Automatic Music Transcription

Zhiyao Duan [1] and Emmanouil Benetos [2]

[1] Department of Electrical and Computer Engineering,
University of Rochester

[2] Centre for Digital Music, Queen Mary University of London

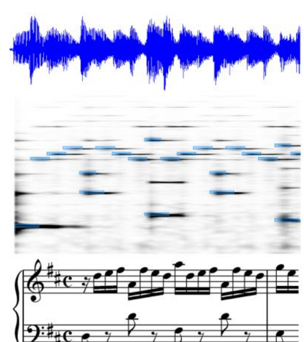
Tutorial at ISMIR 2015
Malaga, Spain
October 26, 2015

Tutorial Outline

1. Introduction
2. How do humans transcribe music?
3. State-of-the-art research on AMT (1st part)

   Break

4. State-of-the-art research on AMT (2nd part)
5. Datasets and evaluation measures
5. Relations and applications to other problems
6. Software & Demo
7. Challenges and research directions
8. Conclusions + Q&A



Tutorial Website:
http://c4dm.eecs.qmul.ac.uk/ismir15-amt-tutorial/

Introduction

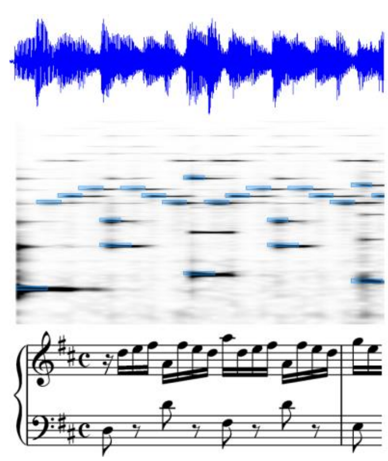
AMT - Introduction (1)

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)

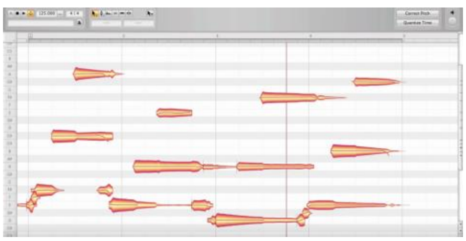
Music notation
- Note name, key, rhythm, instrument
- Uses score time (beat)



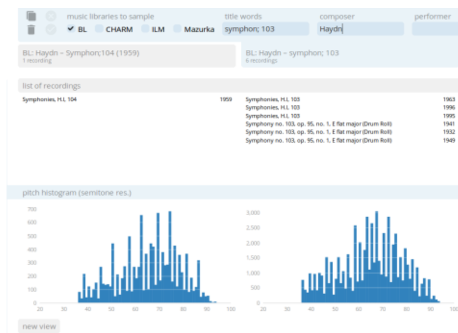
AMT - Introduction (2)

Fundamental (and open) problem in music information research

Applications:
- Search/annotation of musical information
- Interactive music systems
- Systematic/computational musicology



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



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

5

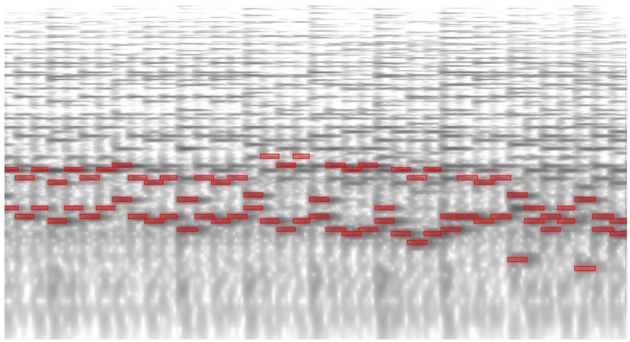
AMT - Introduction (3)

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

6

AMT - Introduction (4)

Core problem: multi-pitch detection



AMT - Introduction (5)

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

AMT - Introduction (6)

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

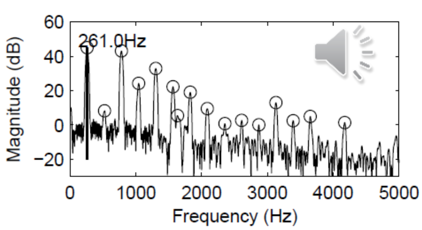
- Can we make computers compete with Mozart?

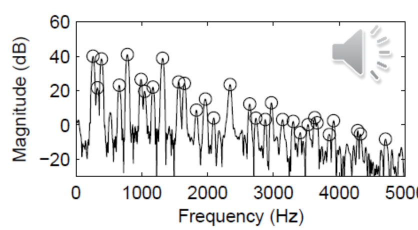
9

AMT - Introduction (7)

Challenges:

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





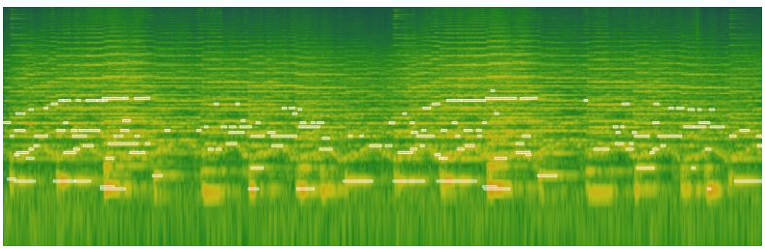
- Large variety of music
  - Music pieces: style, form, etc.
  - Instrumentation: bowed/plucked strings, winds, brass, percussive, etc.
  - Playing technique: legato, staccato, vibrato, etc.

10

AMT - Introduction (8)

State of the Art - Limitations:

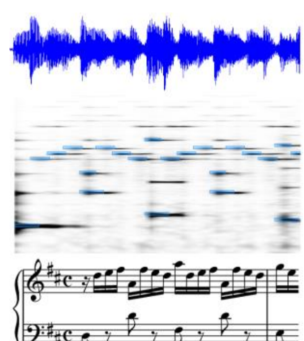
- Performance clearly below of a human expert - especially for multiple-instrument music
- Lack of dataset size/diversity
- No unified methodology (as e.g., automatic speech recognition)
- Little input beyond CS/EE (musicology, music cognition, music acoustics)



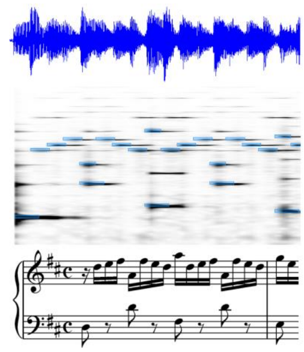
Automatic transcription of B. Smetana – Má vlast (Vltava)

Tutorial Focus/Objectives

- Focusing (mostly) on polyphonic music transcription
- Most work on Western tonal music! We'll try to go beyond that.
- Presenting an overview of representative AMT research (+ related problems)
- Discussion on limitations, challenges, and future directions
- Resources: bibliography, datasets, code, demos
- Tutorial website: http://c4dm.eecs.qmul.ac.uk/ismir15-amt-tutorial/



Tutorial Outline



1. Introduction
2. How do humans transcribe music?
3. State-of-the-art research on AMT (1st part)

   Break

4. State-of-the-art research on AMT (2nd part)
5. Datasets and evaluation measures
5. Relations and applications to other problems
6. Software & Demo
7. Challenges and research directions
8. Conclusions + Q&A

Tutorial Website:
http://c4dm.eecs.qmul.ac.uk/ismir15-amt-tutorial/

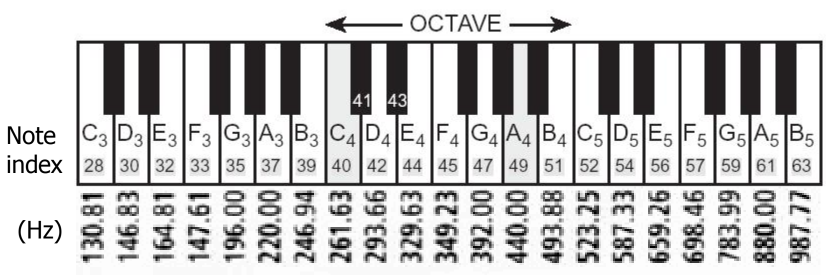
13

How do humans transcribe music?

14

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



15

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

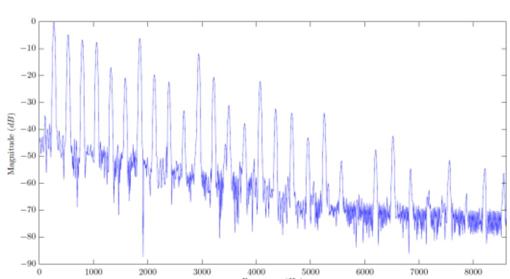
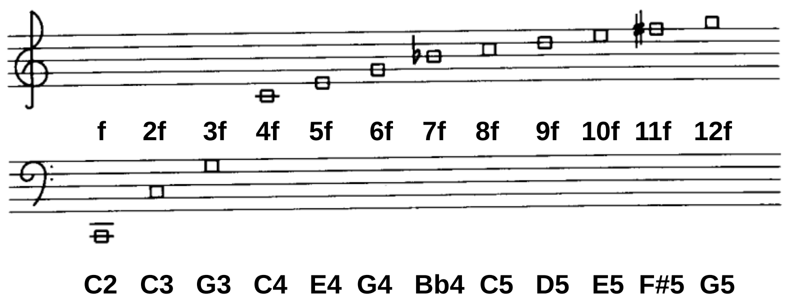


Figure:
spectrum of a C4 piano note. The fundamental is located at 261.6Hz.

