Automatic Music Transcription

Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London

http://www.eecs.qmul.ac.uk/~emmanouilb/

January 2019

Queen Mary University of London

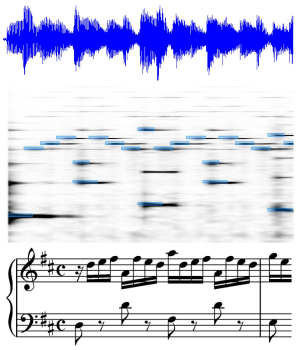
1

Tutorial Material

Tutorial Website:
http://c4dm.eecs.qmul.ac.uk/nus-amt-tutorial/

Tutorial sources:

- Paper: “Automatic music transcription: challenges and future directions”, by E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, 2013.
- ISMIR 2015 tutorial, “Automatic music transcription”, by Z. Duan and E. Benetos.
- Tutorial at UFRGS, Brazil: "Music information retrieval and automatic music transcription“, by E. Benetos and R. Schramm, 2017.
- Paper: “Automatic music transcription: an overview” by E. Benetos, S. Dixon, Z. Duan, and S. Ewert, 2019.

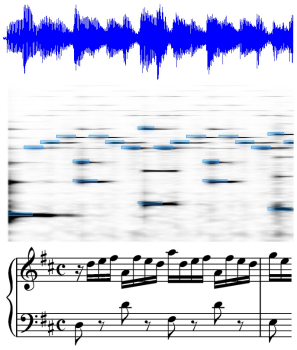


2

2

Tutorial Outline

1. Introduction
2. How do humans transcribe music?
3. State-of-the-art research on AMT
4. Datasets and evaluation measures
5. Relations and applications to other problems
6. Software & Demo
7. Further extensions and advanced topics
8. Conclusions + Q&A



3

3

Introduction

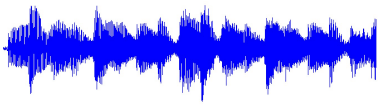
4

4

AMT - Introduction (1)

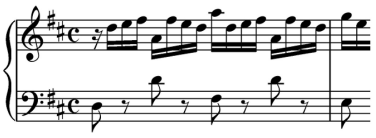
Automatic music transcription (AMT): the process of converting an acoustic musical signal into some form of music notation (e.g. staff notation, MIDI file, piano-roll,…)

Music audio



Mid-level & Parametric representation
- Pitch, onset, offset, stream, loudness
- Uses audio time (ms)

Music notation
- Note name, key, rhythm, instrument
- Uses score time (beat)



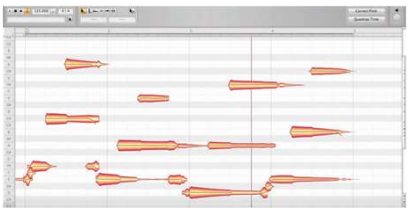
5

5

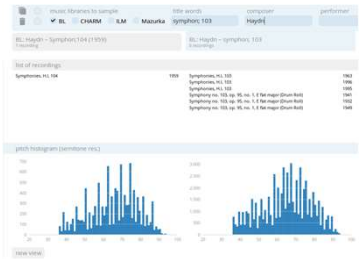
AMT - Introduction (2)

Fundamental (and open) problem in music information research

Applications:
- Search/annotation of musical information
- Interactive music systems
- Music education
- Music production
- Digital/computational musicology



http://www.celemony.com/en/melodyne/



http://dml.city.ac.uk/vis/

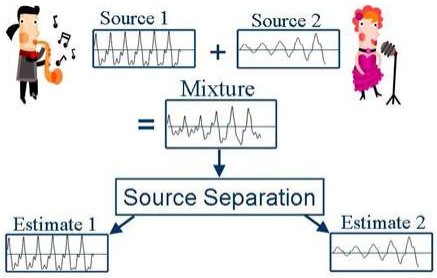
6

6

AMT - Introduction (3)

Relations to other music information research tasks:

- Audio source separation
- Score following
- Structural segmentation
- Music similarity
- Cover song detection
- …



Provides a link between music signal processing and symbolic music processing

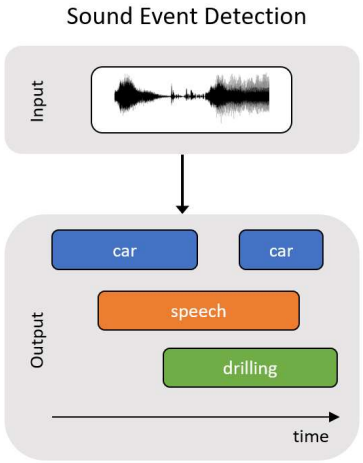
7

7

AMT - Introduction (4)

Relations to other research fields/tasks:

- Automatic speech recognition
- Sound event detection
- Computer vision
- Natural language processing

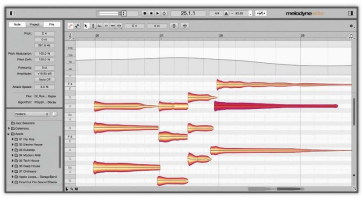


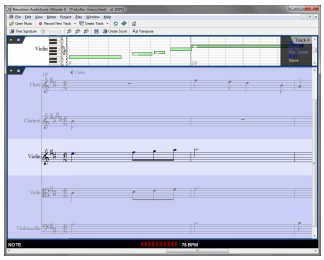
8

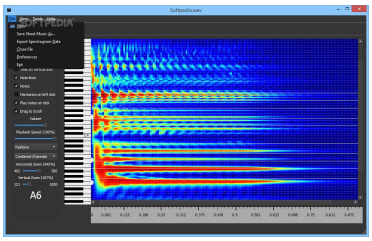
8

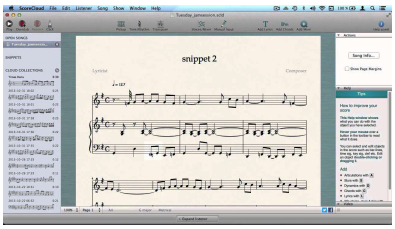
AMT - Introduction (5)

AMT is not just about academic research!









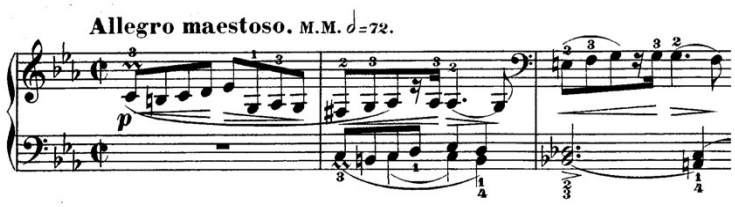
9

9

AMT - Introduction (6)

Subtasks:

- Pitch detection
- Onset/offset detection
- Instrument identification
- Rhythm parsing
- Identification of dynamics/expression
- Typesetting

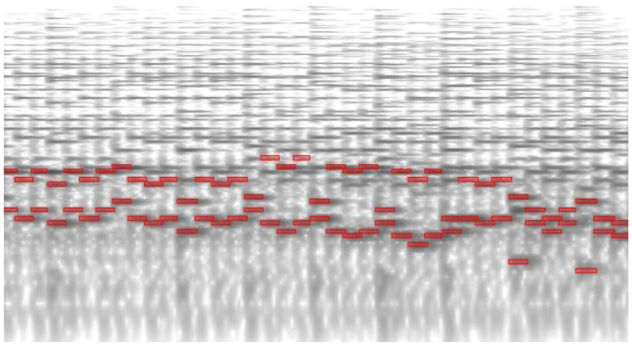


10

10

AMT - Introduction (7)

Core problem: multi-pitch detection



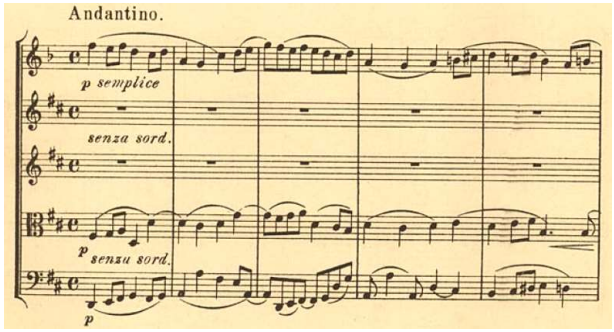
11

AMT - Introduction (8)

How difficult is it?

- Let's listen to a piece and try to transcribe (hum) the different tracks

J. Brahms, Clarinet Quintet in B minor, op.115. 3rd movement



12

AMT - Introduction (9)

We humans are amazing!

- "In Rome, he (14 years old) heard Gregorio Allegri's *Miserere* **once** in performance in the Sistine Chapel. He wrote it out **entirely from memory**, only returning to correct **minor errors**..."

  -- Gutman, Robert (2000). *Mozart: A Cultural Biography*

Wolfgang Amadeus Mozart

- Can we make computers compete with Mozart?

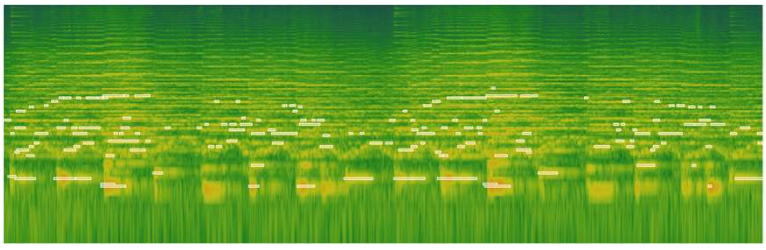
13

13

AMT - Introduction (10)

Challenges:

- Contrary to speech recognition, computer vision etc, musical attributes are not independent!
- Inferring musical attributes from a mixture is an extremely underdetermined problem
- Data annotation is extremely time-consuming, leading research towards specific sub-problems, e.g. piano transcription



Automatic transcription of B. Smetana – Má vlast (Vltava)

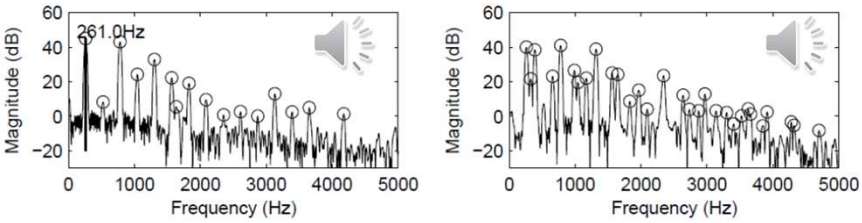
14

14

AMT - Introduction (11)

Challenges (continued):

- Concurrent sound sources interfere with each other
  - Overlapping harmonics: C4 (46.7%), E4 (33.3%), G4 (60%)



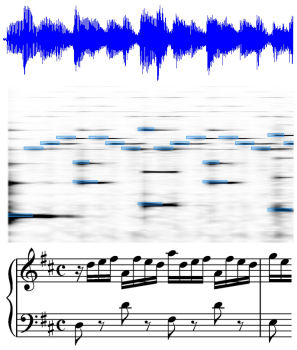
- Large variety of music
  - Music pieces: style, form, etc.
  - Instrumentation: bowed/plucked strings, winds, brass, percussive, etc.
  - Playing techniques: legato, staccato, vibrato, etc.

15

15

Tutorial Focus/Objectives

- Focusing (mostly) on polyphonic music transcription
- Presenting an overview of representative AMT research (+ related problems)
- Overview of current trends and topics
- Discussion on limitations, challenges, and future directions
- Tutorial website:
  http://c4dm.eecs.qmul.ac.uk/nus-amt-tutorial/

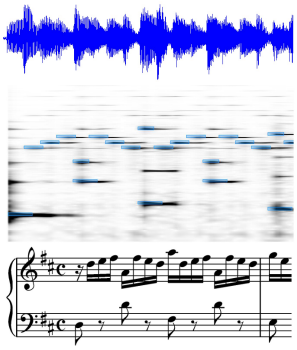


16

16

Tutorial Outline

1. Introduction
2. How do humans transcribe music?
3. State-of-the-art research on AMT
4. Datasets and evaluation measures
5. Relations and applications to other problems
6. Software & Demo
7. Further extensions and advanced topics
8. Conclusions + Q&A



17

17

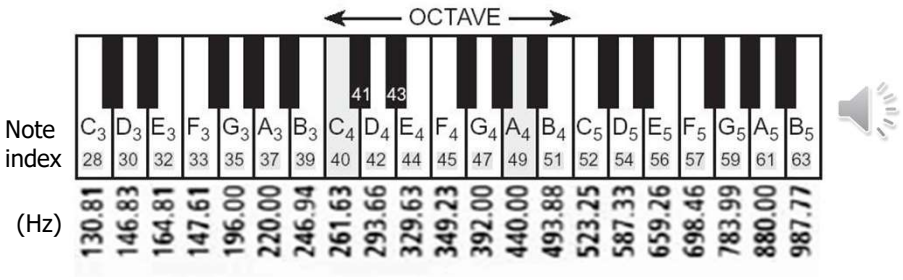
How do humans transcribe music?

