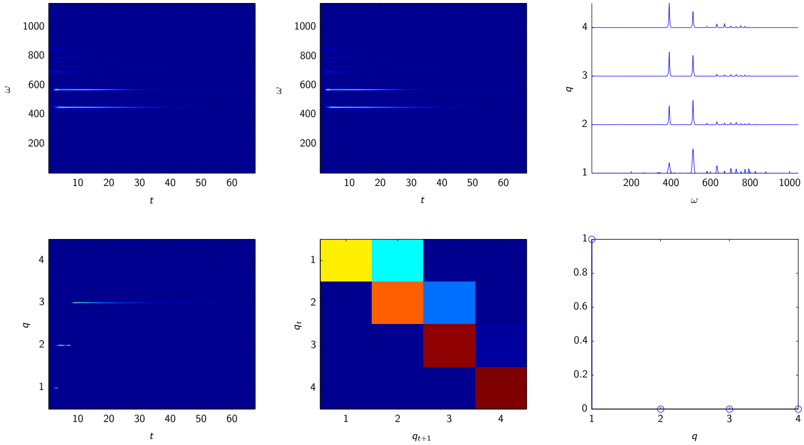


Figure: Decomposition of a piano sound (C4) using the proposed model.

centre for digital music

Queen Mary University of London

HMM-constrained SIPLCA for pitch detection (8)

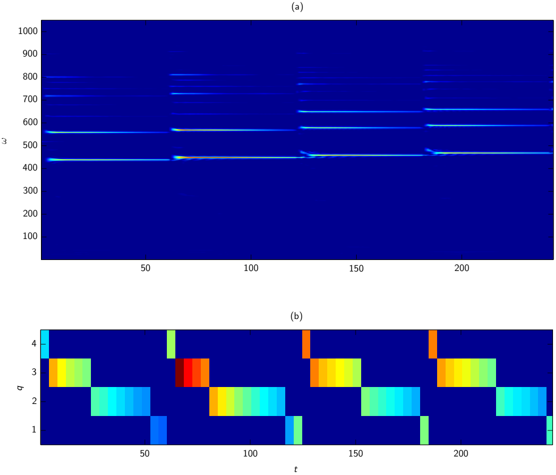
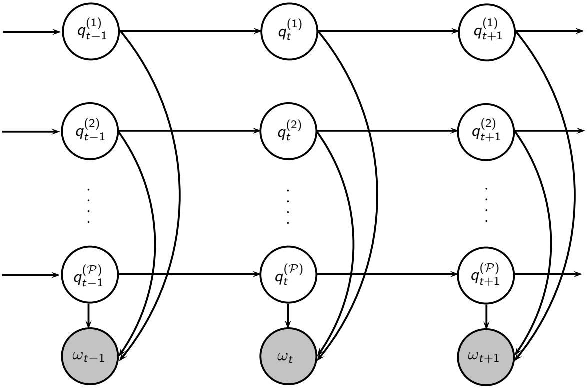


Figure: Sound state contribution for a piano melody.

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:

$$V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{s,p} P_t(p) P_t(s|p) \sum_{q_t^{(p)}} P_t(q_t^{(p)}|p, \bar{\omega}) P(\omega|s, p, q_t^{(p)}) *_\omega P_t(f|p) \tag{13}$$

- Temporal constraints through pitch-wise HMMs:

$$P(\bar{\omega}) = \sum_{\bar{q}^{(p)}} \sum_{\bar{s}} \sum_{\bar{p}} \sum_{\bar{f}} P(q_1^{(p)}) \prod_t P(q_{t+1}^{(p)}|q_t^{(p)}) \prod_t P_t(\omega_t|q_t^{(p)}) \tag{14}$$

centre for digital music

Queen Mary University of London

HMM-constrained SIPLCA for multi-pitch detection (3)

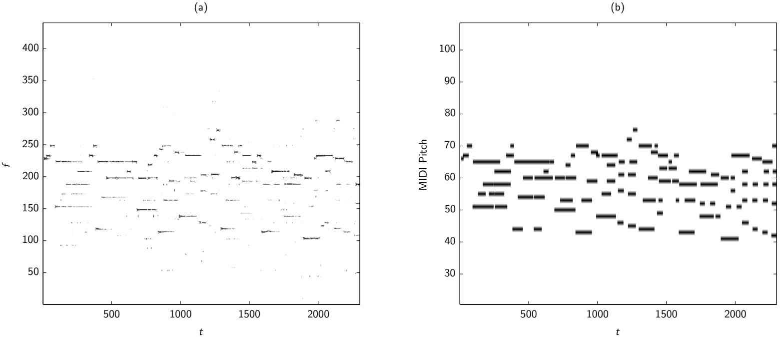


Figure: (a) Pitch spectrogram $P(f,t)$ of an excerpt of "RWC-MDB-J-2001 No. 7" (guitar). (b) The pitch ground truth of the same recording. The abscissa corresponds to 10ms. Original recording: 🔊 Synthesized transcription: 🔊

centre for digital music

Queen Mary University of London

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