16

Pitch Perception (3)

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



17

Pitch Perception (4)

Relative pitch: Ability to recognise 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]

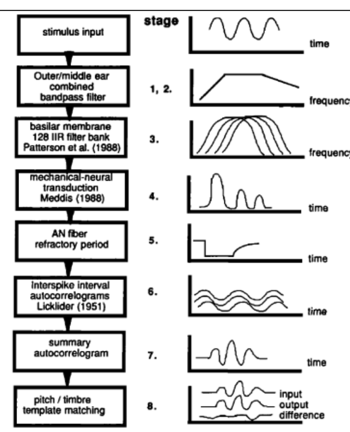


Figure from Meddis & Hewitt, 1991

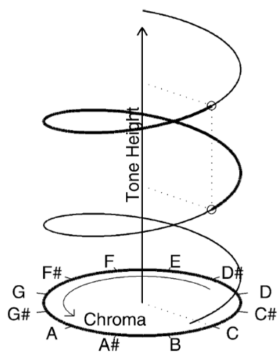
18

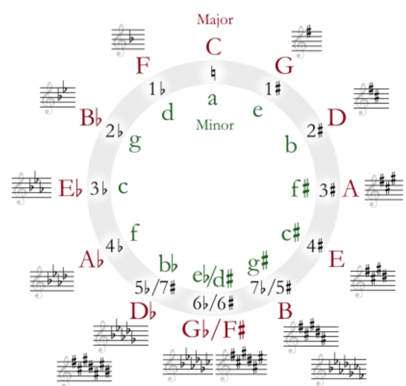
Pitch Perception (5)

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]

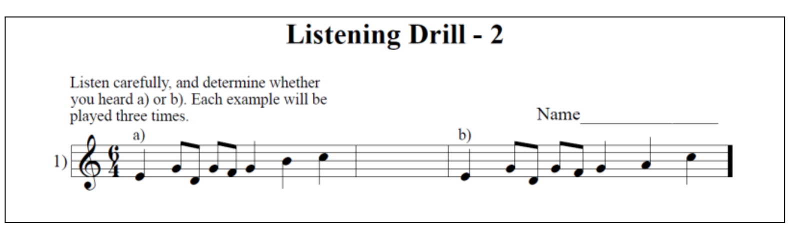




Human Transcription (1)

- 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

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

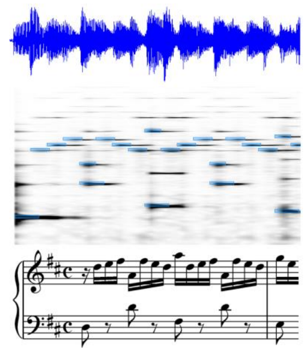
State-of-the-art research in AMT

State-of-the-art Outline

1. Multi-pitch analysis
   A. Frame-level
   B. Note-level
   C. Stream-level

2. Percussive instruments transcription

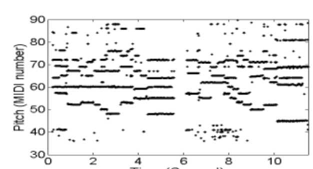
3. Towards a complete music notation



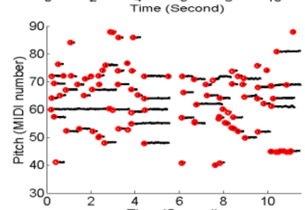
23

State of the Art of Multi-pitch Analysis

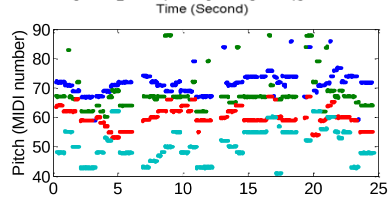
- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods



- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods



- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods



24

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

25

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

26

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

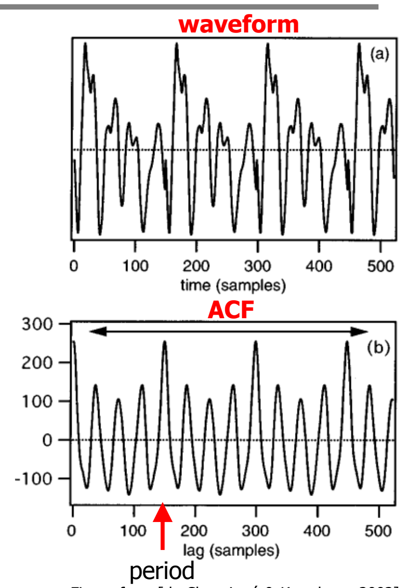


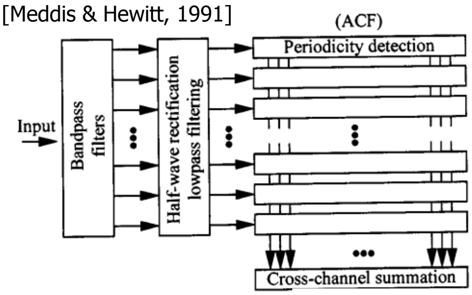
Figure from [de Cheveigné & Kawahara, 2002]

27

Time Domain - Autocorrelation

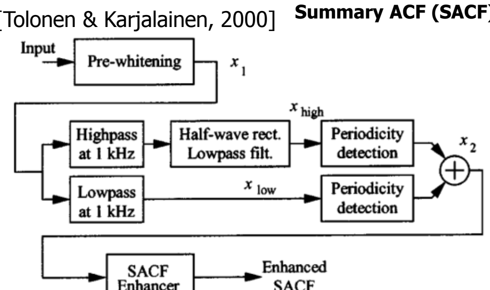
- Detailed simulation of human auditory system
  - Outer- and middle-ear freq. attenuation effect
  - ~100 channels with critical bandwidth
  - Inner hair cell response
- Simplified version
  - Only 2 channels
  - Enhanced SACF: remove SACF peaks due to integer multiples of periods

[Meddis & Hewitt, 1991]



Summary ACF (SACF)

[Tolonen & Karjalainen, 2000]



Figures from [Tolonen & Karjalainen, 2000]

28

Time Domain – Adaptive Oscillators

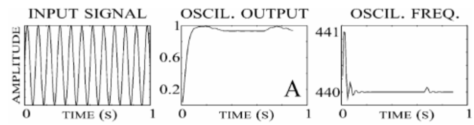
- Oscillator
  - Parameters: *freq.* and phase
- Adaptive oscillator
  - Adapts its freq. and phase to input signal
- Oscillator networks
  - Each network tracks a group of harmonically related partials
  - In total 88 networks for 88 pitches

Pros: good performance on piano

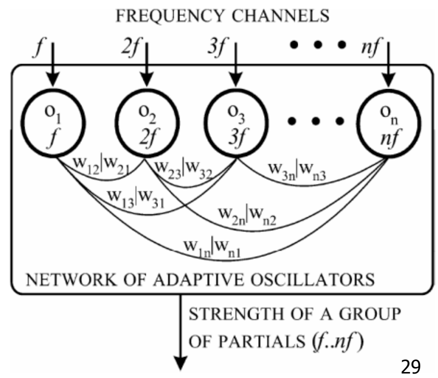
Cons: may not deal well with frequency deviation and modulations

[Marolt, 2004]

A single oscillator



An oscillator network



29

Time Domain – Probabilistic Modeling (1)

- 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

30

Time Domain – Probabilistic Modeling (2)

- A more detailed model [Davy & Godsill, 2003]

$$y_t = \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{i=1}^{I} \underline{a_{k,m,i}\phi_{i,t}} \cos\left[(m + \underline{\delta_{k,m}})\omega_{0,k} t\right] + \underline{b_{k,m,i}\phi_{i,t}} \sin\left[(m + \underline{\delta_{k,m}})\omega_{0,k} t\right] \right\} + v_t$$

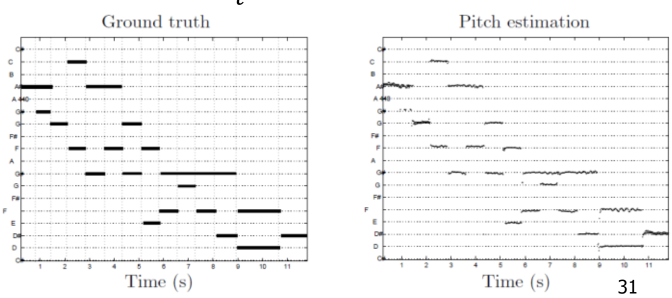
Allows harmonics to change amplitude within a note

Deals with detuning

  - Auto-regressive model for v_t.

Pros: Promising result on real recording (#notes K is provided)

Cons: computational intensive



Time Domain – Probabilistic Modeling (3)

- Damped note model [Cemgil et al., 2006]

state

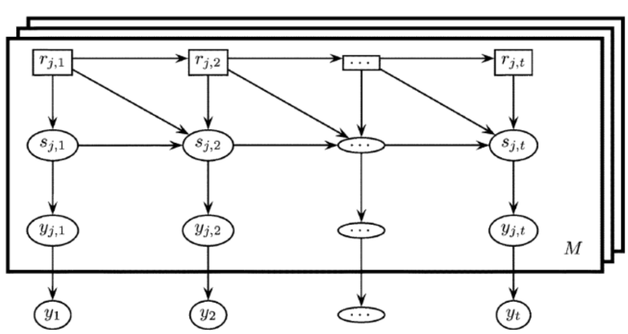
waveform

Note activity (sound/mute)

state

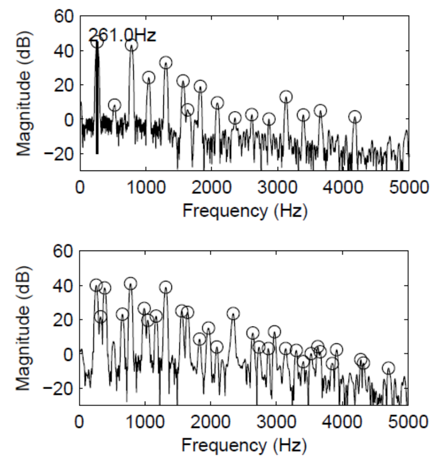
j-th note waveform

Audio mixture



Frequency Domain Methods

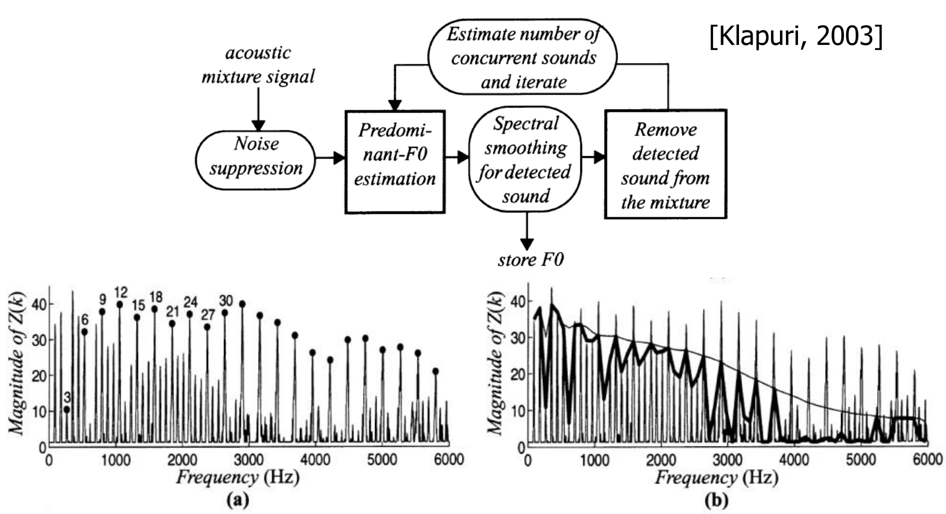
- Key idea
  - Each pitch has a set of harmonics
  - Recognize the harmonic patterns



- Difficulty
  - Tend to have harmonic errors
  - Harmonic amplitude varies
  - Overlapping harmonics

Iterative Spectral Subtraction

[Klapuri, 2003]

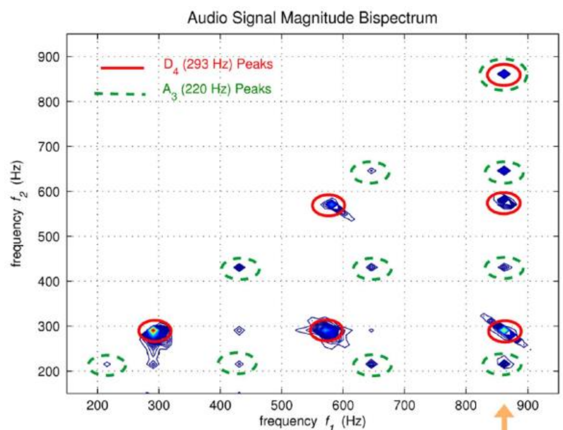


Pros: good performance, simple, fast
Cons: hard to subtract the appropriate amount of energy

Iterative Bispectral Subtraction

- Bispectrum [Argenti et al., 2011]
  - 2-D Fourier transform of the 3rd order cumulant of the signal, or equivalently, $B_x(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2)$
  - Account for nonlinear partial interactions

Algorithm
1. Calculate Constant Q bispectrum of signal
2. Perform 2-d correlation between bispectra of signal and a template
3. Highest correlation gives a pitch estimate
4. Cancel entries of signal bispectrum corresponding to harmonics of the pitch
5. Repeat 2-4.