18

18

Pitch Perception (1)

Pitch:
- That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high (ANSI)
- (Operational) A sound has a certain pitch if it can be reliably matched to a sine tone of a given frequency at 40 dB SPL
- People hear pitch in a logarithmic scale



19

19

Pitch Perception (2)

Fundamental frequency (F0): is defined as the reciprocal of the period of a periodic signal.

Properties of pitch perception [de Cheveigné, 2006; Houtsma, 1995]:
- Range: Pitch may be salient as long as the F0 is within about 30Hz-5kHz
- Missing fundamental: the fundamental frequency need not be present in for a pitch to be perceived
- Harmonics: For a sound with harmonic partials to be heard as a musical tone, its spectrum must include at least 3 successive harmonics of a common frequency

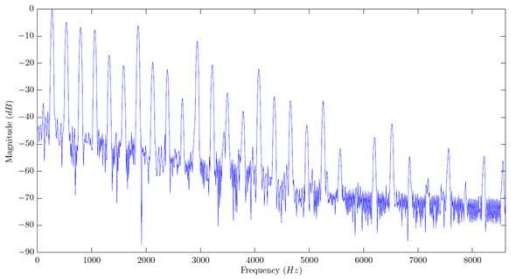


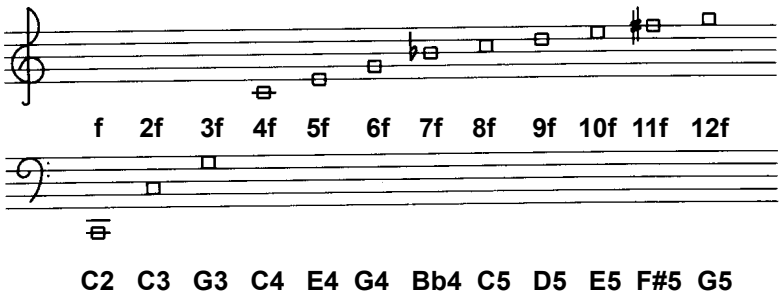
Figure: spectrum of a C4 piano note. The fundamental is located at 261.6Hz.

20

20

Pitch Perception (3)

- Harmonics make tones more pleasant, but may confuse pitch perception, especially in polyphonic settings (octave/harmonic errors)



21

21

Pitch Perception (4)

Relative pitch: Ability to recognize and reproduce frequency ratios
Absolute pitch: Identifying pitch on an absolute nominal scale without explicit external reference

Pitch perception theories have informed the creation of AMT systems.

Modern theories:

- Pattern matching [de Boer, 1956; Wightman, 1973; Terhardt, 1974]
- Autocorrelation model [Licklider, 1951; Meddis & Hewitt, 1991; de Cheveigné, 1998]

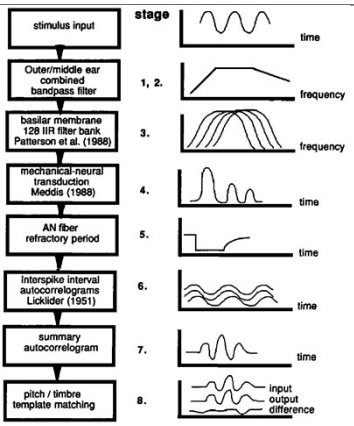


Figure from Meddis & Hewitt, 1991

22

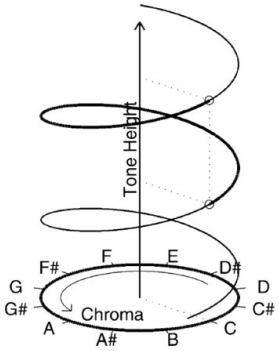
22

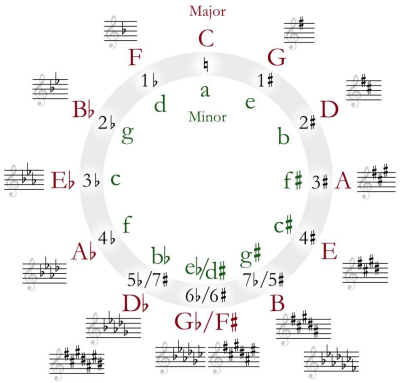
Pitch Perception (5)

Pitch is not a one-dimensional entity! (low/high)

Multidimensional aspects of pitch:

- Octave similarity – helix representation [Revesz, 1954]
- Pitch distance – circle of fifths representation [Shepard, 1982]



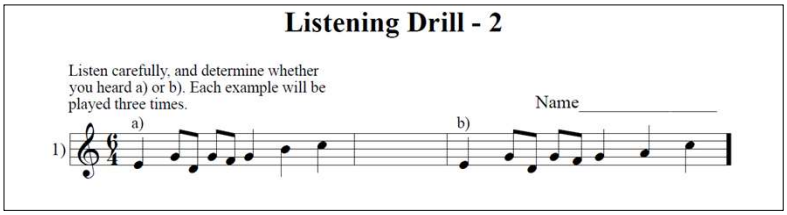


23

Human Transcription (1)

- Called **musical dictation** in ear training pedagogy
- **Definition**: a skill by which musicians learn to identify, solely by hearing, pitches, intervals, melody, chords, rhythms, and other elements of music.
- Required in all college-level music curriculums; general expectation after 4-5 semesters' training:

> "they can transcribe an excerpt of a quartet (e.g. four measures) with quite complex harmonies, after listening to it four or five times"
>
> ---- Temperley, 2013



source: http://www.sheetmusic1.com/ear.training.html

24

24

Human Transcription (2)

- For accurate transcription, a great deal of practice is often necessary!
- How trained musicians transcribe music [Hainsworth03]:
  - Some use a transcription aid: musical instrument, tape recorder, software
  - Faithful transcription vs. reduction/arrangement
  - Implicitly: style detection, instrument identification, beat tracking
  - Process:
    1. Rough sketch of the piece
    2. Chord scheme / bass line
    3. Melody + counter-melodies

25

25

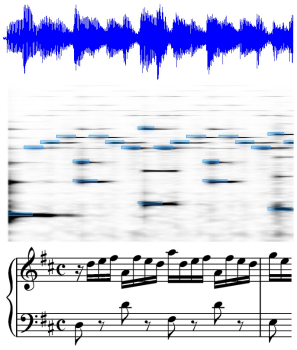
State-of-the-art research in AMT

26

26

State-of-the-art Outline

1. Frame-level transcription
   A. Time & Frequency domain methods
   B. Spectrogram decomposition methods
   C. Classification-based methods
2. Note-level transcription
3. Stream-level transcription

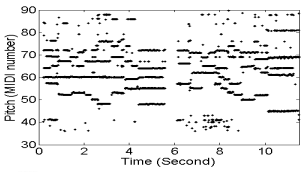


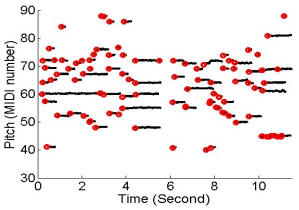
27

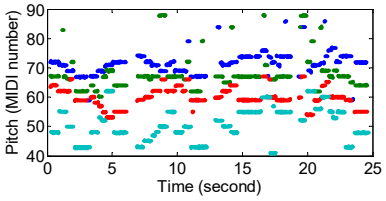
27

State of the Art in Multi-pitch Analysis

- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods
- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods
- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods







28

28

How difficult is it?

- Let's do a test!
  - Q1: How many pitches are there?
  - Q2: What are their pitches?
  - Q3: Can you find a pitch in Chord 1 and a pitch in Chord 2 that are played by the same instrument?

Chord 1	Chord 2
2	3
C4/G4	C4/F4/A4
Clarinet G4 Horn C4	Clarinet A4 Viola F4 Horn C4

29

29

Frame-level: Multi-pitch Estimation

Categorization of methods

- Domain of operation: time, frequency, hybrid
- Representation:
  - Time domain: raw waveform, auditory filterbank
  - Frequency domain: STFT spectrum, CQT spectrum, ERB filterbank, specmurt, spectral peaks
- Core algorithm: rule-based, signal processing approaches, maximum likelihood, Bayesian, spectrogram decomposition, sparse coding, classification-based, etc.
- Iterative vs. joint estimation of pitches

30

30

Time Domain Methods

- Key idea
  - Harmonic sounds are periodic
  - Use autocorrelation function (ACF) to find signal period
- Difficulty
  - Tend to have subharmonic errors
  - Periodicity is unclear when multiple harmonic sounds are mixed

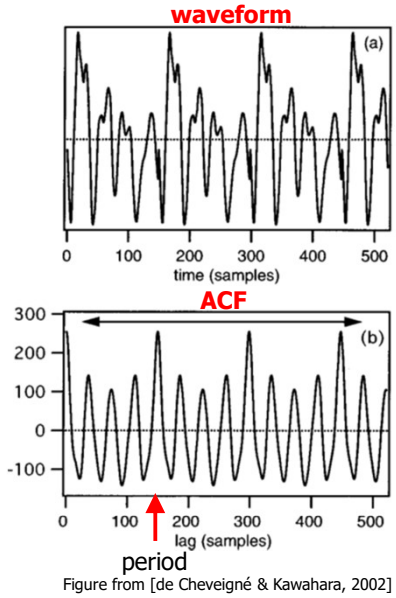


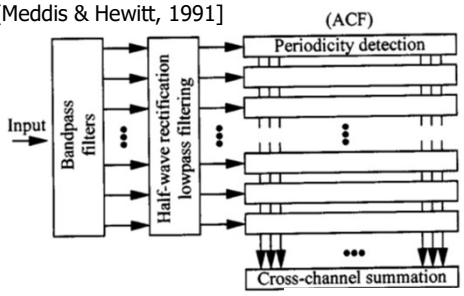
Figure from [de Cheveigné & Kawahara, 2002]

31

Time Domain - Autocorrelation

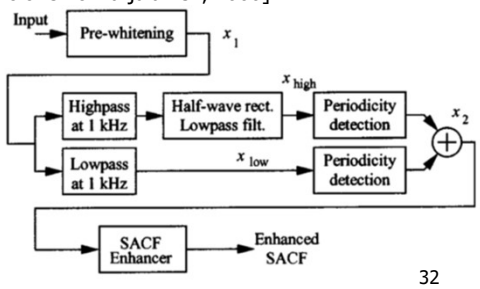
- Detailed simulation of human auditory system
  - Outer- and middle-ear freq. attenuation effect
  - ~100 channels with critical bandwidth
  - Inner hair cell response
- Simplified version
  - Only 2 channels
  - Enhanced SACF: remove SACF peaks due to integer multiples of periods

[Meddis & Hewitt, 1991]



Summary ACF (SACF)

[Tolonen & Karjalainen, 2000]



Figures from [Tolonen & Karjalainen, 2000]

32

Time Domain – Probabilistic Modeling

- Harmonic model [Walmsley et al., 1999]

#notes #harmonics Harmonic amplitude and phase Gaussian noise (i.i.d.)

$$y_t = \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M_k} \alpha_m \cos(m\omega_{0,k} t) + \beta_m \sin(m\omega_{0,k} t) \right\} + v_t$$

F0

  - Parameters: $K, \{M_k\}, \{\alpha_m\}, \{\beta_m\}, \{\omega_{0,k}\}$, variance of v_t
  - Impose priors on parameters
  - Bayesian inference by Markov Chain Monte Carlo (MCMC)

  Pros: rigorous mathematical mode

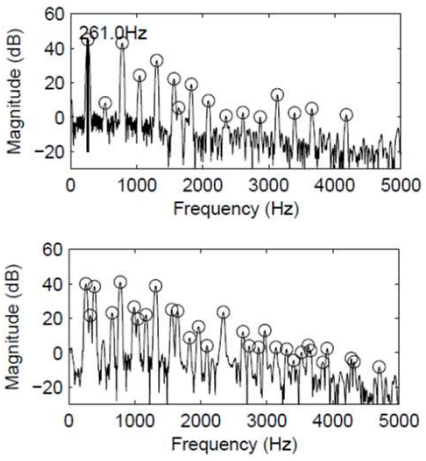
  Cons: computationally expensive; purely harmonic model

