Polyphonic Music Transcription Using Note Onset and Offset Detection

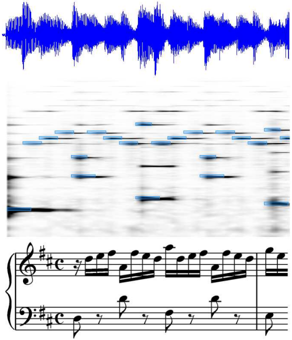
Emmanouil Benetos and Simon Dixon

{emmanouilb, simond}@eecs.qmul.ac.uk

Centre for Digital Music
Queen Mary, University of London

Introduction (1)

- Automatic music transcription (AMT): audio recording → music notation
- Applications:
    - Music information retrieval
    - Interactive music systems
    - Musicological analysis
- Subtasks:
    - Pitch estimation
    - Onset/offset detection
    - Instrument identification
    - Rhythmic parsing
- Still remains an open problem



Introduction (2)

Related Work on automatic music transcription:

- Iterative spectral subtraction-based system in Klapuri03
- Rule-based system in Zhou06, also proposed the Resonator-Time Frequency Image (RTFI)
- Joint multiple-F0 estimation in Yeh10
- Iterative estimation exploiting temporal evolution by the authors

Introduction (3)

Related Work on onset detection:

- Onset detection function combining energy and phase in Bello05
- Combining onset features using late fusion in Holzapfel10

Proposed approach:

- System for joint multiple-F0 estimation, exploiting onset and offset detection for improved multipitch estimation
- Novel onset detection features derived from transcription preprocessing steps
- Addressing offset detection using HMMs

Introduction (4)

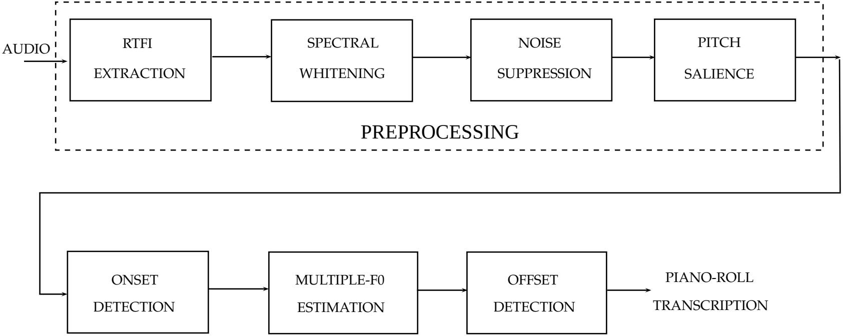
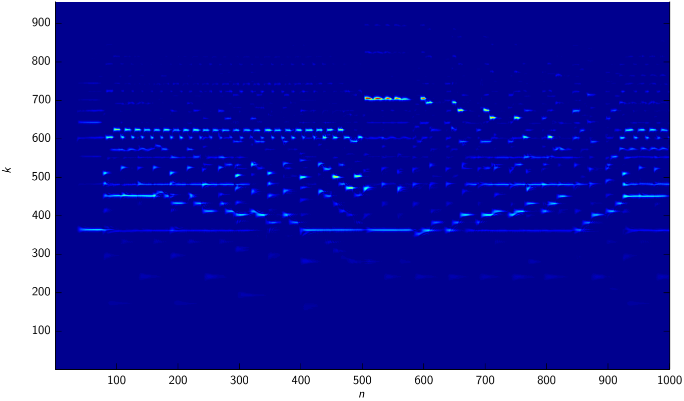


Figure: Transcription System Diagram

Preprocessing (1)

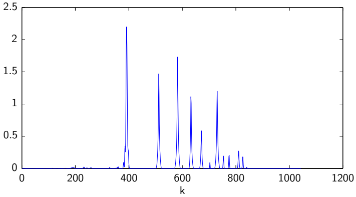
An RTFI with 120 bins/octave and 40msec frame interval is employed.