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 pitche*
- Choose the best ones

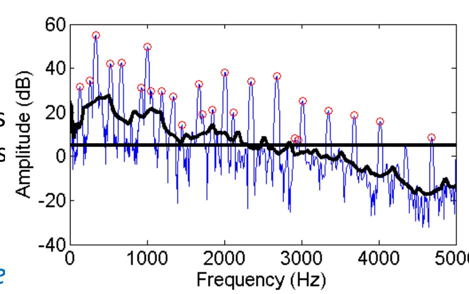


Figure from [Duan et al., 2010]

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

Spectral Peak Modeling – Rule-based

- Rule-based approaches
  - [Pertusa & Iñesta, 2008]
    - Salience(pitch) = Loudness(partials) * Smoothness(partials)
    - Salience(pitch combination) = Sum(saliences of pitches)
  - [Yeh et al., 2010]
    - Salience of a pitch depends on harmonicity, smoothness, and synchronicity of its partials
- Pros: fast, work well
- Cons: rule-based methods may be hard to adapt to other instruments

37

Spectral Peak Modeling – Maximum Likelihood (1)

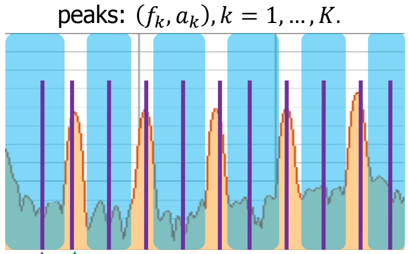
- [Duan et al., 2010]

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

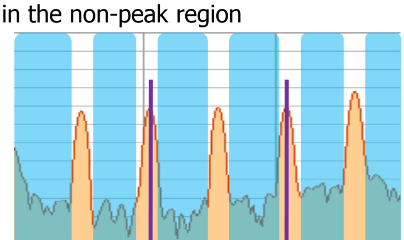
$$p(\boldsymbol{O}|\boldsymbol{\theta}) = p\left(\boldsymbol{O}^{\text{peak}}|\boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{O}^{\text{non-peak}}|\boldsymbol{\theta}\right)$$

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

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

38

Spectral Peak Modeling – Maximum Likelihood (2)

- [Emiya et al., 2007]
  - Auto-Regressive (AR) model for harmonics of pitches
  - Moving-Average (MA) model for residual

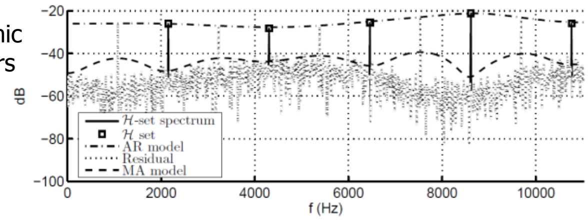
Both tend to be smooth!

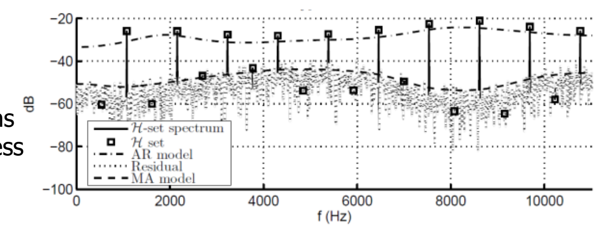
Pros: balances harmonic and subharmonic errors

Harmonic error
- AR model fits well
- MA model doesn't

Subharmonic error
- MA model fits well
- AR model doesn't

Cons: the assumptions on spectral smoothness is not always true





39

Spectral Peak Modeling – Maximum Likelihood (3)

- [Peeling & Godsill, 2011]
  - Assumes that the number of partials of the i-th note is a non-homogenous Poisson process on the frequency axis with a rate of $\lambda_i(f)$, which is the expected partial density at frequency f
  - Assumes that concurrent notes are independent
  - So the number of partials of all notes is a superposition of multiple independent Poisson processes, hence another Poisson process with rate $\lambda(f) = \sum_i \lambda_i(f)$
  - Models $\lambda_i(f)$ with a GMM, with Gaussians centered at harmonics of the i-th note

Likelihood function

$$p(f_1, \ldots, f_N, N|\lambda(f)) = \exp\left(-\int_0^{F_{\max}} \lambda(f)\,\mathrm{d}f\right) \prod_{n=1}^{N} \lambda(f_n)$$

Frequency and number of detected partials (peaks)

Rate function, dependent on pitch hypotheses

Pros: mathematically interesting
Cons: strong assumption

40

Full Spectrum Modeling – Probabilistic (1)

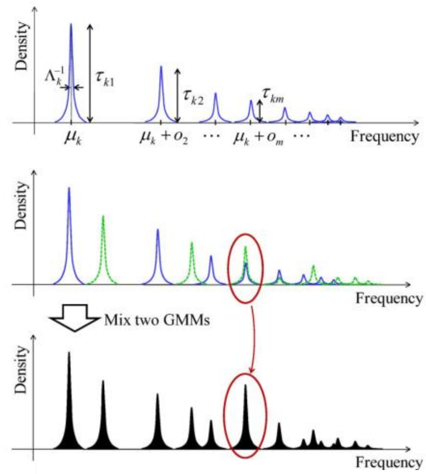
- Key idea: view spectra as (parametric) probabilistic distributions

- Each note = tied- Gaussian Mixture Model (tied-GMM)

$$\mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)$$

- Signal = Mixture of GMMs

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{dk} \mathcal{M}_k(\boldsymbol{x})$$



Figures from [Yoshii & Goto, 2012]

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

Full Spectrum Modeling – Probabilistic (2)

- PreFEst [Goto, 2004]
  - Gaussian models are given; estimate Gaussian mixing weights and note mixing weights
- Harmonic Clustering (HC) [Kameoka et al., 2004]
  - Estimate all parameters
  - Use Akaike Information Criterion (AIC) to decide number of notes
- Infinite Latent Harmonic Allocation (iLHA) [Yoshii & Goto, 2012]
  - Model allows arbitrary number of Gaussians and notes
  - Automatically decide their numbers using non-parametric Bayesian inference

Full Spectrum Modeling – Probabilistic (3)

Non-parametric model

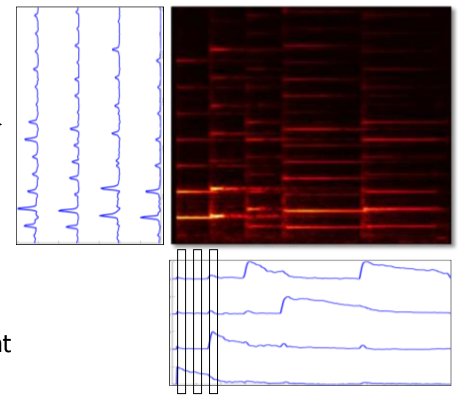
[Smaragdis & Raj, 2006]

- 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

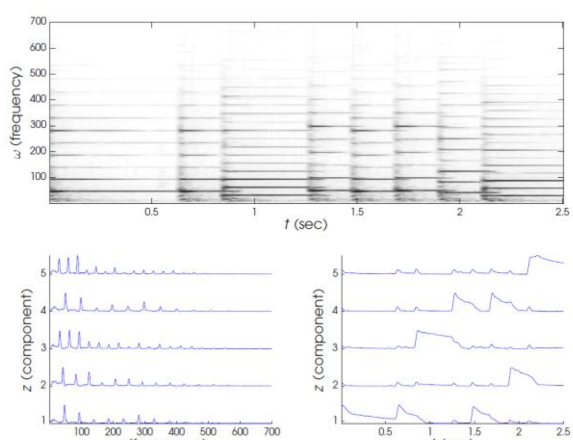
Activation weights $P_t(z)$

Distribution of components

43

Spectrogram Decomposition (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

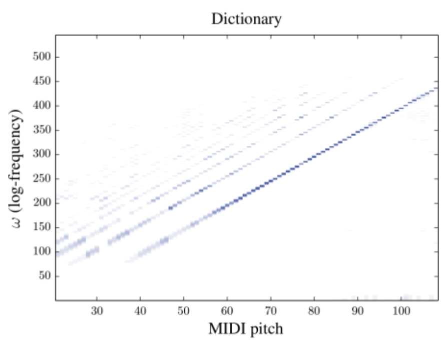


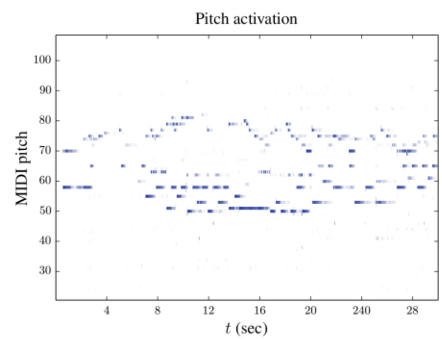
Spectrogram Decomposition (2)

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])