33

33

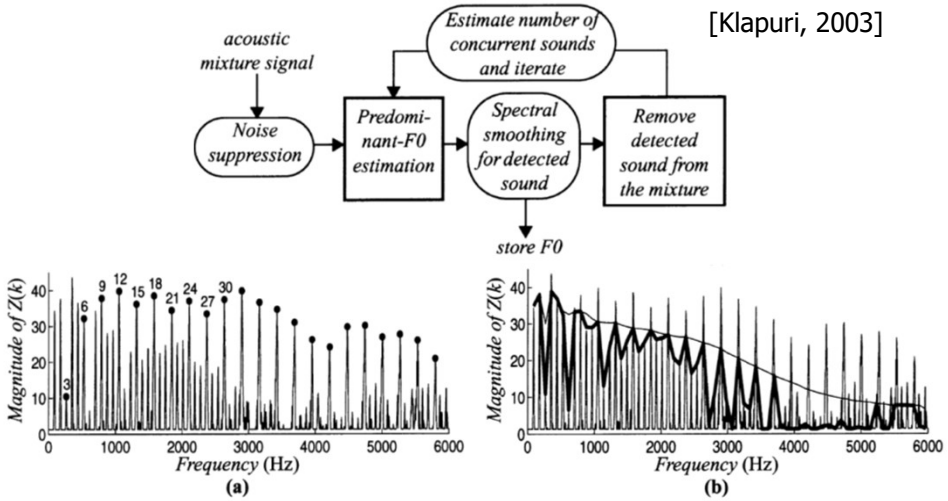
Frequency Domain Methods

- Key idea
  - Each pitch has a set of harmonics
  - Recognize the harmonic patterns
- Difficulty
  - Tend to have harmonic errors
  - Harmonic amplitude varies
  - Overlapping harmonics



34

Iterative Spectral Subtraction



[Klapuri, 2003]

Pros: good performance, simple, fast
Cons: hard to subtract the appropriate amount of energy

35

35

Spectral Peak Modeling

- Peak picking
- Choose pitch candidates
  - Around first several peaks and their integer fractions
- Calculate salience (or likelihood) of *each pitch* or each *combination of pitches*
- Choose the best ones

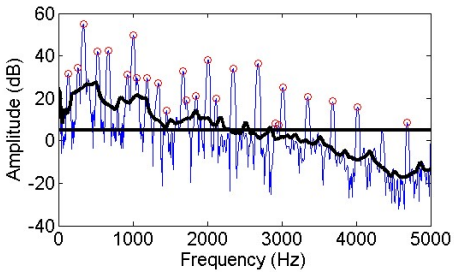


Figure from [Duan et al., 2010]

- Pros: intuitive; works well; more compact representation of audio
- Cons: sensitive to peak detection; has difficulty in dealing with sources with different loudness

36

36

Spectral Peak Modeling – Maximum Likelihood

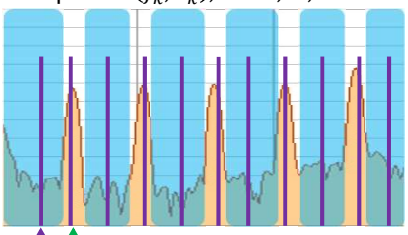
- [Duan et al., 2010]

Pros: balances harmonic and subharmonic errors
Cons: soft notes may be masked by others

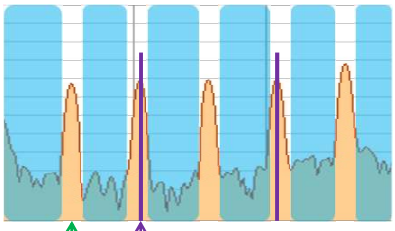
$$p(\boldsymbol{O}|\boldsymbol{\theta}) = p\left(\boldsymbol{O}^{\text{peak}}|\boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{O}^{\text{non-peak}}|\boldsymbol{\theta}\right)$$

Probability of observing these peaks: $(f_k, a_k), k = 1, \ldots, K.$

Probability of not having any harmonics in the non-peak region



$p\left(\boldsymbol{O}^{\text{peak}}|\boldsymbol{\theta}\right)$ is large
$p\left(\boldsymbol{O}^{\text{non-peak}}|\boldsymbol{\theta}\right)$ is small



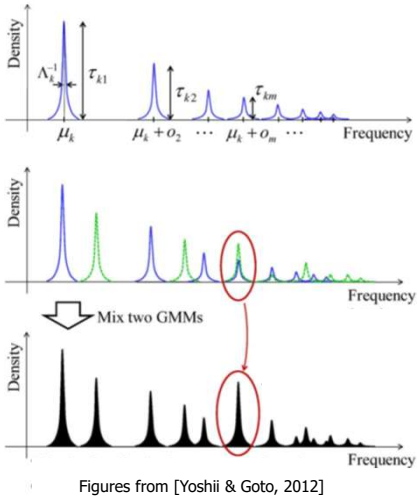
$p\left(\boldsymbol{O}^{\text{peak}}|\boldsymbol{\theta}\right)$ is small
$p\left(\boldsymbol{O}^{\text{non-peak}}|\boldsymbol{\theta}\right)$ is large

37

37

Full Spectrum Modeling – Probabilistic

- Key idea: view spectra as (parametric) probabilistic distributions
- Each note = tied- Gaussian Mixture Model (tied-GMM)
- Signal = Mixture of GMMs



Pros: flexible to incorporate priors on parameters
Cons: doesn't model inharmonicity and transients; many parameters to optimize

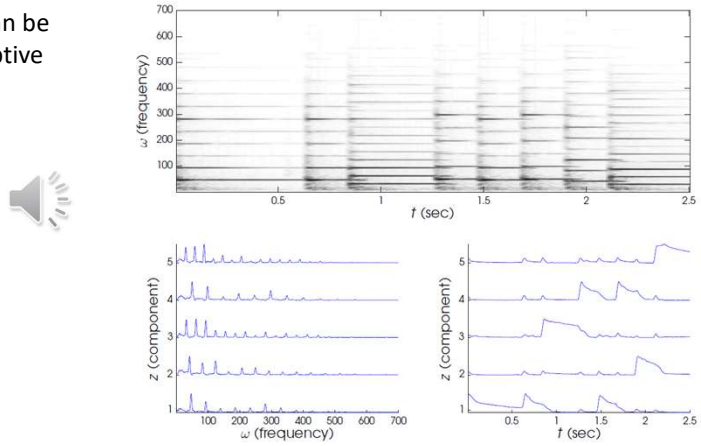
Figures from [Yoshii & Goto, 2012]

38

38

Spectrogram Decomposition (1)

- **Non-negative Matrix Factorization (NMF)** applied to magnitude spectrograms [Smaragdis03]
- Related methods: **Probabilistic Latent Component Analysis (PLCA)**, **sparse coding**
- Dictionary can be fixed or adaptive



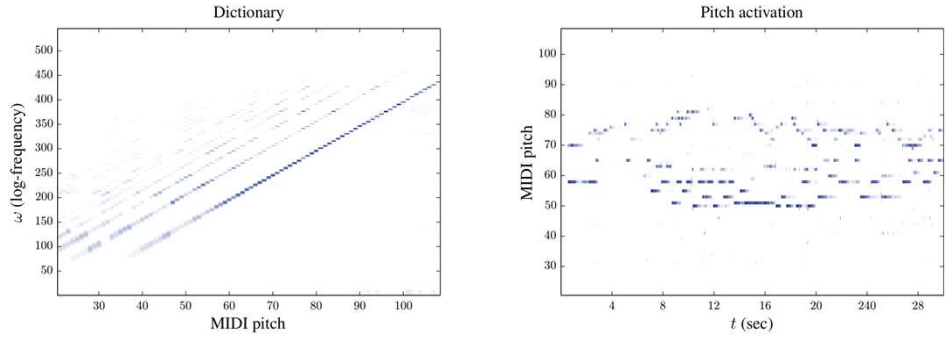
39

Spectrogram Decomposition (2)

NMF model: Given a non-negative matrix *V* find non-negative matrix factors *W* and *H* such that: $V \approx WH$

AMT Models with Fixed Templates

- *W*: note dictionary; *H*: pitch activation
- Keep *W* fixed, only estimate *H* (e.g. [Dessein10; Ari12])

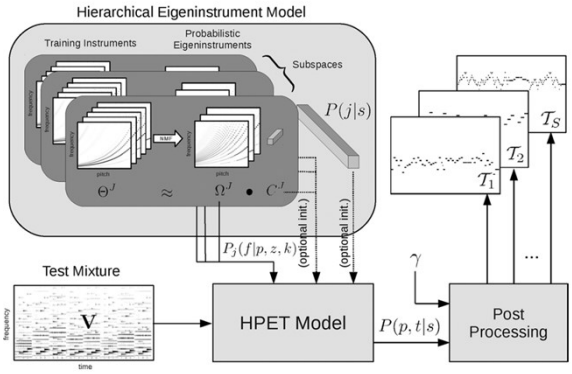


40

Spectrogram Decomposition (3)

Fixed Templates (continued)

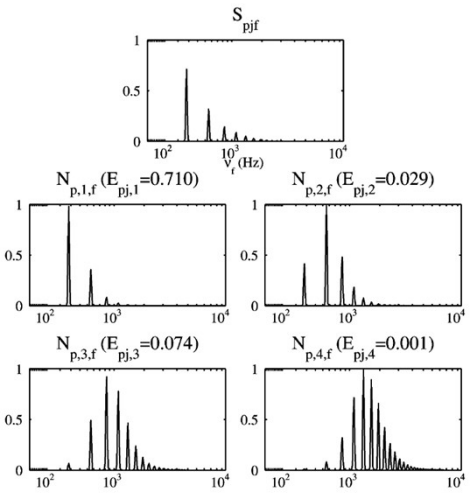
- PLCA + eigeninstruments [Grindlay11]
- PLCA + sparsity/continuity priors [Bay12]
- Pros: dictionary incorporates prior knowledge on instrument model + acoustics, good performance in a source-dependent scenario
- Cons: models perform poorly if test audio doesn't match the dictionary



Spectrogram Decomposition (4)

Adaptive templates

- Bayesian NMF + harmonicity/smoothness [Bertin10]
- NMF with adaptive harmonic decomposition [Vincent10]
- PLCA with template adaptation [Benetos14]
- Pros: dictionary closely matches test audio, potentially improving AMT performance
- Cons: strong assumptions (e.g. strictly harmonic spectra, lack of transient components, relying on a good initial estimate…)



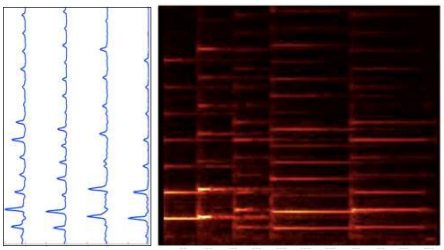
Spectrogram Decomposition (5)

[Smaragdis & Raj, 2006]

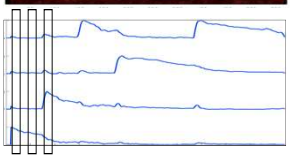
- Probabilistic Latent Component Analysis (PLCA)

Sound quanta distribution at t → $P_t(f) \approx \sum_z P(f|z)P_t(z)$

Dictionary Elements $P(f|z)$ →



Time-invariant sound quanta distribution for each component



← Activation weights $P_t(z)$

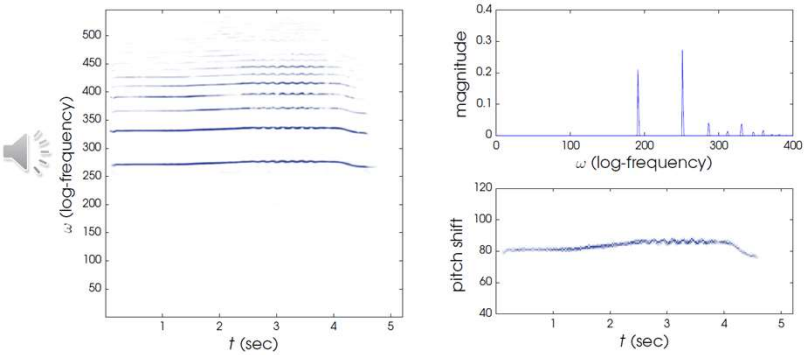
Distribution of components

43

Spectrogram Decomposition (6)

Convolutive models (NMD, Shift-Invariant PLCA)