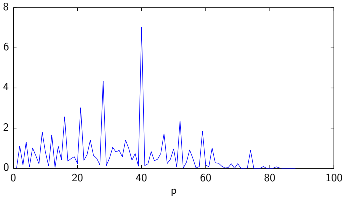
Figure: the RTFI $X[n,k]$ from the first 10sec of the MIREX multi-F0 recording



Preprocessing (2)

- Spectral whitening: suppressing timbral information. The whitening method proposed in Klapuri03 is used
- Noise suppression: a $\frac{1}{3}$ octave span median filtering procedure is employed
- Pitch salience: a pitch salience (or pitch strength) function $s[p]$, $p \in [21, \dots, 108]$ is extracted, along with tuning and inharmonicity coefficients





Onset Detection (1)

- Two proposed onset descriptors utilizing information from multiple-F0 preprocessing
- Spectral flux-based descriptor with tuning information:

$$SF[n] = \sum_{p=21}^{108} HW(\psi[p,n] - \psi[p,n-1]) \tag{1}$$

where HW is a half-wave rectifier and $\psi[p,n]$ is a semitone-resolution filterbank with tuning information.

- Onsets can be detected by peak picking on $SF[n]$.

Onset Detection (2)

- For detecting soft onsets, a pitch-based descriptor is also proposed, based on $s[p]$:

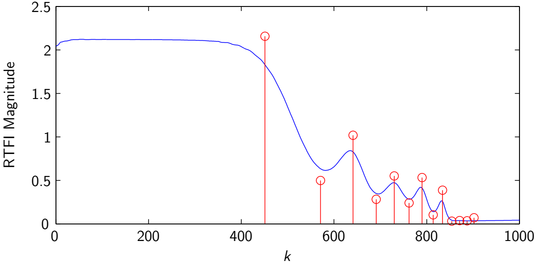
$$SD[n] = \sum_{i=1}^{12} HW(Chr[i,n] - Chr[i,n-1]) \qquad (2)$$

  where $Chr[i,n]$ is a chroma-warped and smooth version of the pitch salience function.

- Late fusion is applied in order to combine the 2 onset descriptors
- Development set from Ghent University for tuning onset detection parameters

Multiple-F0 Estimation (1)

- For each frame, a pitch candidate set **C** is selected, and overlapping partial treatment is applied for each subset $C \subseteq \mathbf{C}$
- A partial collision list is computed
- Amplitudes of overlapped partials estimated by discrete cepstrum-based spectral envelope estimation



Multiple-F0 Estimation (2)

- Score function for selecting the optimal pitch candidate set $C \subseteq \mathbf{C}$:

$$\begin{aligned}
\mathcal{L}(C) &= \sum_{i=1}^{|C|} (\mathcal{L}_{p(i)}) + \mathcal{L}_{res} \\
\mathcal{L}_p &= w_1 Fl[p] + w_2 Sm[p] - w_3 SC[p] + w_4 PR[p] \\
\mathcal{L}_{res} &= w_5 Fl[Res] \qquad (3)
\end{aligned}$$

Fl: spectral flatness of the harmonic partial sequence
Sm: smoothness measure
SC: spectral centroid
PR: harmonically-related pitch ratio

Res: residual spectrum

centre for digital music

Queen Mary University of London

Multiple-F0 Estimation (3)

- Optimal pitch set:

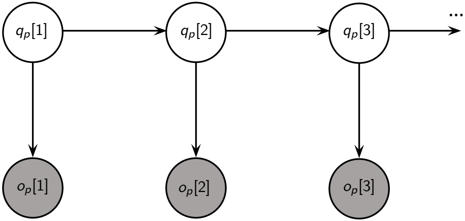
$$\hat{C} = \arg\max_{C \subseteq \mathbf{C}} \mathcal{L}(C) \tag{4}$$

- Weight parameters $w_i, i = 1, \ldots, 5$ trained using the Nelder-Mead search algorithm
- Training set for weight parameters consists of 100 piano chords from the MAPS database

Offset Detection

- Proposed offset detection using 2-state HMMs for each pitch p
- State priors $P(q_p[1])$ and transitions $P(q_p[n]|q_p[n-1])$ computed from MIDI files from the RWC database
- Observation probability for an active pitch from pitch salience:

$$P(o_p[n]|q_p[n]=1) = \frac{1}{1+e^{-(s[p,n]-1)}} \tag{5}$$



Evaluation (1)