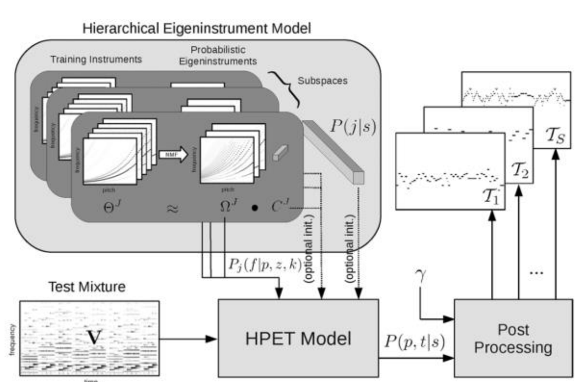




Spectrogram Decomposition (3)

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

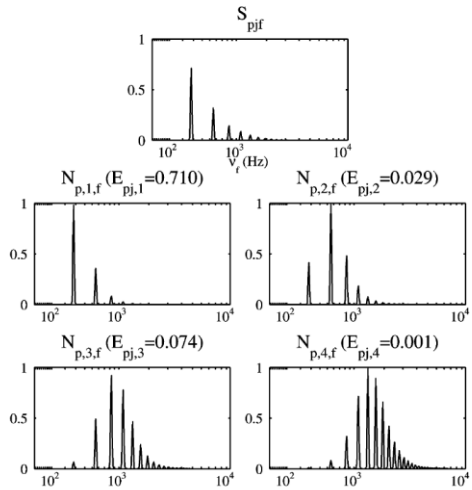


46

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…)

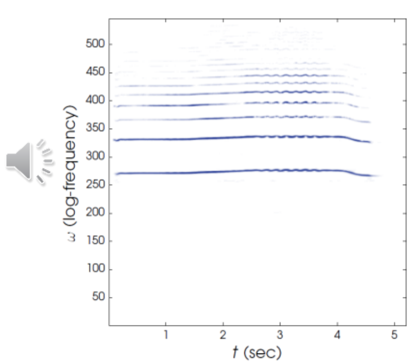


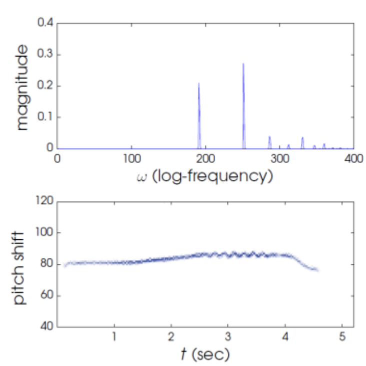
47

Spectrogram Decomposition (5)

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)



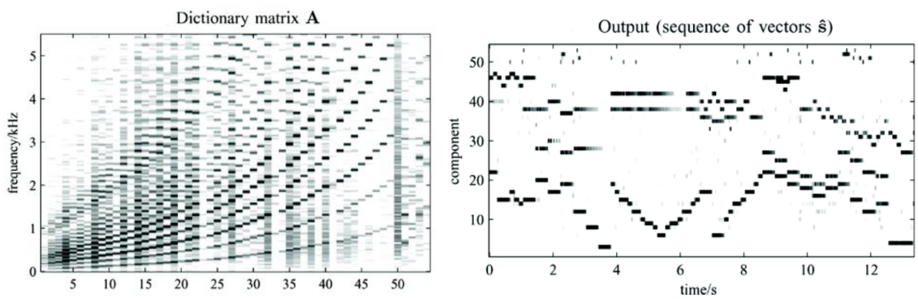


48

Spectrogram Decomposition (6)

Sparse coding

- Key concept: spectral templates are sparse; pitch activation is sparse for each time frame
- Sparse coding [Abdallah06]
- Group sparsity [O'Hanlon12]
- Pros: handling large dictionaries, computationally efficient methods
- Cons: little support on incorporating prior knowledge



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

50

Classification-based Methods (1)

[Marolt, 2004]

- 76 neural networks for piano notes (except for the lowest 12 notes)
- Input: output of partial tracking networks across multiple frames

neural network model	correct	spurious
time-delay NNs	96.8%	13.1%
Elman's NNs	95.2%	13.5%
multilayer perceptrons	96.4%	16.0%
RBF NNs	88.2%	14.6%
fuzzy-ARTMAP	84.1%	18.9%

- Combined with onset detection modules to achieve note-level transcription → SONIC

Classification-based Methods (2)

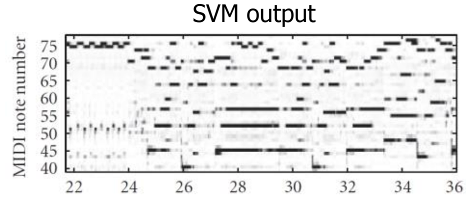
[Poliner & Ellis, 2007]

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

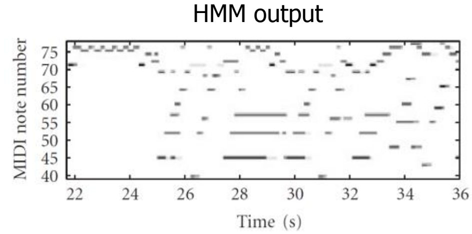
$$\begin{cases} 0\text{–}2 \text{ kHz, for notes} \le \text{B5 (988Hz)} \\ 1\text{–}3 \text{ kHz, for C6} \le \text{notes} \le \text{B6} \\ 2\text{–}4 \text{ kHz, for notes} \ge \text{C7 (2093Hz)} \end{cases}$$

- HMM smoothing for each class independently

SVM output



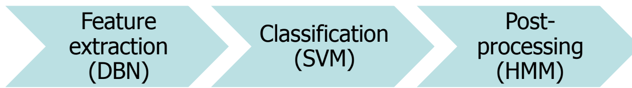
HMM output



Classification-based Methods (3)

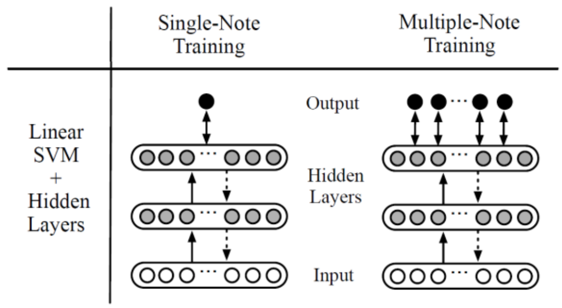
[Nam et al., 2011]

- Automatic feature learning by deep belief network (DBN)



2 hidden layers with 256 nodes each

Input: magnitude spectrum



Fine tune DBN weights for each note separately

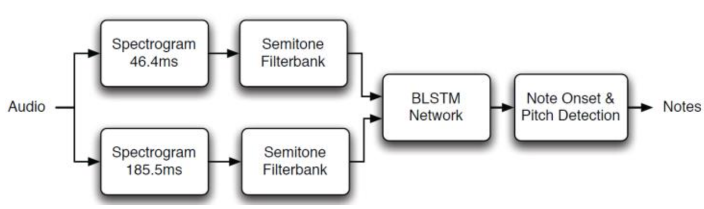
Fine tune DBN weights for all notes together

53

Classification-based Methods (4)

[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

54

Classification-based Methods (5)

[Raphael, 2002] for piano transcription

- Hidden Markov model (HMM)
  - States: note combinations
  - Observations: spectral features (energy, spectral flux, mean and variance of frequency distribution in each frequency band)
- Training: unsupervised training using piano audio and non-aligned MIDI scores (Baum-Welch algorithm)
  - Initialize states using score
  - Iteratively adjust model parameters and states
- Recognition: state space is huge, even after some pruning!
  - Restrict state space by multi-pitch estimation using observation model
  - Viterbi decoding

Pros: captures note transitions

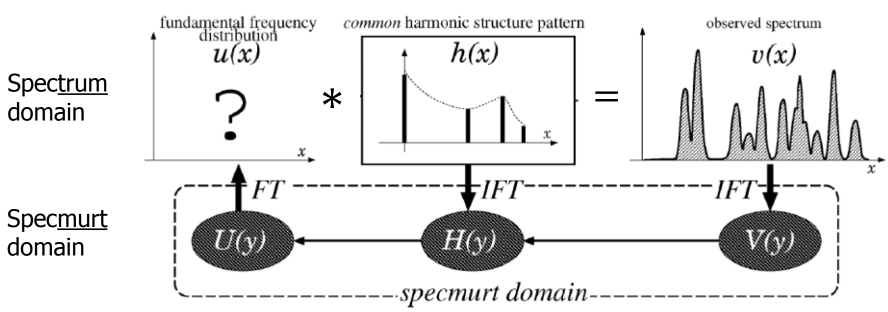
Cons: computationally expensive

55

Other Interesting Approaches

Specmurt Analysis: IFT of log-freq power spectrum [Saito et al., 2008]

- Assumes a common harmonic structure of all notes
- Iterative estimation of $u(x)$ and $h(x)$



- Harmonic structure is shared by all notes in the same frame, but not necessarily in different frames, in contrast to many other methods e.g., NMF methods

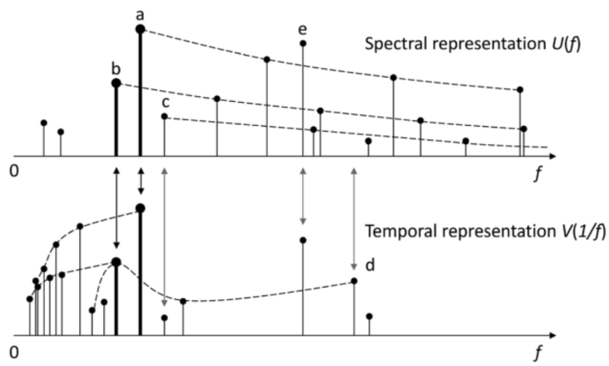
56

Other Interesting Approaches

- **Combining spectral and temporal representations** [Su & Yang, 2015]

Peaks in log-amplitude spectrum (harmonic errors)

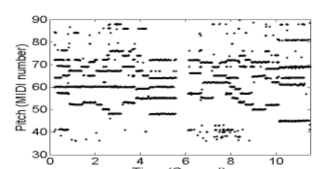
Peaks in autocorrelation function (subharmonic errors)



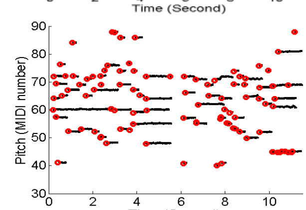
- Rules are designed to find F0s that have a prominent harmonic series in $U(f)$ and a prominent subharmonic series in $V(1/f)$

State of the Art

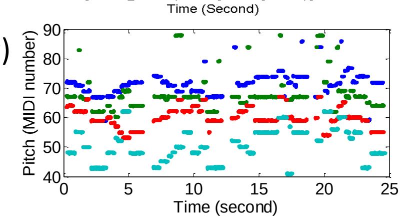
- Frame-level (multi-pitch estimation)
  - Estimate pitches and polyphony in each frame
  - Many methods