- SIPLCA – fixed templates [Benetos12]
- SIPLCA – adaptive templates [Fuentes13]
- Pros: can model tuning changes & frequency modulations
- Cons: computationally expensive; no improvement over linear models in some cases (e.g. tuned piano)



44

Classification-based Methods

- Basic idea
  - View polyphonic music transcription as multi-label classification
  - Each quantized pitch (e.g., MIDI number) is a class
  - Positive/negative examples: frames contain/not contain the pitch
- Pros:
  - Simple idea
  - Requires no acoustical prior knowledge
- Cons:
  - Only outputs quantized pitch
  - Requires lots of training data given the many class combinations
  - May overfit training data; hard to adapt to different datasets/instruments

45

45

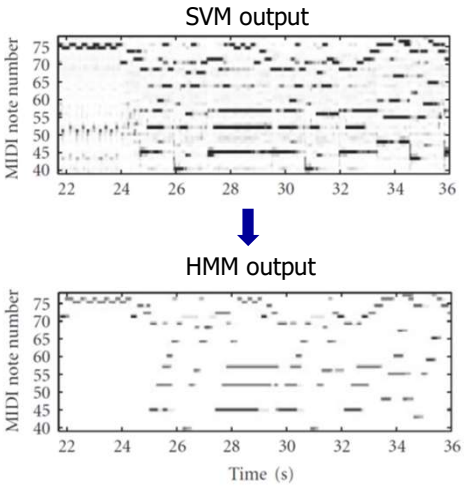
Classification-based Methods (1)

[Poliner & Ellis, 2007]

- 87 independent one-vs-all SVMs for piano (except for the highest note C8)
- Trained on MIDI-synthesized piano performances
- Features: magnitude spectrum within

$$\begin{cases} 0\text{–}2 \text{ kHz, for notes} \le \text{B5 (988Hz)} \\ 1\text{–}3 \text{ kHz, for C6} \le \text{notes} \le \text{B6} \\ 2\text{–}4 \text{ kHz, for notes} \ge \text{C7 (2093Hz)} \end{cases}$$

- HMM smoothing for each class independently



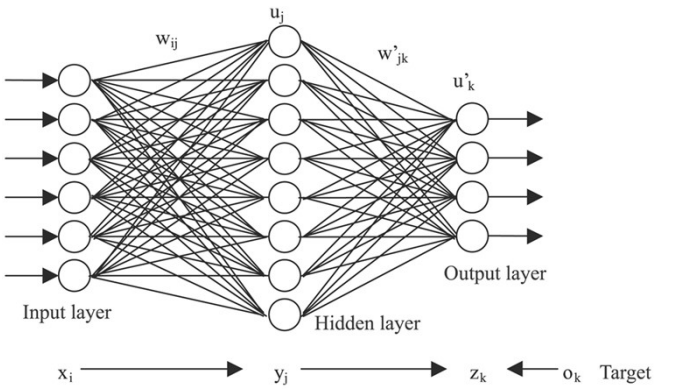
46

46

Classification-based Methods (2)

Neural Networks (NNs)

- Systems vauely inspired by biological neural networks
- Based on a collection of connected units or nodes called artificial neurons
- NNs are able to learn a nonlinear function from input to output via an optimization algorithm



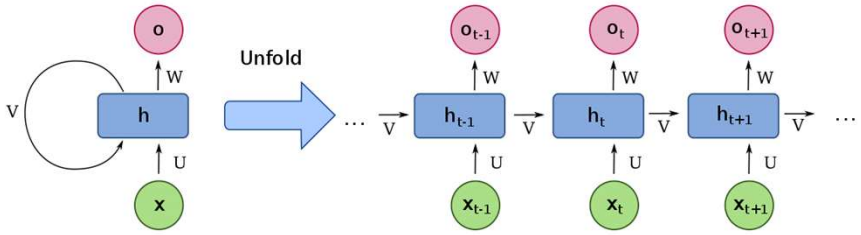
47

47

Classification-based Methods (3)

Recurrent Neural Networks (RNNs)

- Connections between nodes form a directed graph along a sequence
- Used to model sequential/timeseries data
- Variants: gated recurrent units (GRUs), long short-term memory networks (LSTMs)…



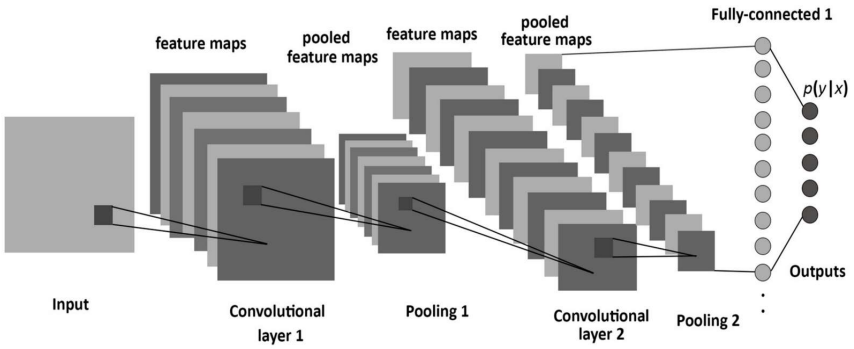
48

48

Classification-based Methods (4)

Convolutional Neural Networks (CNNs)

- Also known as **shift invariant neural networks**
- Convolutional layers apply a convolution operation to the input, passing the result to the next layer



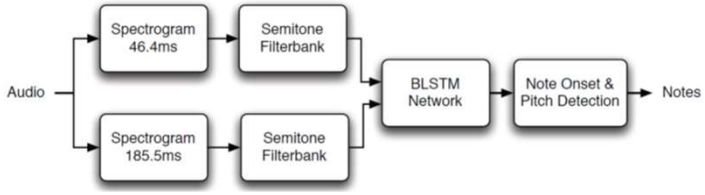
49

49

Classification-based Methods (5)

[Böck & Schedl, 2012] for piano transcription

- Bidirectional long short-term memory (BLSTM) network
  - Input layer: spectrum and its first-order time difference
  - 3 bidirectional hidden layers, 88 LSTM units each
  - 88 units in the regression output layer
  - Thresholding and pick picking for onset detection



- Pros: output notes jointly
- Cons: method does not compute note durations

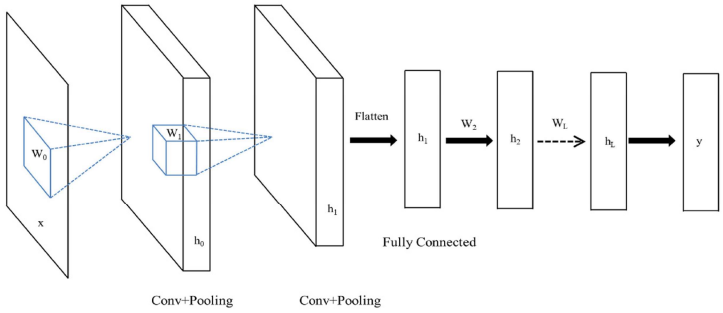
50

50

Classification-based Methods (6)

[Sigtia et al, 2016; Keltz et al, 2016] for piano transcription

- Comparison between a DNN, RNN and CNN for piano transcription
- Uses the constant-Q transform as input representation



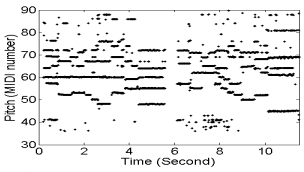
- Pros: CNNs report state-of-the-art performance
- Cons: low temporal resolution

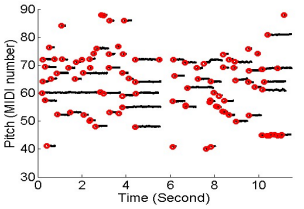
51

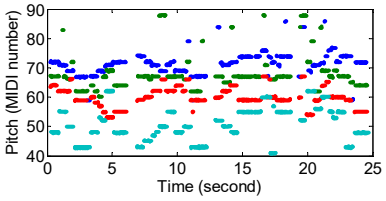
51

State of the Art recap

- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods
- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods
- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods







52

52

Note Tracking

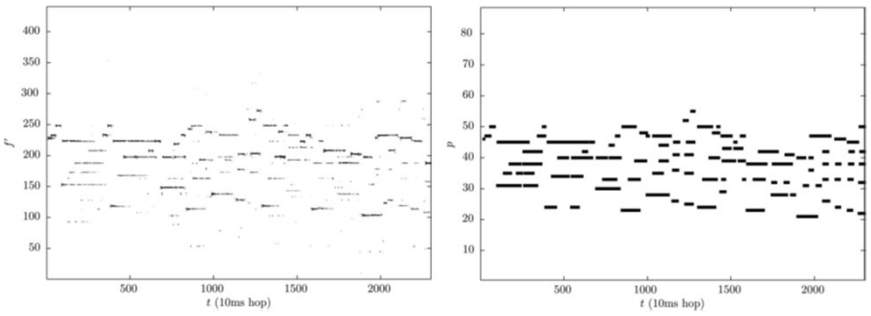
- Onset detection followed by multi-pitch estimation between onsets
  - [Marolt, 2004; Emiya et al., 2010; Grosche et al., 2012; O'Hanlon et al., 2012; Cogliati & Duan, 2015a]
  - Can be sensitive to onset detection accuracy
- As post-processing of frame-level pitch estimates
  - Form notes independently by connecting nearby pitches
  - Ignores interactions between simultaneous pitches
  - Consider interactions between simultaneous pitches
- Directly from audio

53

53

Frame Level → Note Level (1)

- Based on pitch salience/likelihood/activations
  - Thresholding, filling, pruning: [Bertin et al., 2010; Dessein et al., 2010; Grindlay & Ellis, 2011; Böck & Schedl, 2012;]
  - Median filtering: [Su & Yang, 2015]
  - Pitch-wise on/off HMMs - [Poliner & Ellis, 2007; Nam et al., 2011; Benetos & Dixon, 2013]

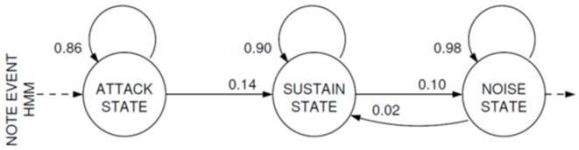


54

54

Frame Level → Note Level (2)

- Based on pitch salience/likelihood/activations
  - HMM smoothing: [Ryynanen & Klapuri, 2005]
  - Model each note with a note event HMM (3 states)
  - Observation: pitch deviation, pitch salience, onset strength



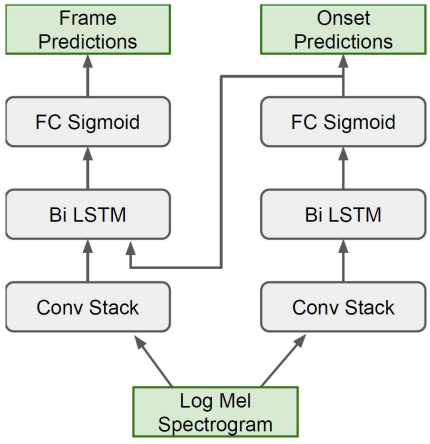
  - Model silence with a silence HMM (1 state)
  - Model transition between notes←→notes and notes←→silence with a musicological HMM
    - Note transition is key-dependent
    - Note sequence: starts with silence→note and ends with note→silence
    - Greedy iterative algorithm to find multiple note sequences

55

55

End-to-end note tracking

- Google's response to piano transcription [Hawthorne et al, 2018]
- Combines two networks, one for detecting onsets and one for detecting pitches
- Current state-of-the-art in piano transcription

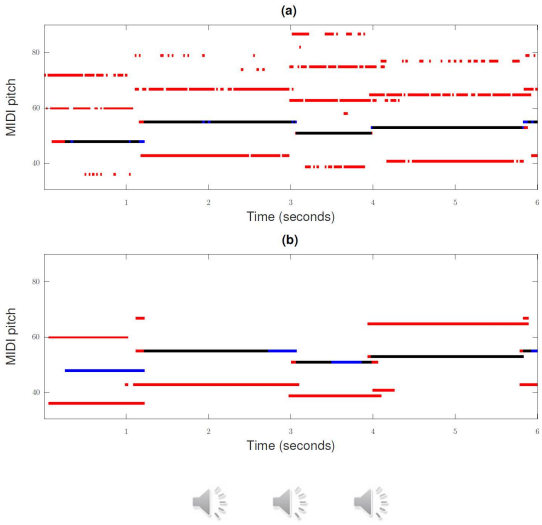


56

56

Comparison between NMF and NNs

- NNs are able to represent complex manifolds in a robust and efficient way
- NNs can be trained in an end-to-end fashion
- NMF performs well with small datasets
- NMF easily adapts to new acoustic conditions

