Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

Complex audio feature extraction: Transcription

Anja Volk

Sound and Music Technology,
19 Dec, 2019

Outline

- What is transcription?
  - basic representations of musical content
- Application: query by Humming
- Audio and symbolic representations in folk song
- (multiple) F0 estimation
- Melody transcription
- Chord transcription

Recap

- Audio features for corpus analysis
  - Why do we undertake corpus analysis?
  - Examples: What makes a chorus? What makes a hook?
    - What type of audio features:
      - Psycho-acoustic audio features
      - Corpus-relative features

Recap

- Audio features for corpus analysis
  - Why do we undertake corpus analysis?
  - Examples: What makes a chorus? What makes a hook?
    - What type of audio features:
      - Psycho-acoustic audio features
      - Corpus-relative features

🔊 sharpness 🔊

Recap

- Audio features for corpus analysis
  - Why do we undertake corpus analysis?
  - Examples: What makes a chorus? What makes a hook?
    - What type of audio features:
      - Psycho-acoustic audio features
      - Corpus-relative features

🔊 roughness 🔊

Recap

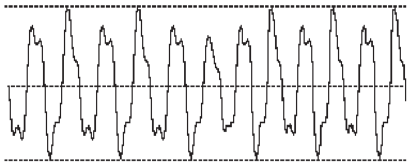
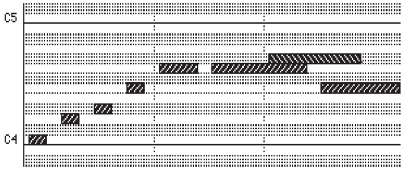
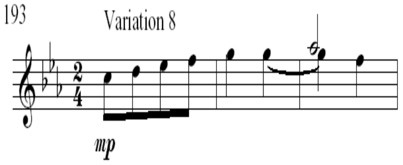
- Audio features for corpus analysis
  - Why do we undertake corpus analysis?
  - Examples: What makes a chorus? What makes a hook?
    - What type of audio features:
      - Psycho-acoustic audio features
      - Corpus-relative features

Recap

- Audio features for corpus analysis
  - Why do we undertake corpus analysis?
  - Examples: What makes a chorus? What makes a hook?
    - What type of audio features:
      - Psycho-acoustic audio features
      - Corpus-relative features
- Trends in the evolution for popular music and classical music (student presentations)
  - Features used: tonal and timbre descriptiors, e.g. chord progressions, tonal analysis
  - Always an important question: what can we conclude from a specific dataset?

Today: Transcription

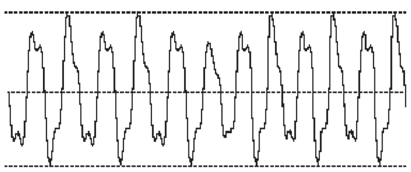
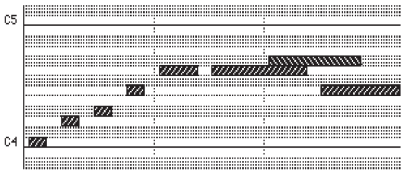
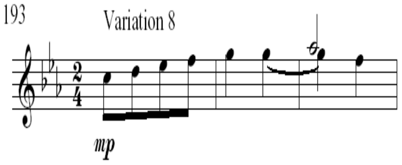
Basic representations of musical content

musical content	example	compare image	compare text	structure	convert to above	convert to below
Digital audio (MP3, Wav)		level 1: primitive features	speech	none	-	hard
Time-stamped events (MIDI)		level 2: objects	text	little	easy	fairly hard (OK job)
Music notation (Finale, Sibelius, MusicXML)		level 2: compound objects	text + markup	much	easy (OK job)	-

Transcription:

Basic representations of musical content

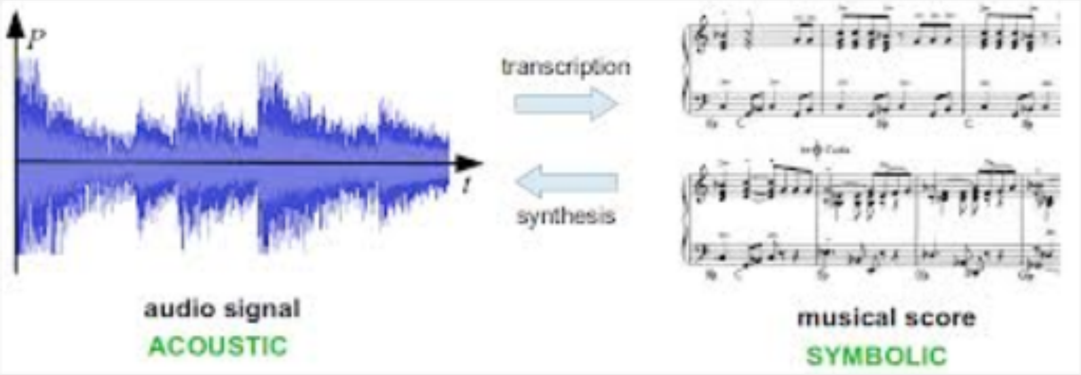
transcription

musical content	example	compare image	compare text	structure	convert to above	convert to below
Digital audio (MP3, Wav)		level 1: primitive features	speech	none	-	hard
Time-stamped events (MIDI)		level 2: objects	text	little	easy	fairly hard (OK job)
Music notation (Finale, Sibelius, MusicXML)		level 2: compound objects	text + markup	much	easy (OK job)	-

Audio transcription

- reconstruction of *sound events*, or even music notation, from audio signals
- sound events is what we perceive
  - *music notation* can be seen as an approximation of our perception
  - but *other symbolic representations* of sound events may be equally useful
- audio transcription problem is a major obstacle
  - lots of research
  - usable solutions exist for controlled situations

What does audio transcription involve?

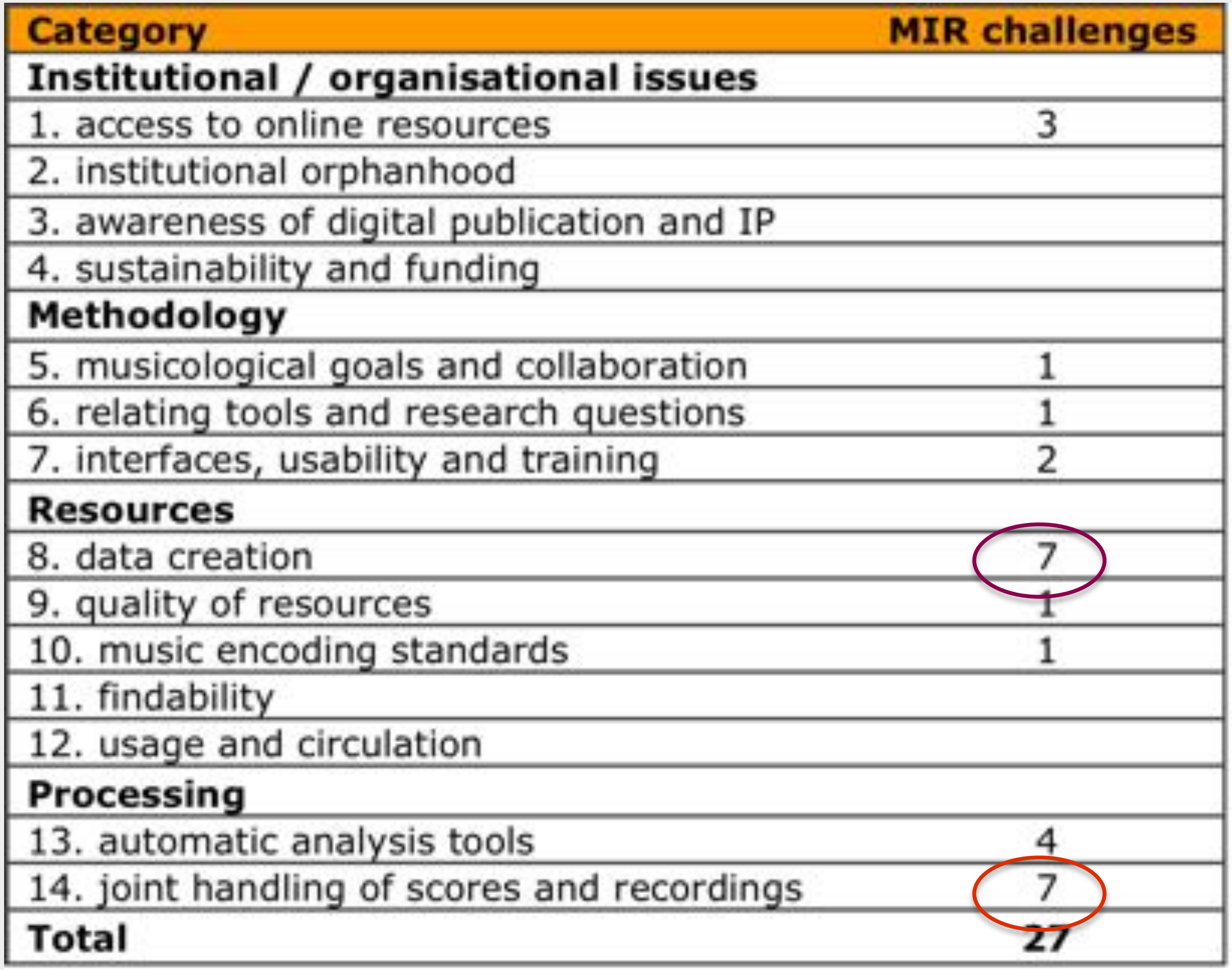


- many different tasks, often covered by MIREX
  - (multiple) F0 estimation
  - onset detection
  - melody extraction
  - audio chord estimation
  - instrument recognition
  - …