- Note-level (note tracking)
  - Estimate pitch, onset, offset of notes
  - Fewer methods



- Stream-level (multi-pitch streaming)
  - Stream pitches by sources
  - Very few methods



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

59

Frame Level → Note Level (1)

- Based on pitch salience/likelihood/activations
  - Thresholding, filling, pruning: [Bertin et al., 2010; Dessein et al., 2010; Carabias-Orti et al., 2011; Grindlay & Ellis, 2011; Böck & Schedl, 2012; Fuentes et al., 2013; Weninger et al., 2013]
  - Median filtering: [Su & Yang, 2015]

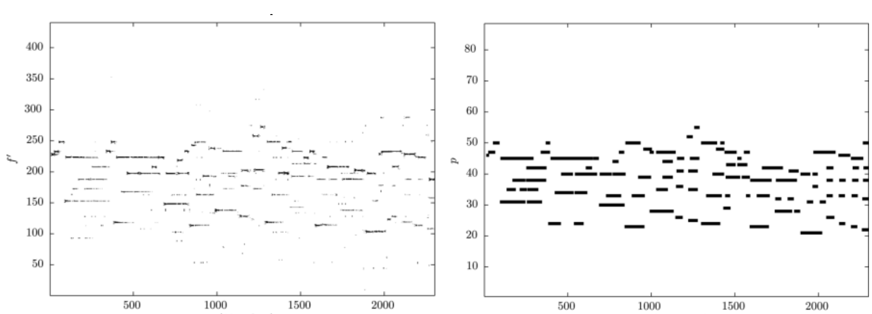
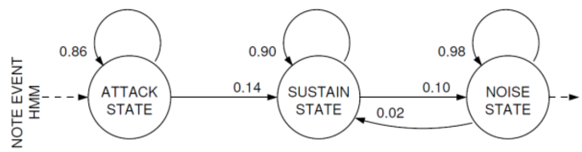


Figure from [Benetos & Dixon, 2013]

60

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

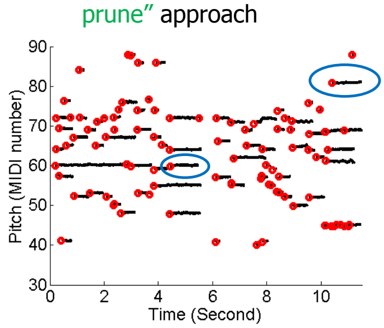
61

Frame Level → Note Level (3)

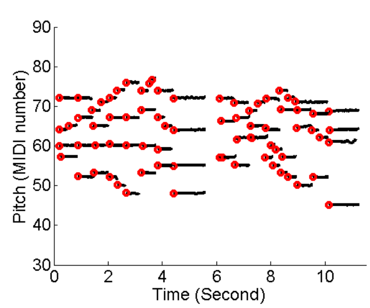
Problems of forming notes independently

- Contains many spurious notes caused by consistent MPE errors (usually octave/harmonic errors)
- Often violates instantaneous polyphony constraints

Results from the "connect-fill-prune" approach



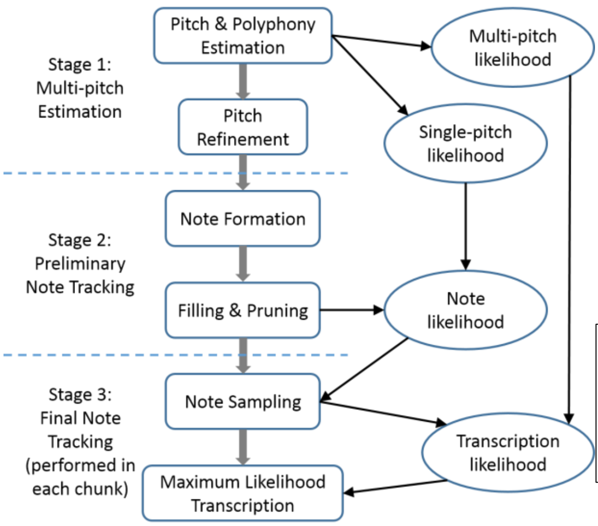
Ground-truth

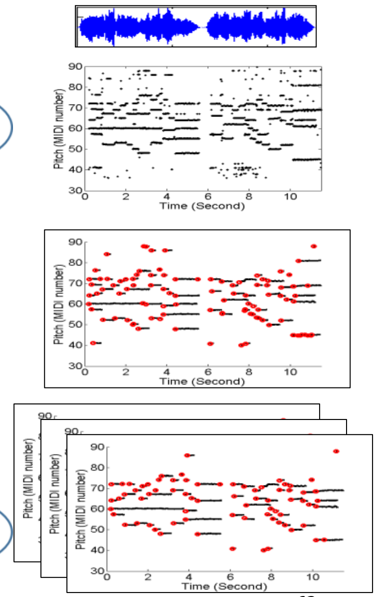


62

Frame Level → Note Level (4)

- **[Duan & Temperley, 2014] considering note interactions**



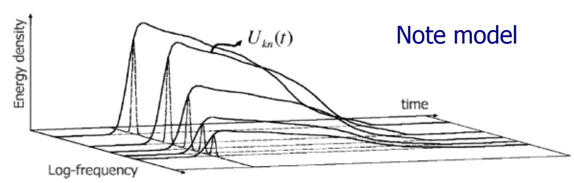


63

Note Tracking from Audio Directly (1)

[Kameoka et al., 2007]

- Harmonic temporal structured clustering (HTC)



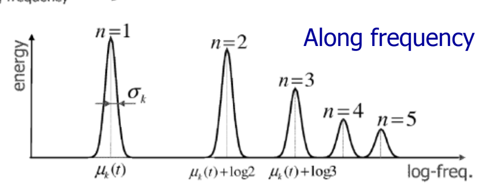
Mixture spectrogram

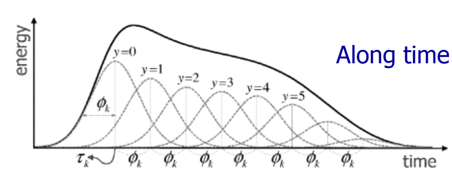
Activation of sources (latent variables)

$$\iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{q_k(x,t;\boldsymbol{\Theta})}\mathrm{d}x\mathrm{d}t$$

Source signal

parameters





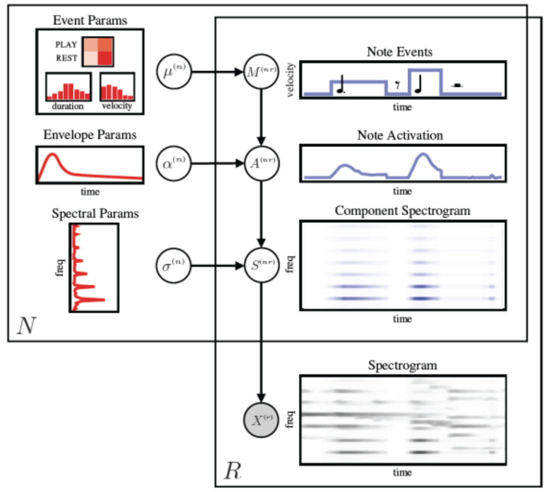
- EM algorithm

64

Note Tracking from Audio Directly (2)

[Berg-Kirkpatrick et al., 2014]

- An NMF-like approach for piano transcription
  - Each note is modeled by a spectral profile and an activation envelope
  - Duration and global velocity of activation envelope is generated from an HMM with two states (play and rest)
- Spectral profiles and activation envelopes are initialized using other pianos



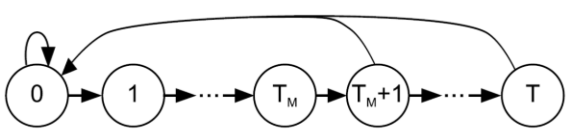
65

Note Tracking from Audio Directly (3)

[Ewert et al., 2015] for piano transcription

- Model each note as a series of log-freq magnitude spectra (states)

State space of a note



Silence

Minimun note length $= T_M$

$$\text{Mixture spectrum} = \sum_{88\text{ notes}} \text{spectrum(state)} * \text{activation}$$

unknown

- Too many state combinations!
- Greedy algorithm
  - Step 1: Estimate all state sequences for each note independent
  - Step 2: Decompose mixture spectrum into active notes to estimate activations

66

Note Tracking from Audio Directly (4)

[Cogliati et al., 2015] for piano transcription

- Time domain convolutional sparse coding

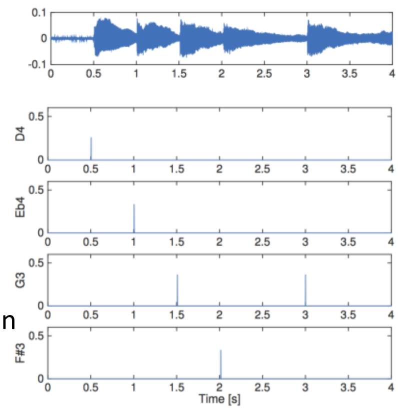
Note activation weights (i.e., the transcription)

Sparsity regularization

$$\arg\min_{\{x_m\}} \frac{1}{2} \left\| \sum_m \boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s} \right\|_2^2 + \lambda \sum_m \|\boldsymbol{x}_m\|_1$$

Note templates (pre-recorded)

Music signal to be transcribed

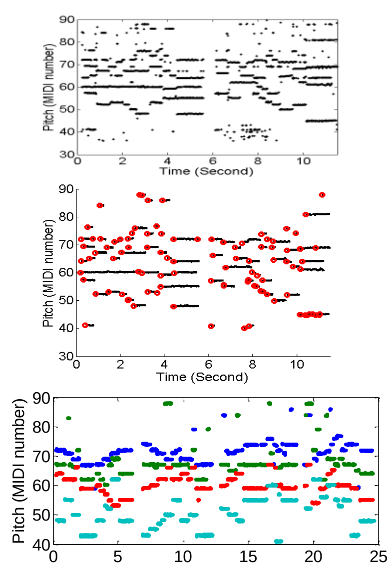


- Pros: high accuracy and onset precision
- Cons: piano/environment-dependent; doesn't estimate offset

67

State of the Art

- 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



68

Multi-pitch Streaming (Timbre Tracking)