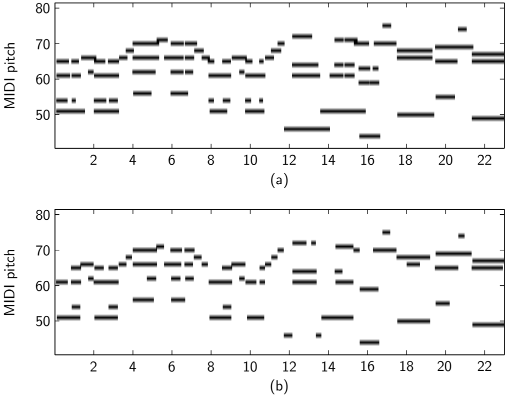
- Test set: Twelve 23sec excerpts from the RWC database (classic and jazz music)
- Aligned MIDI ground truth created using Sonic Visualizer

	RWC ID	Instruments
1	RWC-MDB-J-2001 No. 1	Piano
2	RWC-MDB-J-2001 No. 2	Piano
3	RWC-MDB-J-2001 No. 6	Guitar
4	RWC-MDB-J-2001 No. 7	Guitar
5	RWC-MDB-J-2001 No. 8	Guitar
6	RWC-MDB-J-2001 No. 9	Guitar
7	RWC-MDB-C-2001 No. 30	Piano
8	RWC-MDB-C-2001 No. 35	Piano
9	RWC-MDB-J-2001 No. 12	Flute + Piano
10	RWC-MDB-C-2001 No. 12	Flute + String Quartet
11	RWC-MDB-C-2001 No. 42	Cello + Piano
12	RWC-MDB-C-2001 No. 49	Tenor + Piano

Table: The RWC data used for transcription experiments.

Evaluation (2)

Figure: (a) The pitch ground-truth of an excerpt from ‘RWC MDB-J-2001 No. 9’ (guitar) 🔊 (b) The transcription output of the same recording 🔊



Evaluation (3)

	Frame-based	Onsets only	Onsets+offsets	Cañadas10	Saito08	Kameoka07
Mean	60.5%	59.7%	**61.2%**	59.1%	56.2%	59.6%
Std.	11.5%	11.5%	11.2%	11.5%	12.9%	16.9%

Table: Transcription results (*Acc*) for the 12 RWC recordings.

Method	Acc	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Onsets only	59.7%	40.3%	8.4%	24.6%	7.3%
Onsets+offsets	61.2%	38.8%	7.3%	24.8%	6.7%

Table: Transcription error metrics.

Features	Pre	Rec	F
$SF + SD$	52.85%	86.84%	63.17%
SF	66.29%	81.69%	70.56%
SD	55.36%	82.42%	63.80%

Table: Onset detection results.

centre for digital music

Queen Mary
University of London

Evaluation (3)

	Frame-based	Onsets only	Onsets+offsets	Cañadas10	Saito08	Kameoka07
Mean	60.5%	59.7%	**61.2%**	59.1%	56.2%	59.6%
Std.	11.5%	11.5%	11.2%	11.5%	12.9%	16.9%

Table: Transcription results (*Acc*) for the 12 RWC recordings.

Method	*Acc*	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Onsets only	59.7%	40.3%	8.4%	24.6%	7.3%
Onsets+offsets	61.2%	38.8%	7.3%	24.8%	6.7%

Table: Transcription error metrics.

Features	*Pre*	*Rec*	*F*
$SF + SD$	52.85%	86.84%	63.17%
SF	66.29%	81.69%	70.56%
SD	55.36%	82.42%	63.80%

Table: Onset detection results.

Evaluation (3)

	Frame-based	Onsets only	Onsets+offsets	Cañadas10	Saito08	Kameoka07
Mean	60.5%	59.7%	**61.2%**	59.1%	56.2%	59.6%
Std.	11.5%	11.5%	11.2%	11.5%	12.9%	16.9%

Table: Transcription results (*Acc*) for the 12 RWC recordings.

Method	*Acc*	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Onsets only	59.7%	40.3%	8.4%	24.6%	7.3%
Onsets+offsets	61.2%	38.8%	7.3%	24.8%	6.7%

Table: Transcription error metrics.

Features	*Pre*	*Rec*	*F*
$SF + SD$	52.85%	86.84%	63.17%
SF	66.29%	81.69%	70.56%
SD	55.36%	82.42%	63.80%

Table: Onset detection results.

Conclusions

Contributions:

- Onset detection features derived from multiple-F0 preprocessing
- Score function combining several features for multiple-F0 estimation
- Offset detection using HMMs
- Transcription results on RWC excerpts outperform state-of-the-art

Future work:

- Explicitly modelling sound states (attack, transient, sustain, release)
- Joint multiple-F0 estimation and note tracking
- Public evaluation through MIREX framework