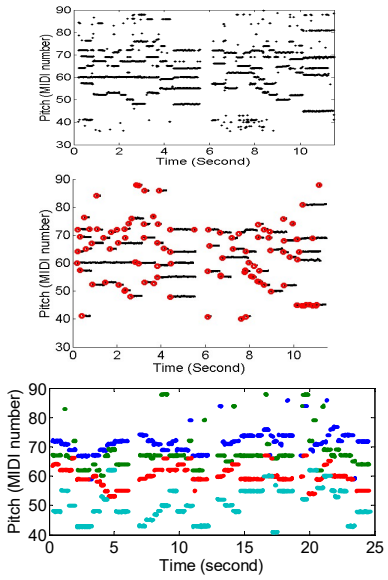


57

57

State of the Art recap

- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods
- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods
- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods



58

58

Multi-pitch Streaming (Timbre Tracking)

- Supervised
  - Train timbre models of sound sources
  - Apply timbre models during pitch estimation: [Cont et al., 2007; Benetos et al., 2013]
  - Classify estimated pitches/notes: [Wu et al. 2011]
- Supervised with timbre adaptation
  - Adapt trained timbre models to sources in mixture: [Grindlay & Ellis, 2011]
- Unsupervised
  - Cluster pitch estimates according to timbre [Duan et al., 2014; Mysore & Smaragdis, 2009]

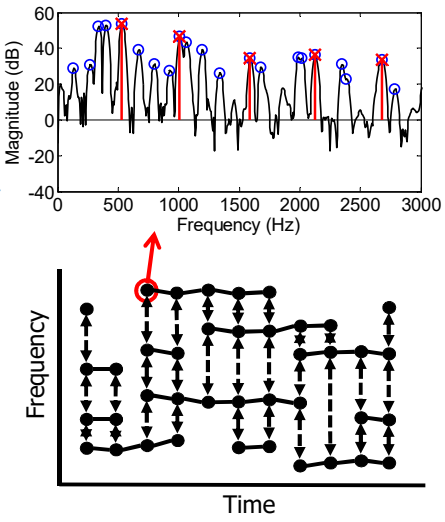
59

59

Timbre Tracking – Unsupervised (1)

[Duan et al., 2014]

- Constrained clustering
  - Objective: maximize timbre consistency within clusters
  - Constraints based on pitch locations: must-links and cannot-links
- Timbre representation: harmonic structure feature
- Iterative algorithm: update clustering to monotonically decrease objective function and satisfy more constraints



60

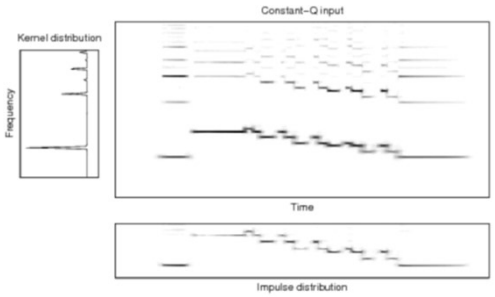
60

Timbre Tracking – Unsupervised (2)

[Mysore & Smaragdis, 2009] for relative pitch tracking

- Shift-invariant PLCA on constant-Q spectrogram
  - Assumption: instrument spectrum shape invariant to pitch
  - Constraints: 1) note activation over frequency shift is unimodal; 2) note activation over time is smooth
- Can be viewed as a pitch clustering algorithm

- Pros: pitch estimation and timbre tracking are performed at the same time
- Cons: does not recognize the absolute pitch



61

61

Datasets

62

62

Datasets (1)

- Hard to come by!
- Annotations can be generated:
  - Automatically (e.g. from a Disklavier piano, or by single-pitch detection on multi-track recordings)
  - Semi-automatically (e.g. manual corrections from F0 tracking or alignment)
  - Manually (e.g. annotating each note, playing back the music on a digital instrument [Su15b])
- Dataset types:
  1. Chords/isolated notes
  2. Music pieces

63

63

Datasets (2)

Polyphonic datasets – chords/isolated notes

1. UIOWA Musical Instrument Samples

   http://theremin.music.uiowa.edu/MIS.html
   - mono/stereo recordings for woodwind, brass, and string - instruments + percussion (isolated notes)

2. RWC Musical Instrument Sounds

   https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-i.html
   - Isolated sounds for 50 instruments (incl. percussion)
   - Covers different playing styles, dynamics, instrument models

64

64

Datasets (3)

Polyphonic datasets – chords/isolated notes

3. McGill University Master Samples

   - 3 DVDs – cover orchestral instruments + percussion
   - Available through select libraries – dataset owned by Garritan

4. MAPS samples

   http://www.tsi.telecom-paristech.fr/aao/
   - Part of MIDI-aligned Piano Sounds database (MAPS)
   - Isolated notes, random chords, usual chords
   - 9 different piano models (virtual pianos + Disklavier)

65

65

Datasets (4)

Polyphonic datasets – music pieces

1. RWC database - classical subset

   https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-c.html
   - 50 recordings (solo performances, chamber, orchestral music…)
   - Non-aligned MIDI provided
   - syncRWC annotations (through automatic alignment):
   https://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SyncRWC/

2. RWC database – jazz subset

   https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-j.html
   - 50 recordings (different instrumentations/style variations)
   - Non-aligned MIDI provided
   - Automatically aligned MIDI (5 recordings incl. percussion):
   http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/37

66

66

Datasets (5)

Polyphonic datasets – music pieces

3. MAPS database

   http://www.tsi.telecom-paristech.fr/aao/
   - 9 different piano models (virtual pianos + Disklavier)
   - 9 x 30 complete classical pieces + MIDI ground truth

4. TRIOS dataset

   http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27
   - 5 multitrack recordings of classical/jazz trios
   - MIDI ground truth provided

67

67

Datasets (6)

Polyphonic datasets – music pieces

5. LabROSA Automatic Piano Transcription dataset

   http://labrosa.ee.columbia.edu/projects/piano/
   - Disklavier piano + MIDI ground truth (29 pieces)

6. Bach10 dataset

   http://www.ece.rochester.edu/~zduan/resource/Resources.html
   - 10 multitrack recordings (violin, clarinet, sax, bassoon quartet)
   - MIDI ground truth provided (semi-automatic)

68

68

Datasets (7)

Polyphonic datasets – music pieces

7. MIREX multiF0 development dataset

   http://www.music-ir.org/evaluation/MIREX/data/2007/multiF0/index.htm
   (password required – ask MIREX team!)
   - One woodwind quintet multitrack recording + manual MIDI annotation

8. MusicNet

   https://homes.cs.washington.edu/~thickstn/musicnet.html
   - 330 freely-licensed classical music recordings
   - Automatically aligned pitch & instrument annotations

69

69

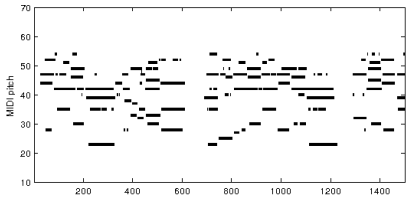
Evaluation Metrics

70

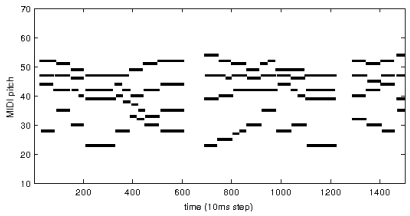
70

Evaluation Metrics (1)

- Typically comparing piano-rolls or MIDI-like representations (e.g. onset-offset-pitch)




```
0.2100    0.8000    47.0000
0.2300    0.7600    44.0000
0.4100    0.8800    52.0000
0.4300    0.7200    28.0000
0.7900    1.6800    42.0000
0.8500    0.9800    47.0000
0.9100    1.4600    35.0000
0.9900    1.3000    47.0000
1.1500    1.4400    51.0000
1.4700    2.0000    46.0000
```




```
0.2100    0.8800    52.0000
0.2100    0.8000    47.0000
0.2200    0.7700    44.0000
0.2800    0.8300    28.0000
0.7800    1.4500    42.0000
0.8400    1.4500    47.0000
0.8900    1.4500    51.0000
0.8900    1.4500    35.0000
1.4800    2.0100    49.0000
1.4800    2.0100    46.0000
```

71

71

Evaluation Metrics (2)

- Evaluation on:
  - Multi-pitch detection
  - Instrument assignment
    (i.e. assign each detected note to an instrument source)
  - Polyphony level estimation (e.g. [Klapuri03, Duan10])

- Evaluation methodologies:
  - Frame-based
  - Note-based

72

72

Evaluation Metrics (3)

Frame-based evaluation

- Comparing the transcribed output and the ground truth frame-by-frame, typically at 10ms step (as in MIREX MultiF0 task).
- Accuracy [Dixon, 2000]:

$$Acc_1 = \frac{\sum_n N_{tp}[n]}{\sum_n N_{fp}[n] + N_{fn}[n] + N_{tp}[n]}$$

  - $N_{tp}[n]$: # true positives
  - $N_{fp}[n]$: # false positives
  - $N_{fn}[n]$: # false negatives

73

73

Evaluation Metrics (4)

Frame-based evaluation

- Accuracy (alternative metric – Kameoka et al, 2007):

$$Acc_2 = \frac{\sum_n N_{ref}[n] - N_{fn}[n] - N_{fp}[n] + N_{subs}[n]}{\sum_n N_{ref}[n]}$$

  - $N_{subs}[n] = \min(N_{fn}[n], N_{fp}[n])$ (# pitch substitutions)
  - $N_{ref}[n]$: # ground-truth pitches at frame n

- Chroma accuracy: pitches warped into one octave
- Precision – Recall – F-measure:

$$Pre = \frac{\sum_n N_{tp}[n]}{\sum_n N_{sys}[n]} \quad Rec = \frac{\sum_n N_{tp}[n]}{\sum_n N_{ref}[n]} \quad \mathcal{F} = \frac{2 \cdot Rec \cdot Pre}{Rec + Pre}$$

  - $N_{sys}[n]$: # detected pitches

74

74

Evaluation Metrics (5)

Note-based evaluation

- Each note is characterized by its onset, offset, and pitch
- Onset-only evaluation: a note event is considered correct if its onset is within a tolerance (e.g. +/-50ms) and its pitch within a tolerance (e.g. quarter tone) of a ground truth pitch
- P-R-F metrics can be defined
- Onset-offset evaluation: additional constraint for offset tolerance (e.g. +/-50ms tolerance **or** offset within 20% of GT note's duration)

```
0.2100    0.8000   47.0000
0.2300    0.7600   44.0000
0.4100    0.8800   52.0000
0.4300    0.7200   28.0000
0.7900    1.6800   42.0000
0.8500    0.9800   47.0000
0.9100    1.4600   35.0000
0.9900    1.3000   47.0000
1.1500    1.4400   51.0000
1.4700    2.0000   46.0000
```

```
0.2100    0.8800   52.0000
0.2100    0.8000   47.0000
0.2200    0.7700   44.0000
0.2800    0.8300   28.0000
0.7800    1.4500   42.0000
0.8400    1.4500   47.0000
0.8900    1.4500   51.0000
0.8900    1.4500   35.0000
1.4800    2.0100   49.0000
1.4800    2.0100   46.0000
```

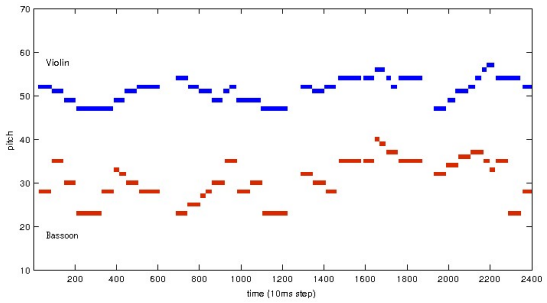
75

75

Evaluation Metrics (6)

Instrument assignment

- A pitch is only considered correct if it occurs at the correct time and is assigned to the proper instrument source
- Similar metrics as in multi-pitch detection can be defined



76

76

Public Evaluation

Public Evaluation (1)

MIREX Multi-F0 Estimation and Note Tracking task

- Subtasks:
  - Task 1: Frame-based evaluation (multiple instruments)
  - Task 2a: Note-based evaluation (multiple instruments)
  - Task 2b: Note-based evaluation (piano only)
  - Task 3: Timbre tracking (i.e. instrument assignment – not run often…)
- Dataset:
  - Woodwind quintet
  - Synthesized pieces using RWC MIDI and RWC samples
  - Polyphonic piano recordings
  - New dataset since 2015:
    piano solo, string quartet, piano quintet, violin sonata

https://www.music-ir.org/mirex

Public Evaluation (2)