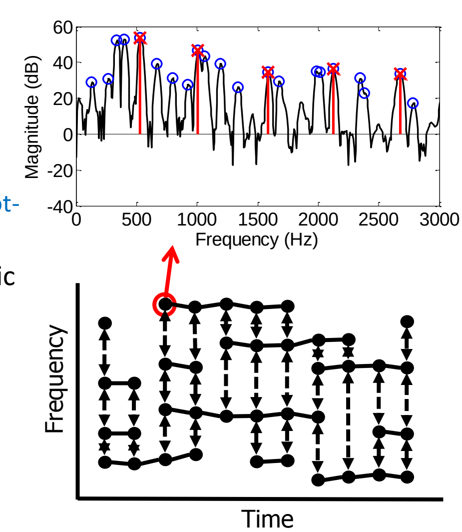
- Supervised
  - Train timbre models of sound sources
  - Apply timbre models during pitch estimation: [Cont et al., 2007; Bay et al., 2012; Benetos et al., 2013]
  - Classify estimated pitches/notes: [Wu et al. 2011]
- Supervised with timbre adaptation
  - Adapt trained timbre models to sources in mixture: [Carabias-Orti et al., 2011; Grindlay & Ellis, 2011]
- Unsupervised
  - Cluster pitch estimates according to timbre: [Duan et al., 2009, 2014; Mysore & Smaragdis, 2009; Arora & Behera, 2015]

69

Timbre Tracking – Unsupervised (1)

[Duan et al., 2009, 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

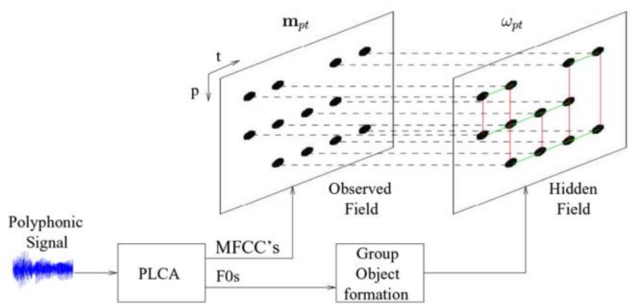


70

Timbre Tracking – Unsupervised (2)

[Arora & Behera, 2015]

- Constrained clustering
  - Objective: maximize timbre consistency within clusters
  - Constraints based on pitch locations: grouping constraints (i.e., pitch continuity) and simultaneity constraints (i.e., simultaneous pitches)
- Timbre representation: MFCC
- Clustering algorithm: hidden Markov random field



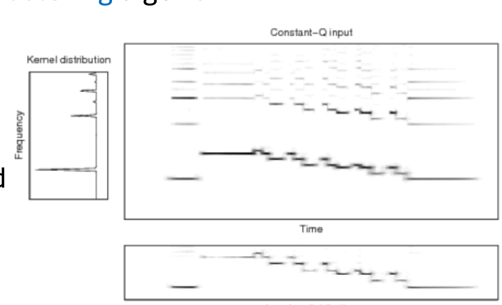
71

Timbre Tracking – Unsupervised (3)

[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



72

State-of-the-art:

Transcribing Percussive Instruments

73

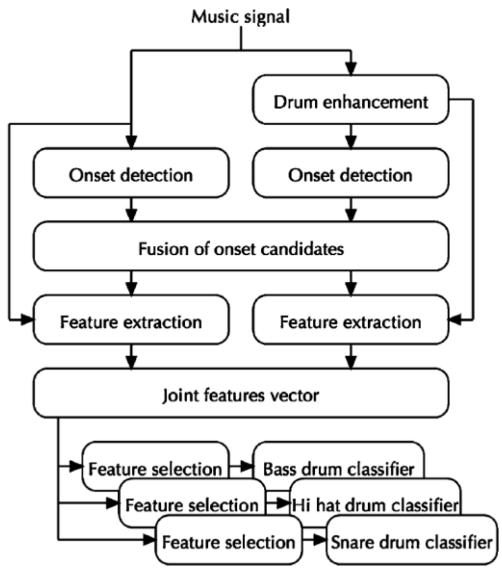
Percussive Instruments Transcription (1)

- **Core application**: transcribing drum kit sounds
- **Literature**:
  - Transcribing solo drums
  - Reducing percussive sounds for transcribing pitched sounds
  - Transcribing drums in the presence of pitched sounds
  - Transcribing drums & pitched sounds

74

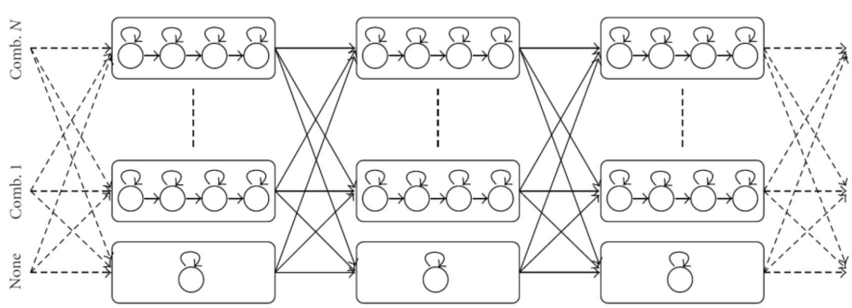
Percussive Instruments Transcription (2)

- **[Gillet and Richard, 2008]**: combines information from the original music signal and a drum track enhanced version obtained by source separation
- Large set of features (temporal, energy, spectral, perceptual…)
- Drum classification using C-support vector machines (C-SVM)
- Separation by harmonic/noise decomposition and time/frequency masking



Percussive Instruments Transcription (3)

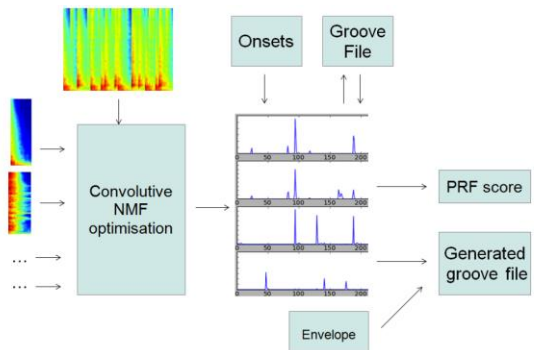
- **[Paulus and Klapuri, 2009]**: using a network of connected hidden Markov models (HMMs)
- HMMs are used to perform the segmentation and recognition jointly
- Features: MFCCs + temporal derivatives



Percussive Instruments Transcription (4)

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



77

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!)

78

State-of-the-art: Towards a Complete Music Notation

Towards a complete music notation (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:

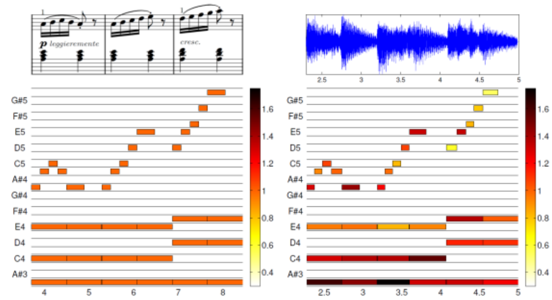
- Detect & integrate rhythmic information
- Detect tuning (per piece/note)
- Extract velocity per detected note
- Transcribe fingering (for specific instruments)
- Quantise pitches over time/beats

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

Towards a complete music notation (2)

Dynamics

- [Ewert11]: extracting note intensities in a score-informed scenario. Mapping with MIDI velocity information.
- [Kosta14]: Mapping between SPL and dynamic markings in the score
- Open problems:
  - Evaluation on intensity/velocity detection for AMT systems
  - Mapping between AMT intensities -> MIDI velocities -> dynamic markings
  - Datasets with audio + MIDI with velocity info + dynamic markings

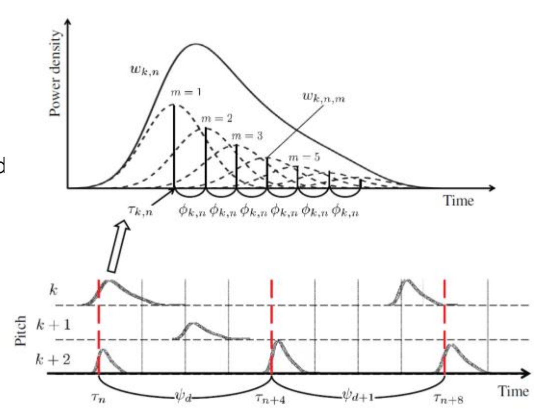


81

Towards a complete music notation (3)

Rhythm quantisation

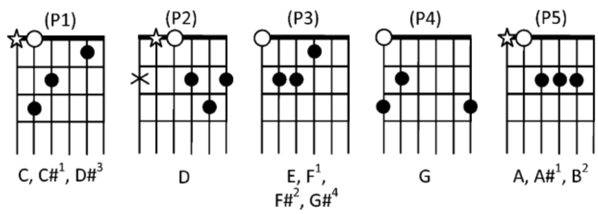
- [Collins14]: Combines multi-pitch detection with beat tracking for creating beat-quantized MIDI (goal: discovery of repeated themes).
- [Ochiai12]: Best structure modelling within an NMF-based multi-pitch detection system.
- Open problems:
  - Joint estimation of rhythmic structure and pitches
  - Exploit onset detection
  - Evaluation of beat-quantized outputs; comparison with scores?



Towards a complete music notation (4)

Fingering / string detection

- [Barbancho12]: extracting fingering configurations automatically from a recorded guitar performance (formulated as an HMM).
- [Maezawa12]: violin fingering transcription (formulated as a GMM-HMM)
- [Dittmar13]: real-time guitar string detection; feature extraction from multi-pitch pre-processing step & SVMs for classification.
- Open problems:
  - Instrument model adaptation
  - Joint estimation of fingerboard location and fingering
  - Integration into a general-purpose AMT system



83

Towards a complete music notation (5)

Computer Music Engraving / Typesetting

- 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:

84

Datasets

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. Polyphonic
  2. Melody/baseline
  3. Percussive
  4. Additional resources (e.g. chord annotations)

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

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)

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

89

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

90

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)

91

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. Score-informed piano transcription dataset

   http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/13
   - 7 Disklavier recordings that contain performance mistakes
   - MIDI ground truth for recordings + "correct" performances

92

Datasets (8)

Melody/baseline datasets

1. RWC database –popular/royalty-free/genre subsets

   https://staff.aist.go.jp/m.goto/RWC-MDB/
   - manual melody annotations for popular/royalty-free subsets
   - some popular/genre recordings also have aligned melody/bass annotations

Datasets (9)

Percussive transcription datasets

1. ENST-Drums

   http://www.tsi.telecom-paristech.fr/aao/en/software-and-database/
   8-channel recordings, 3 drummers, 75min, audiovisual content

2. 200 Drum Machines

   http://colinraffel.com/datasets/200DrumMachines.tar.gz
   Samples collected from 200 different drum machines

Datasets (9)

Percussive transcription datasets

3. DREANSS dataset

   http://mtg.upf.edu/download/datasets/dreanss
   - 22 multi-track excerpts (rock, reggae, metal…) with drum annotations

4. IDMT-SMT-Drums

   http://www.idmt.fraunhofer.de/en/business_units/smt/drums.html
   - 95 polyphonic drum set recordings (real + synthesized)