MIREX Multi-F0 Estimation and Note Tracking task

- Results for Task 1 (frame-based accuracy)

Teams	2011	2012	2013	2014	2015	2017
Yeh and Roebel	0.68	-	-	-	-	-
Dressler	0.63	0.64	-	0.68	-	0.66
Benetos and Dixon/Weyde	0.57	0.58	0.66	0.66	0.66	-
Fuentes et al.	-	0.56	-	-	-	-
Elowsson and Friberg	-	-	-	0.72	-	-
Cheng et al.	-	-	0.62	-	-	-
Su and Yang	-	-	-	0.64	0.59	-
Thickstun et al.	-	-	-	-	-	0.72

79

79

Public Evaluation (3)

MIREX Multi-F0 Estimation and Note Tracking task

- Results for Task 2 (onset/only F-measure)

Teams	2011	2012	2013	2014	2015	2017
Yeh and Roebel	0.56	-	-	-	-	-
Dressler	-	0.65	-	0.66	-	0.69
Benetos and Dixon/Weyde	0.45	0.43	0.55	0.58	0.60	-
Fuentes et al.	-	0.61	-	-	-	-
Elowsson and Friberg	-	-	-	0.82	-	-
Cheng et al.	-	-	0.50	-	-	-
Su and Yang	-	-	-	0.46	0.47	-
Böck	-	0.50	-	0.54	-	-
Duan and Temperley	-	-	-	0.45	-	-
Thome	-	-	-	-	-	0.76

80

80

Relations & Applications to Other Problems

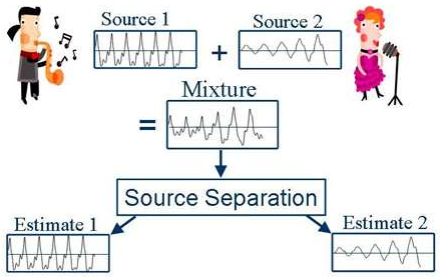
81

81

Relations to Other Problems (1)

Music Source Separation

- Interdependent with multi-pitch detection and instrument identification
- Instrument identification can be improved by separating the source signals [Bosch12]
- Joint instrument identification and separation [Itoyama11]



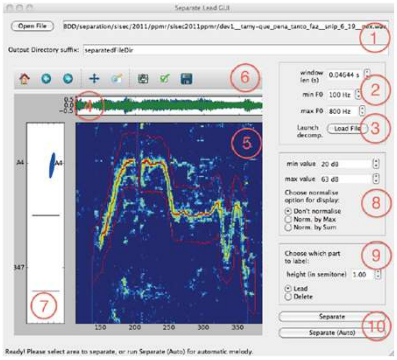
82

82

Relations to Other Problems (2)

Music Source Separation (cont'd)

- Concepts and algorithms from source separation can be utilized for AMT [Durrieu12, Ozerov12]
- Semi-automatic source separation & F0 estimation [Durrieu12]
- **But**: a better source separation does not necessarily imply better multi-pitch detection! [Tavares13b]



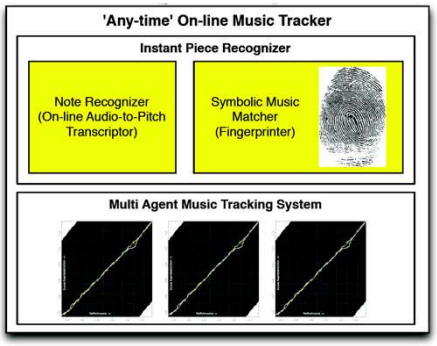
83

83

Relations to Other Problems (3)

Score following

- [Arzt12]: Indentifying score position through transcription-derived pitch- and time-invariant features
- [Duan11]: Use multi-pitch estimation model as the observation model of an HMM for score following (SoundPrism)



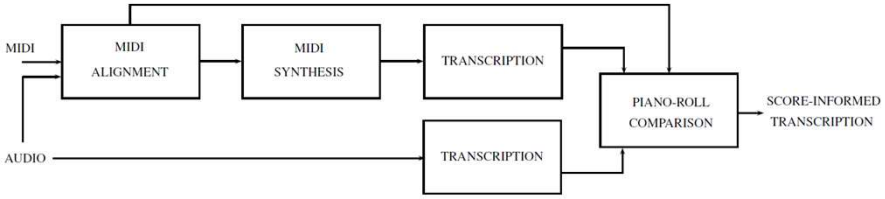
84

84

Relations to Other Problems (4)

Score-informed transcription

- Combining audio-to-score alignment with automatic music transcription
- Applications: automatic instrument tutoring, performance studies
- [Wang08]: Fusing audio & video transcription with score information for violin tutoring
- [Benetos12, Fukuda15]: Score-informed piano tutoring based on NMF
- [Dittmar12]: Songs2See – (based on multi-pitch detection, score-informed source separation, extraction of instrument-specific parameters)



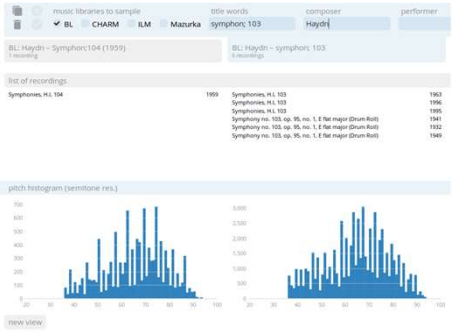
85

85

Relations to Other Problems (5)

Applications to Content-based Music Retrieval

- Deriving high-level features for organising/navigating through audio collections, music similarity & recommendation
- [Lidy07] Music genre classification by combining audio and symbolic descriptors
- [Weyde14] Transcription-derived features for exploring music archives



http://dml.city.ac.uk/vis/

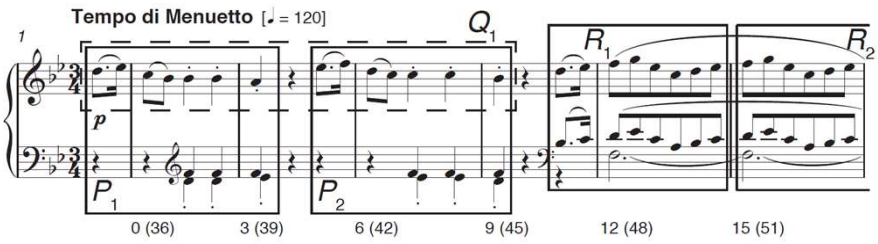
86

86

Relations to Other Problems (6)

Applications to Systematic/Computational Musicology

- [Collins14]: Discovery of repeated themes and patterns from automatically transcribed and beat-quantized MIDI



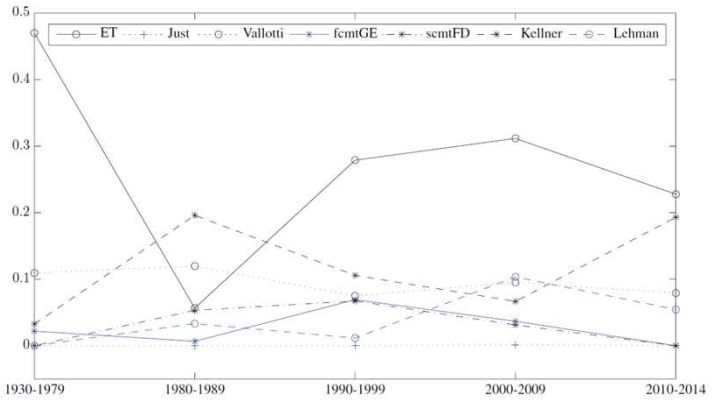
87

87

Relations to Other Problems (7)

Applications to Systematic/Computational Musicology (cont'd)

- [Dixon11; Tidhar14]: Automatic estimation of harpsichord temperament – using a "conservative" transcription as a first step for precise frequency estimation.



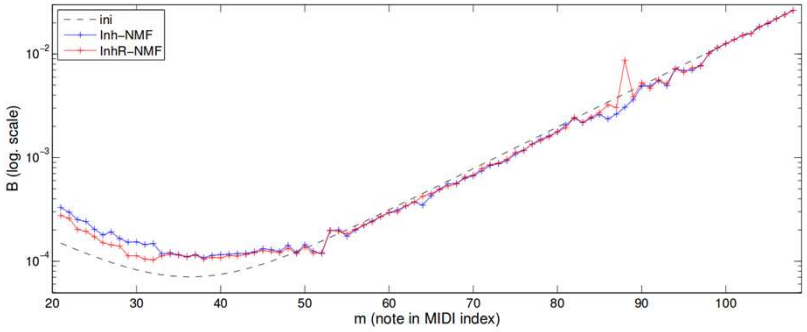
88

88

Relations to Other Problems (8)

Applications to Music Acoustics

- [Rigaud13]: Joint estimation of multiple pitches and inharmonicity for the piano using an NMF-based model



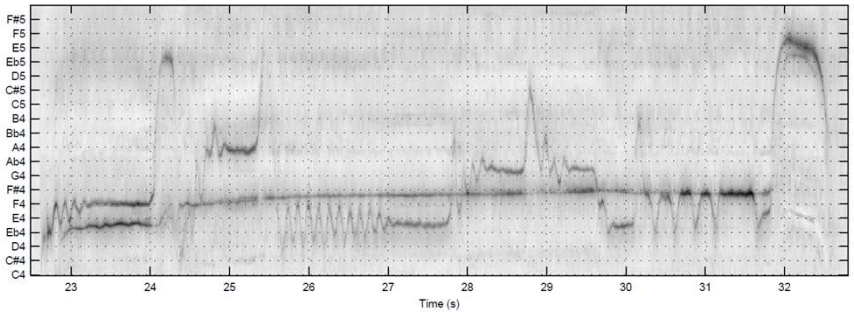
89

89

Relations to Other Problems (9)

Applications to Music Performance Analysis

- [Jure12]: Pitch salience representations for music performance analysis; also used to assist human transcription



90

90

Software & Demo

91

91

AMT Software (1)

Free software / plugins (from academic research)

Authors	Language	URL
Benetos et al	Matlab / C++	http://www.eecs.qmul.ac.uk/~emmanouilb/code.html
Böck	Python	https://github.com/CPJKU/madmom
Duan et al	Matlab	http://www.ece.rochester.edu/~zduan/resource/Resources.html
Fuentes et al	Matlab	http://www.benoit-fuentes.fr/publications.html
Hawthorne et al	Python	https://goo.gl/magenta/onsets-frames-colab
Marolt	win32 executable	http://atlas.fri.uni-lj.si/lgm/transcription-of-polyphonic-piano-music/
Pertusa & Iñesta	Vamp plugin + online prototype	http://grfia.dlsi.ua.es/cm/projects/drims/softwareVAMP.php
Raczyński et al	R / Python	http://versamus.inria.fr/software-and-data/multipitch.tar.bz2
Vincent et al	Matlab	http://www.irisa.fr/metiss/members/evincent/software

92

92

AMT Software (2)

Commercial software / plugins

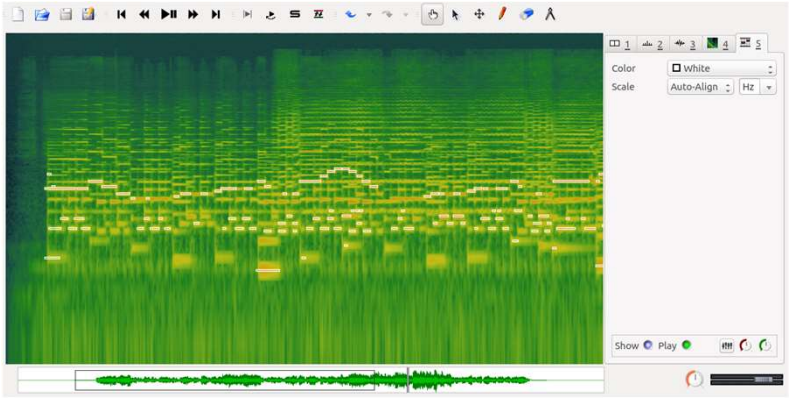
Name	URL
Akoff Sound Labs	http://www.akoff.com/audio-to-midi.html
intelliScore	http://www.intelliscore.net
Melodyne	http://www.celemony.com
PitchScope	http://www.creativedetectors.com/
ScoreCloud	https://scorecloud.com/
Sibelius AudioScore	http://www.sibelius.com/products/audioscore/ultimate.html
Solo Explorer	http://www.recognisoft.com/
Transcribe!	http://www.seventhstring.com/xscribe/
WIDISOFT audio-to-MIDI VST plugin	http://www.widisoft.com/english/translate.html

93

93

Demo

Silvet Vamp plugin



Silvet download: https://code.soundsoftware.ac.uk/projects/silvet/files

Sonic Visualiser download: http://www.sonicvisualiser.org/download.html

94

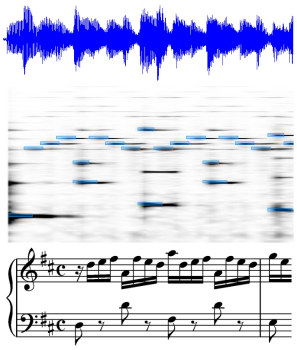
94

Further extensions and advanced topics

95

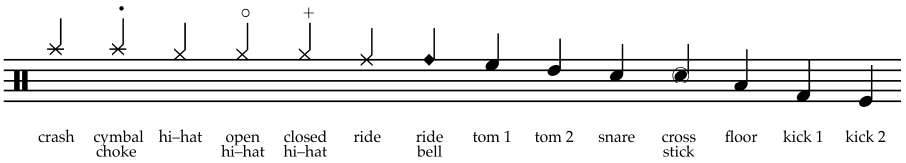
Further Extensions and Advanced Topics

1. Percussive instruments
2. Singing voice
3. Non-Western music
4. Music language models
5. Complete AMT
6. Evaluation measures



96

Transcribing Percussive Instruments



97

97

Percussive Instruments Transcription (1)

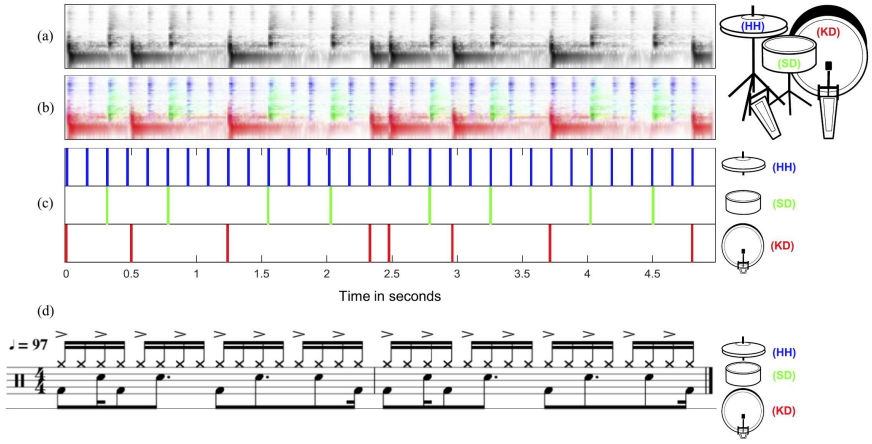
- **Core application**: automatic drum transcription (ADT)
- **Literature**:
  - Transcribing solo drums
  - Reducing percussive sounds for transcribing pitched sounds
  - Transcribing drums in the presence of pitched sounds
  - Transcribing drums & pitched sounds

98

98

Percussive Instruments Transcription (2)

Spectrogram and corresponding transcription of drum sounds:



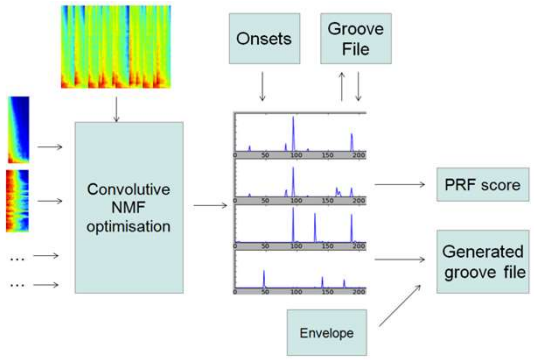
99

99

Percussive Instruments Transcription (3)

Spectrogram decomposition approaches

- [Lindsay-Smith et al, 2012]: convolutive NMF with time-frequency patches
- [Dittmar and Gärtner, 2014]: realtime transcription + separation with NMF and semi-adaptive bases
- [Benetos et al, 2014]: transcribing drums + pitched sounds using supervised PLCA



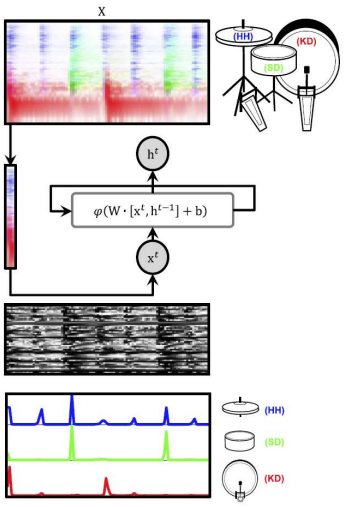
100

100

Percussive Instruments Transcription (4)

Neural network approach

- [Vogl et al, 2017]: recurrent neural networks
- [Southall et al, 2017]: attention mechanisms and convolutional neural networks
- Comparison between NMF and NNs [Wu et al, 2018]: RNNs outperform NMF, however difference becomes marginal for challenging datasets



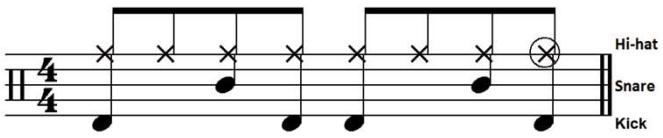
101

101

Percussive Instruments Transcription (5)

Discussion

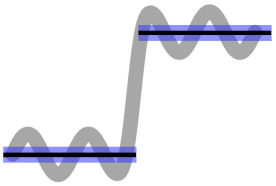
- Good performance for drum transcription in a supervised scenario, even in real-time applications
- Temporal accuracy needed is higher compared to pitched sounds!
- Source adaptation: significant improvement, but more work needed for handling dense drum polyphony & complex patterns
- Open problem: transcribing both drums & pitched sounds (also: lack of data for evaluation!)



102

102

Transcribing Singing Voice



103

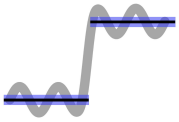
Transcribing Singing Voice (1)

Motivation

- All popular music cultures around the world use singing
- The singing voice is the most expressive of all musical instruments
- Common representations (e.g. MIDI, Western notation) are inadequate for expressive singing

Challenges

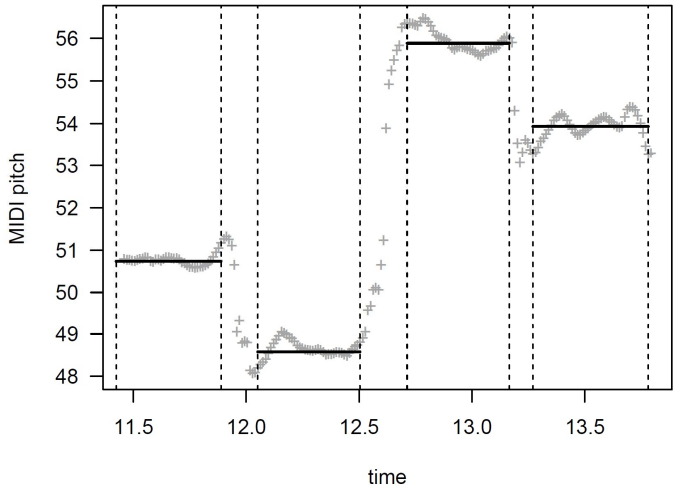
- Phonation modes: voiceless, breathy, flow, pressed, glottal stop…
- Vocal fold oscillation modes: vocal fry, falsetto, modal…
- Different singing styles and techniques (e.g. choral, pop, theatrical, overtone singing…)
- Intonation, drift, poor singing!



104

Transcribing Singing Voice (2)

Example: Note Segmentation and Framewise/Notewise Pitch Estimates

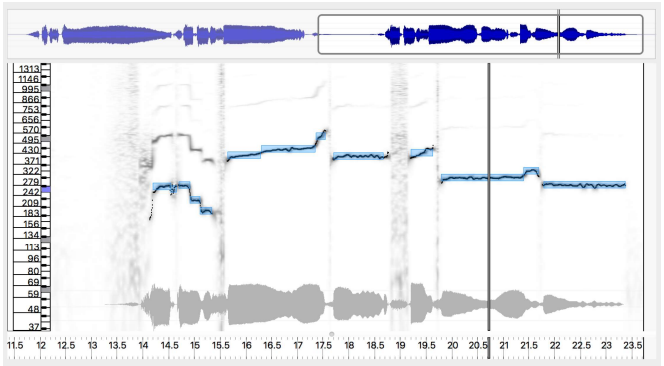


105

105

Transcribing Singing Voice (3)

- Singing voice mostly analyzed in the context of monophonic pitch detection
- Current state-of-the-art: MELODIA [Salamon & Gomez, 2012], pYIN [Mauch & Dixon, 2014], CREPE [Kim et al, 2018]



https://code.soundsoftware.ac.uk/projects/tony

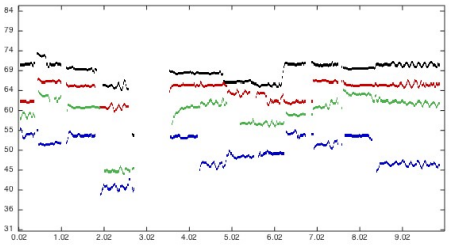
106

106

Transcribing Singing Voice (4)

Challenges

- Note tracking / contour tracking: less explored problem and ill-defined for music that is not sung from a fixed note representation
- Modelling vocal techniques, pitch drift [Mauch et al, 2014]
- Transcribing singing and accompaniment, transcribing multiple singers and choral music



[Schramm & Benetos, 2017]

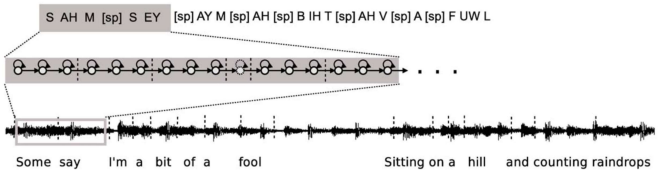
107

107

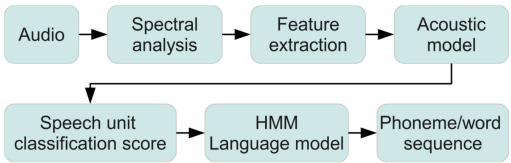
Transcribing Singing Voice (5)

Lyrics Transcription

- Recognizing phonemes/words from singing [Mesaros & Virtanen, 2010]



- Typically based on automatic speech recognition systems
- Vowel types are also important when distinguishing colours in singing voice



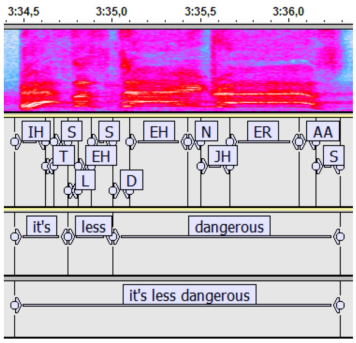
108

108

Transcribing Singing Voice (6)

Lyrics Transcription (continued)

- Figure: sentence-level, word-level, and phoneme-level estimation for singing
- Challenges: lack of annotated datasets, singing voice & instrumental accompaniment, mispronunciation [Gupta et al, 2017; 2018]



[Hansen, 2012]

Vowels		
Error	**Actual word**	**What is spoken**
ow-->ao	golden	gawlden
uw-->uh	fool	full
iy-->ih,ix	seem, sees, sleeping,	sim, sis, slipping
eh-->ae	every	avry

109

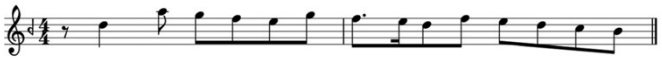
AMT for non-Western music

110

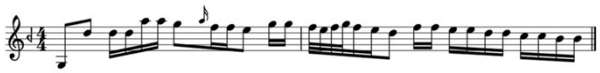
AMT for non-Western music (1)

Automatic transcription of world, folk, and traditional music

- The vast majority of AMT research assumes 12-TET
- Another assumption: monophony/polyphony (whereas in several cultures music is heterophonic), major/minor modes
- Research on transcribing non-Western/traditional music:
  - [Gómez13]: Automatic transcription of (a capella singing) flamenco recordings
  - [Bozkurt08; Benetos15]: Pitch analysis and transcription for Turkish makam music
  - [Srinivasamurthy14]: Transcribing percussion patterns in Chinese opera
  - [Kelleher05]: Transcription & ornament detection for Irish fiddle