Datasets (10)

Additional datasets

1. KSN database

   http://hil.t.u-tokyo.ac.jp/software/KSN/
   - Functional harmony annotations for RWC classical files

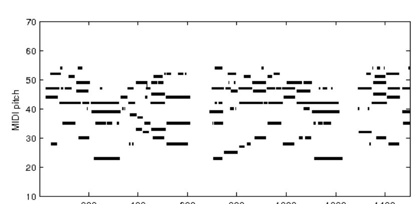
2. AIST RWC annotations

   https://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/
   - Beat/chorus annotations for RWC classical/jazz recordings

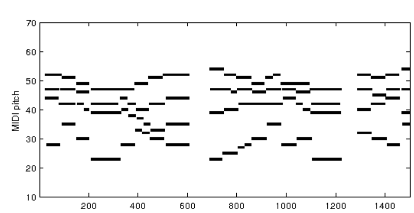
Evaluation Metrics

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

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

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

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

101

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

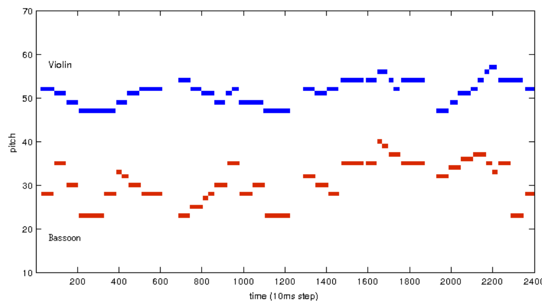
0.2100	0.8800	52.0000
0.2100	0.8000	47.0000
0.2200	0.7700	44.0000
0.2800	0.8300	28.0000
0.7800	1.4500	42.0000
0.8400	1.4500	47.0000
0.8900	1.4500	51.0000
0.8900	1.4500	35.0000
1.4800	2.0100	49.0000
1.4800	2.0100	46.0000

102

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



103

Public Evaluation

104

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 for 2015
    (piano solo, string quartet, piano quintet, violin sonata)

105

Public Evaluation (2)

MIREX Multi-F0 Estimation and Note Tracking task

- Results for Task 1 (frame-based accuracy)

Teams	2009	2010	2011	2012	2013	2014
Yeh and Roebel	0.69	0.69	0.68	-	-	-
Dressler	-	-	0.63	0.64	-	0.68
Canadas-Quesada et al.	-	0.49	-	-	-	-
Benetos and Dixon/Weyde	-	0.47	0.57	0.58	0.66	0.66
Duan et al.	0.57	0.55	-	-	-	-
Fuentes et al.	-	-	-	0.56	-	-
Elowsson and Friberg	-	-	-	-	-	0.72
Cheng et al.	-	-	-	-	0.62	-
Su and Yang	-	-	-	-	-	0.64

106

Public Evaluation (3)

MIREX Multi-F0 Estimation and Note Tracking task

- Results for Task 2 (onset/offset-based F-measure)

Teams	2009	2010	2011	2012	2013	2014
Yeh and Roebel	0.31	0.33	0.35	-	-	-
Dressler	-	-	-	0.45	-	0.44
Benetos and Dixon/Weyde	-	-	0.21	0.23	0.33	0.36
Duan, Han and Pardo	0.22	0.19	-	-	-	-
Fuentes et al.	-	-	-	0.39	-	-
Elowsson and Friberg	-	-	-	-	-	0.58
Cheng et al.	-	-	-	-	0.29	-
Su and Yang	-	-	-	-	-	0.29
Böck	-	-	-	0.09	-	0.14
Dessein et al.	-	0.24	-	-	-	-
Duan and Temperley	-	-	-	-	-	0.28

107

Public Evaluation (4)

MIREX Multi-F0 Estimation and Note Tracking task

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

Teams	2009	2010	2011	2012	2013	2014
Yeh and Roebel	0.50	0.53	0.56	-	-	-
Dressler	-	-	-	0.65	-	0.66
Benetos and Dixon/Weyde	-	-	0.45	0.43	0.55	0.58
Duan, Han and Pardo	0.43	0.41	-	-	-	-
Fuentes et al.	-	-	-	0.61	-	-
Elowsson and Friberg	-	-	-	-	-	0.82
Cheng et al.	-	-	-	-	0.50	-
Su and Yang	-	-	-	-	-	0.46
Böck	-	-	-	0.50	-	0.54
Dessein et al.	-	0.40	-	-	-	-
Duan and Temperley	-	-	-	-	-	0.45

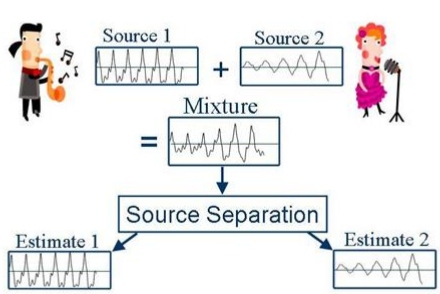
108

Relations & Applications to Other Problems

Relations to Other Problems (1)

Music Source Separation

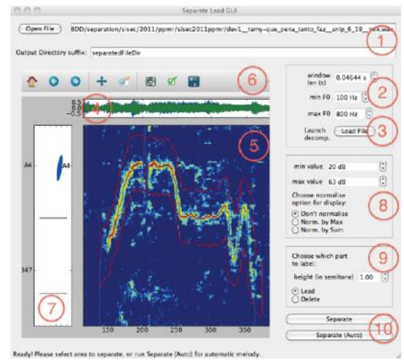
- Interdependent with multi-pitch detection and instrument identification
- Instrument identification can be improved by separating the source signals [Bosch12]
- Joint instrument identification and separation [Itoyama11]



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]

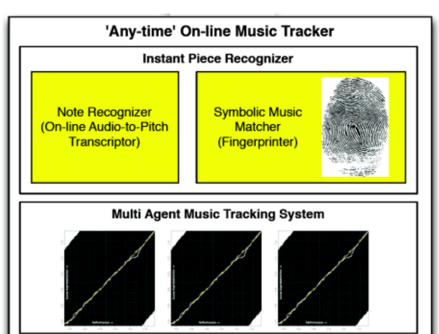


111

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)

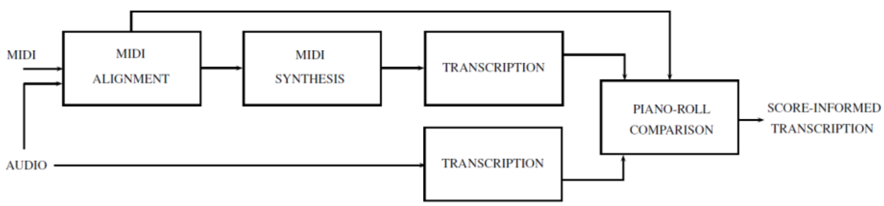


112

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)

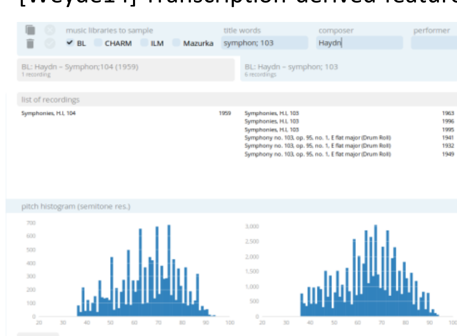


113

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



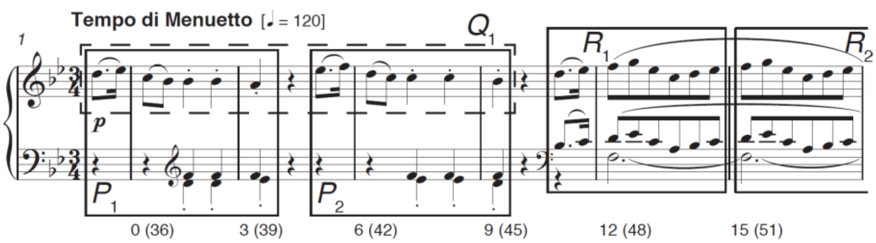
http://dml.city.ac.uk/vis/

114

Relations to Other Problems (6)

Applications to Systematic/Computational Musicology

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

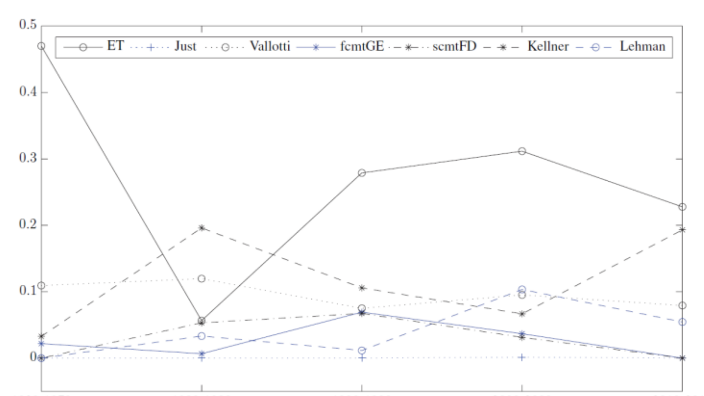


115

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.

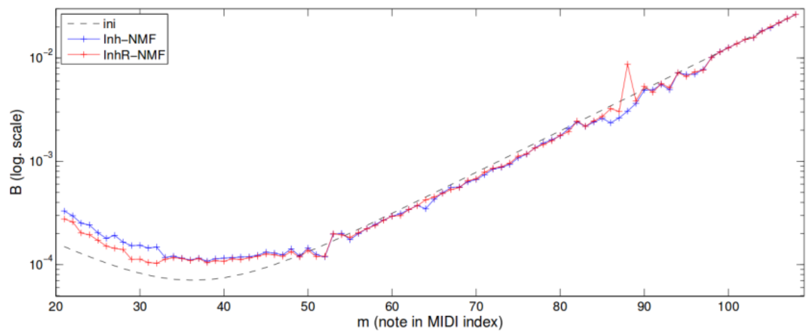


116

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

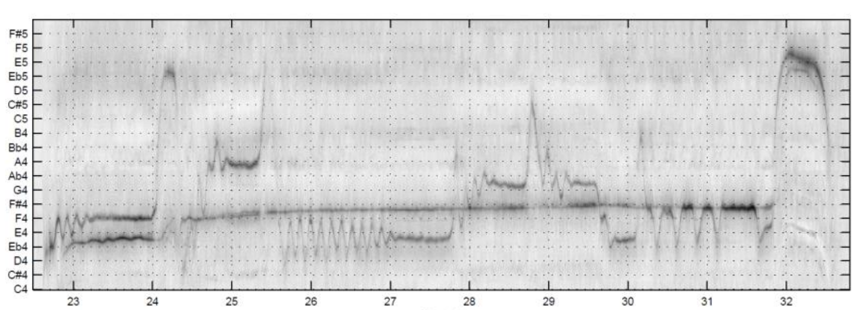


117

Relations to Other Problems (9)

Applications to Music Performance Analysis

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



118

Software & Demo

AMT Software (1)

Free software / plugins (from academic research)

Authors	Language	URL
Benetos et al	Matlab + Vamp plugin	http://www.eecs.qmul.ac.uk/~emmanouilb/code.html
Duan et al	Matlab	http://www.ece.rochester.edu/~zduan/resource/Resources.html
Fuentes et al	Matlab	http://www.benoit-fuentes.fr/publications.html
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
Zhou & Reiss	Vamp plugin	http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html

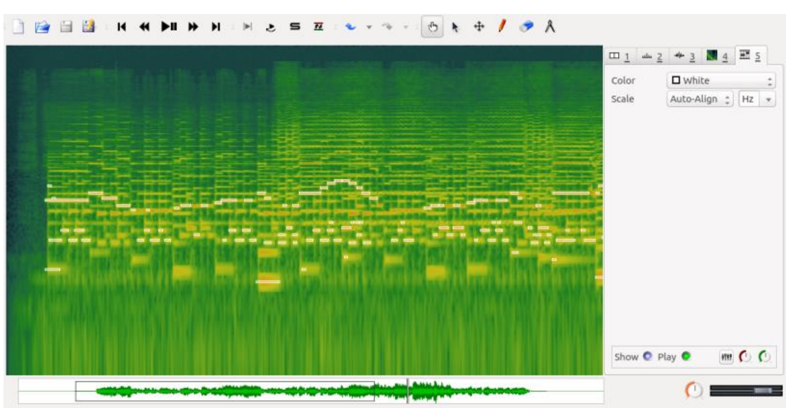
AMT Software (2)

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/
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

Demo

Silvet Vamp plugin



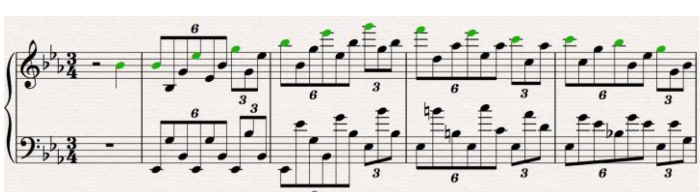
Silvet download: https://code.soundsoftware.ac.uk/projects/silvet/files
Sonic Visualiser download: http://www.sonicvisualiser.org/download.html

Challenges and Future Directions

Challenges and Directions – 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

Challenges and Directions – 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

Challenges and Directions – Musical Knowledge (1)

Incorporating musical knowledge

- Most existing transcription approaches are data-driven (bottom-up)
  - Caused many errors that are not musically meaningful, and hence may be easily avoided by incorporating musical knowledge
- Musicians rely on musical knowledge to transcribe music
  - Key signature, scale
  - Harmonic progression, metrical structure
  - Counterpoint and other composition rules
- Speech recognition successfully integrates *acoustic model* and *language model* through HMM or deep neural networks, 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

Challenges and Directions – Musical Knowledge (2)

Existing attempts in incorporating musical knowledge

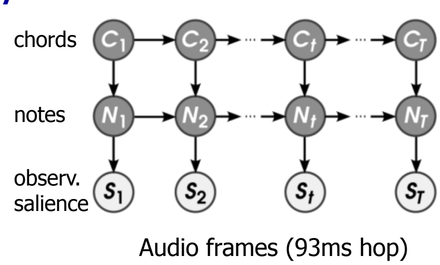
- Blackboard architecture [Martin96; Bello03]
  - Use of competing "knowledge sources"
  - No rigorous mathematical model
- Bayesian networks [Kasino98; Davy06; Cemgil06]
  - Rigorous mathematical models
  - Computationally intensive
  - Very simple musical knowledge (e.g., pitch range, pitch transition)
- More recent approaches

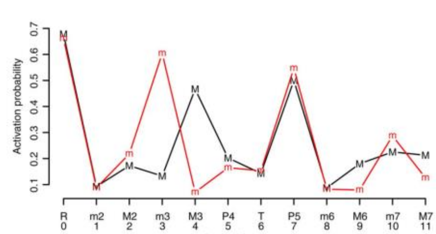
127

Challenges and Directions – Musical Knowledge (3)

[Raczynski et al., 2013] Dynamic Bayesian Networks

- Chord model: chords transition
- Note model: linear combination of the following sub-models:
  - *Harmonic*: pitch on/off based on underlying chord
  - *Duration*: pitch on/off transition
  - *Voice*: pitch jump
  - *Polyphony*: pitch on/off based on previous polyphony
  - *Neighbor*: pitch on/off based on the note directly below
- All models first-order Markovian
- 3% F-measure improvement from an NMF-based AMT approach





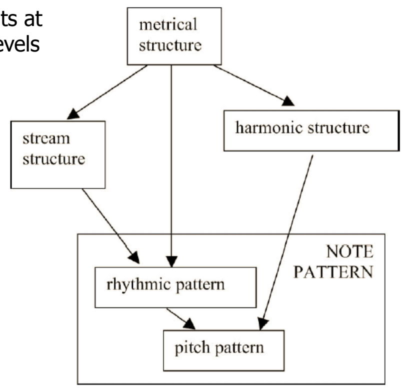
128

Challenges and Directions – Musical Knowledge (4)

[Temperley, 2009] Generative models for deep and interdependent musical structures

Meter, beats at different levels

Note tends to change on beats; note pitch jump; streams begin and end



Harmony tends to change on beats; chord progression

- Parameters are hand coded instead of learned from symbolic data
- Preliminary results (unpublished) show 3% improvement on note-level F-measure, using the acoustic model in [Duan & Temperley, 2014]

129

Challenges and Directions – Musical Knowledge (5)

Model temporal dependencies with RNN-RBM

- 1) Product of experts [Boulanger-Lewandowski12]

Combinations of the best pitch candidates estimated by the acoustic model

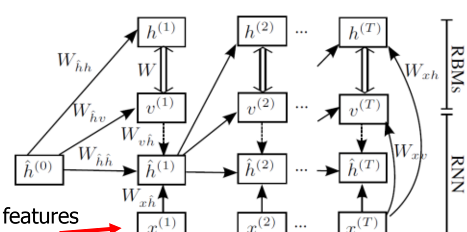
$$C = -\log P_a(v^{(t)}) - \alpha \log P_s(v^{(t)} | \tilde{A}^{(t)})$$

Acoustic model by RBM [Nam11]

Proposed symbolic model

- 2) Joint optimization by I/O RNN-RBM [Boulanger-Lewandowski13]

RBM output features using [Nam11]



Tests on mostly synthetic piano data
- Method 1 achieves 1%-10% frame-level transcription accuracy improvement
- Method 2 achieves 4%-30% improvement

130

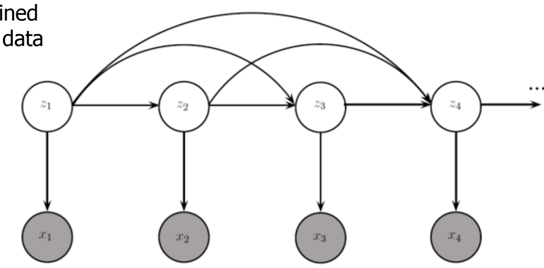
Challenges and Directions – Musical Knowledge (6)

Model music language using RNN

- PLCA + RNN-NADE [Sigtia et al., 2014]
  - RNN-NADE is a variant of RNN-RBM, taking a pitch activity vector sequence as input
  - Impose RNN as a Dirichlet prior for pitch activations into the PLCA framework
  - 3% frame-level transcription accuracy improvement on real data
- RNN + RNN [Sigtia et al., 2015]

RNN-NADE trained from symbolic data

DNN or RNN or DNN+RNN trained from acoustic data



131

Challenges and Directions – User Assisted Approach (1)

User-assisted (semi-automatic) music transcription

- What information is helpful and is easy to provide by users?
  - Key, tempo, time signature, structural information, timbre
- How to make the interaction easy for users to annotate?
  - Typing information
  - Editing through graphical user interface
  - Singing/humming melodic lines
  - Playing on a keyboard
- How to reduce the amount of information that users need to provide?
  - The system needs to learn from user annotations quickly and actively
  - An iterative approach is preferred

132

Challenges and Directions – User Assisted Approach (2)

Existing approaches

- Ask users to provide instrument labels for some notes to learn instrument models using shift-invariant NMF [Kirchhoff et al., 2012]
- Ask users to provide transcription of some segments of the piece to learn a PLCA-based model [Scatolini et al., 2015]

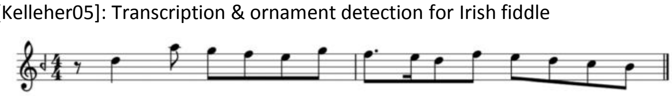
- In source separation
  - Singing voice / accompaniment separation through humming [Mysore & Smaragdis, 2009]
  - Music source separation with user-selected F0 track [Durrieu & Thiran, 2012]
  - Interactive Source Separation Editor with user selected spectrogram regions PLCA [Bryan et al., 2014]

133

Challenges and Directions – Non-Western (1)

Automatic transcription of non-Western/non-Eurogenetic/traditional music

- The vast majority of AMT research assumes 12 TET
- Another assumption: monophony/polyphony (whereas in several cultures music is **heterophonic**)
- 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

134

Challenges and Directions – Non-Western (2)

Automatic transcription of non-Western/non-Eurogenetic/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…)

135

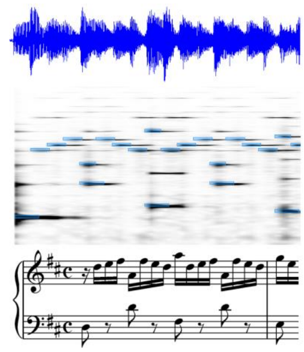
Conclusions

136

Conclusions (1)

State of the field

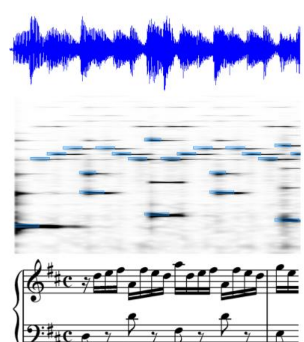
- Continues to attract attention in the MIR and music signal processing research communities + emerging topic for music language modelling
- Performance (objective + perceptual) has increased over the last decade
- Instrument- and style-specific AMT systems have sufficiently good performance for end-user 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



Thanks for listening!