(a) Melody as notated



(b) Transcription of *oud* performance

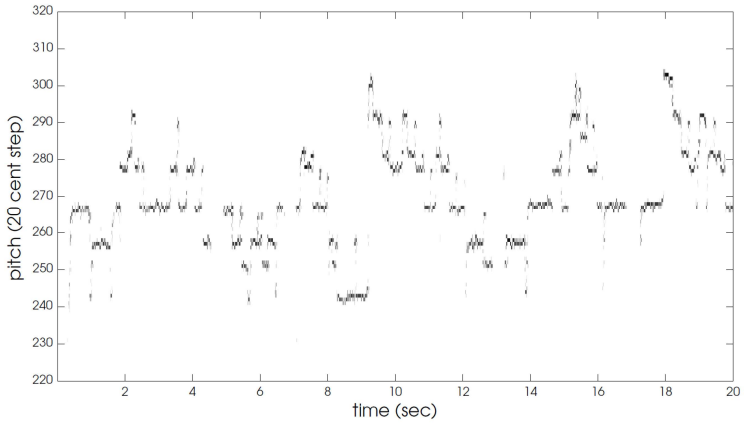
111

111

AMT for non-Western music (2)

Automatic transcription Turkish microtonal music

- Mode-informed system, supporting 20 cent resolution, tonic detection, and heterophony [Benetos & Holzapfel, 2015]



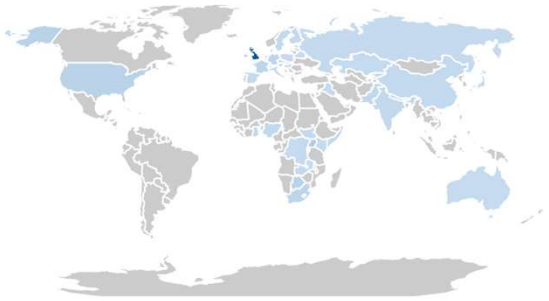
112

112

AMT for non-Western music (3)

Automatic transcription of world & traditional music

- DML system: 20-cent time-pitch representations for 60k recordings of the British Library Sound Archive - http://dml.city.ac.uk/vis/
- Open problems:
  - Data! (recordings & annotations)
  - Methodology: culture-specific vs. general-purpose systems
  - Prescriptive vs descriptive notation
  - Engagement from the ethnomusicology community (changing: FMA, AAWM…)



113

113

Music Language Models

114

114

Music Language Models (1)

Incorporating musical knowledge

- Most existing transcription approaches are data-driven
  - Errors are not musically meaningful; could be avoided by incorporating musical knowledge
- Musicians rely on musical knowledge to transcribe music
  - Key signature, harmony, metrical structure
  - Counterpoint and other composition rules
- Speech recognition successfully integrates acoustic models and language models, although these models cannot be directly applied to AMT:
  - Music is polyphonic
  - Music rhythm involves much longer temporal dependencies
  - Music harmony arrangement involves rich music theory

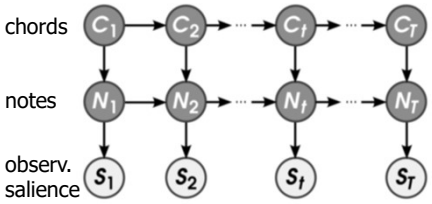
115

115

Music Language Models (2)

[Raczynski et al., 2013] Dynamic Bayesian Networks

- Chord model: chord transition
- Note model: linear combination of the following sub-models:
  - *Harmonic*
  - *Duration*
  - *Voice*
  - *Polyphony*
  - *Neighbor*
- All models first-order Markovian
- 3% F-measure improvement from an NMF-based AMT approach



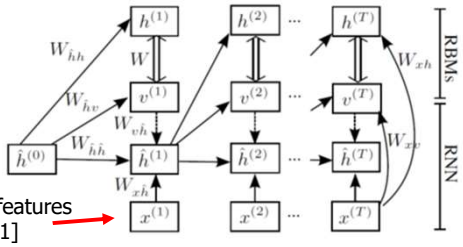
116

116

Music Language Models (3)

Model temporal dependencies with RNN-RBM

- Joint optimization by RNN-RBM [Boulanger-Lewandowski13]



RBM output features using [Nam11]

Tests on mostly synthetic piano data

- Method achieves 1%-10% frame-level transcription accuracy improvement

117

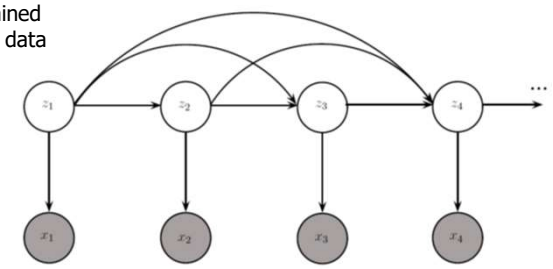
Music Language Models (4)

Hybrid acoustic-language AMT model [Sigtia et al, 2016]

- DNNs/RNNs/CNNs for acoustic model
- RNN-NADE for language model
- Probabilistic graphical model for combining acoustic and language models

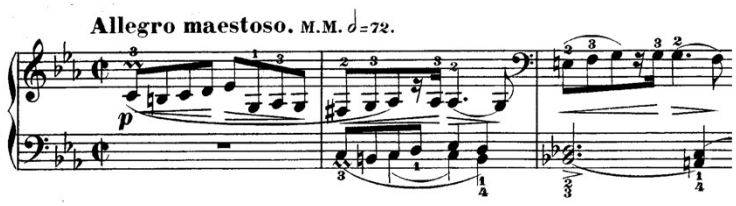
RNN-NADE trained from symbolic data

DNN or RNN or DNN+RNN trained from acoustic data



118

Towards a Complete Music Transcription



119

119

Complete AMT (1)

Current AMT systems can (up to a point!):

- Detect (multiple) pitches, onsets, offsets
- Identify instruments in polyphonic music
- Assign detected notes to a specific instrument

Also, some systems are able to:

- Detect & integrate rhythmic information
- Detect tuning (per piece/note)
- Extract velocity per detected note
- Transcribe fingering (for specific instruments)
- Quantise pitches over time/beats

Significant work needs to be done in order to extract a complete score

120

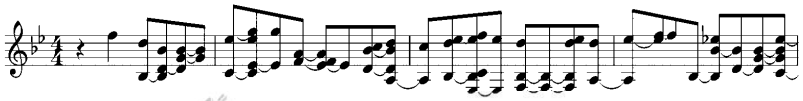
120

Complete AMT (2)

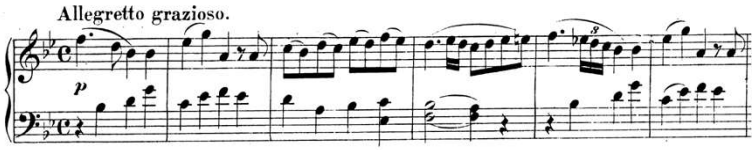
Computer Music Engraving / Typesetting from MIDI

- Various software tools:
  Sibelius, MuseScore, Finale, LilyPond, MaxScore, ScoreCloud…
- Most literature from the point of software development – little information on objective/user evaluation
- Unknown performance on engraving "noisy" scores from AMT systems

MuseScore-generated score of a MIDI transcription (MAPS_MUS-mz_333_3)



Synthesized MIDI:

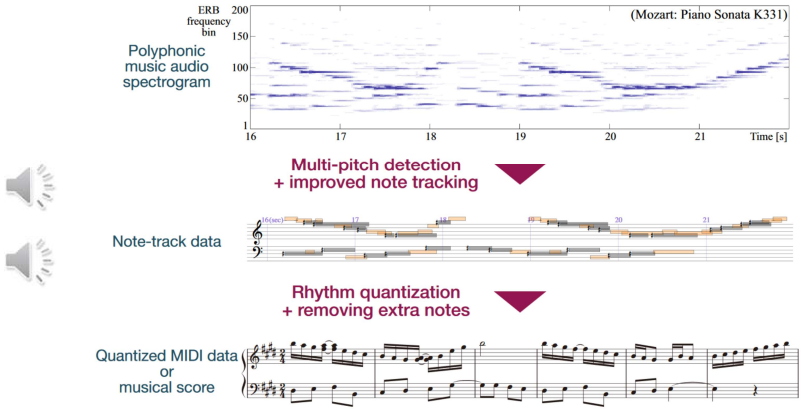


121

Complete AMT (3)

Integrating multi-pitch detection & rhythm quantization

- NMF-based multi-pitch detection followed by HMM-based rhythm quantization and removing extra notes [Nakamura et al, 2018]

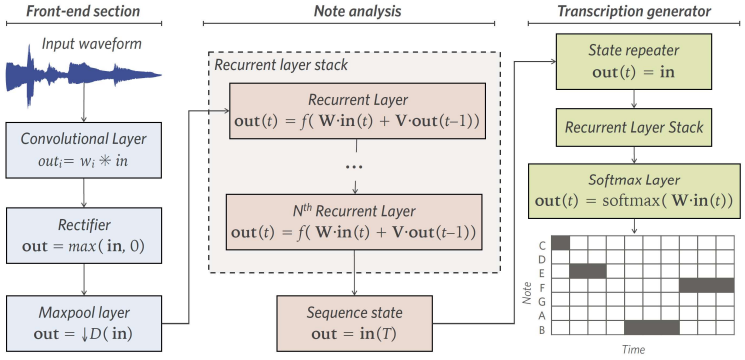


122

Complete AMT (4)

End-to-end complete AMT [Carvalho & Smaragdis, 2017]

- Sequence-to-sequence model for monophonic transcription
- Output in lilypond notation
- Potentially useful for complete polyphonic AMT



123

123

Evaluation Measures

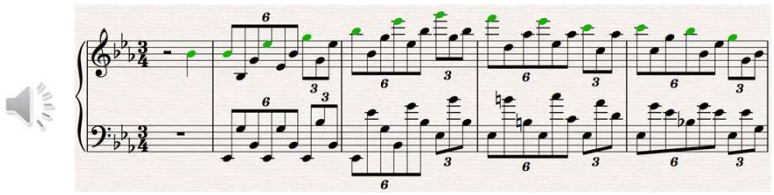
124

124

Evaluation Measures (1)

Design musically meaningful evaluation measures

- Some notes are more musically important



- Some errors are more musically annoying
  - Inharmonic errors > harmonic/octave errors
  - Wrong notes outside the scale > wrong notes within the scale
- The annoyingness depends on the application
  - For music re-synthesis: insertion errors > miss errors
  - For music search: octave errors > semitone errors

125

125

Evaluation Measures (2)

Some ideas for designing musically meaningful measures

- Observation approach: Analyze how music teachers grade music dictation exams
  - Quantitative analysis of music teachers' evaluation measures
  - Well supported by music theory and music education practice
  - Depends on the type of music
  - Errors made by music students cannot represent errors made by computers
- Experiment approach: Subjective listening tests on different types of algorithmically generated errors
  - Analyze correlations between the presence of errors and the listening experience
  - Full control and easy generation of different types of error
  - Difficult to find enough qualified subjects

126

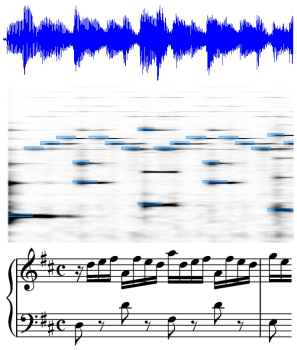
126

Conclusions

Conclusions (1)

State of the field

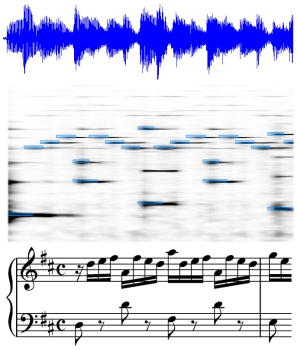
- Continues to attract attention in the MIR, audio and ML/AI research communities
- Performance (objective + perceptual) has increased over the last decade
- Instrument- and style-specific AMT systems have sufficiently good performance for end-user applications
- AMT-derived features are useful for computing high-level music descriptors



Conclusions (2)

State of the field (cont'd)

- As the scope of AMT research continues to grow – increasing number of open problems & sub-problems!
- Agreement that a successful AMT system cannot rely only on information from the acoustic signal. Input needed from:
  - Music acoustics
  - Music theory/language
  - Music perception
- Unified methodology



129

129

Thanks for listening!

130

130