Applications

- Cover song detection
- Query by humming
- Audio notation alignment (score following)
  - Example: Score following https://www.youtube.com/watch?v=ZnpOYikF0qE
  - Example: PHENICX project on enhancing experience of concert audience: https://www.youtube.com/watch?v=7kxS8nblDYk
  - Performance analysis
- Investigating oral traditions
- not all of these require a full transcription

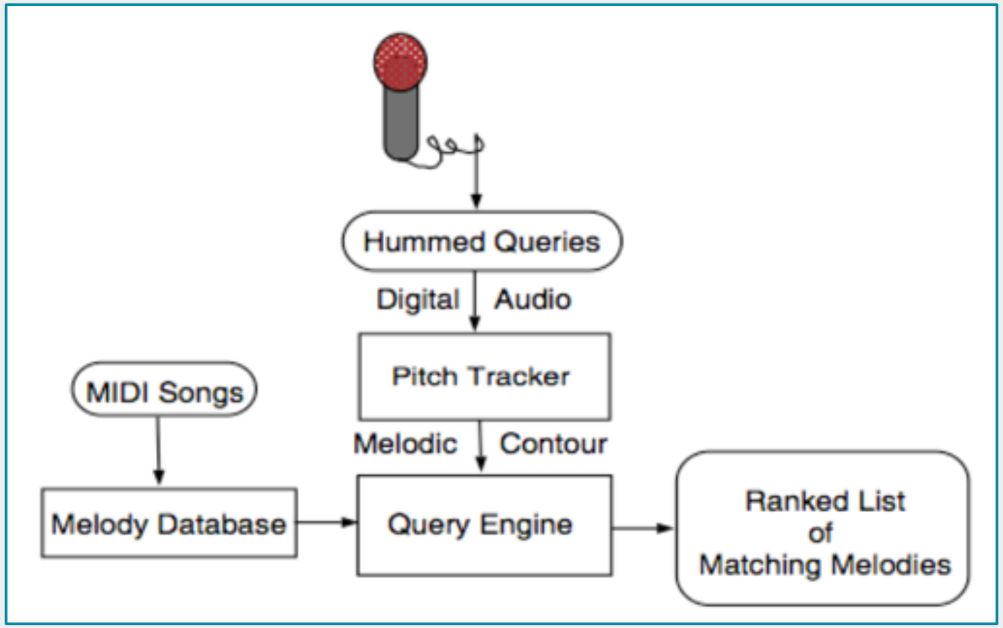
Questionnaire: musicology challenges for MIR



Category	MIR challenges
Institutional / organisational issues	
1. access to online resources	3
2. institutional orphanhood	
3. awareness of digital publication and IP	
4. sustainability and funding	
Methodology	
5. musicological goals and collaboration	1
6. relating tools and research questions	1
7. interfaces, usability and training	2
Resources	
8. data creation	7
9. quality of resources	1
10. music encoding standards	1
11. findability	
12. usage and circulation	
Processing	
13. automatic analysis tools	4
14. joint handling of scores and recordings	7
Total	**27**

Where it started…

- Ghias et al., *Query By Humming: Musical Information Retrieval in an Audio Database* (1995)



- actually, searching MIDI files
  - but the query is user-generated audio
  - matching after audio transcription

Query by Humming

- How was this done (Ghias et al. 1995)?
  - F0 estimation
    - human voice has peculiar resonance properties
    - formants (high frequency ranges that we perceive as vowels)
    - autocorrelation based method worked best
  - output turned into 3-letter alphabet a.k.a. Parson's code
    - U, D, S (up, down, same)
  - approximate string matching

- results
  - 'sequences of 10-12 pitch transitions were sufficient to discriminate 90% of the songs'
- high hopes for MIR on the Internet

QBH as a research paradigm

- attractive scenario with very strong impact in the next 10 years (and even up to now)
  - audio-to-symbolic matching
  - audio-to-audio matching
- how realistic is it?

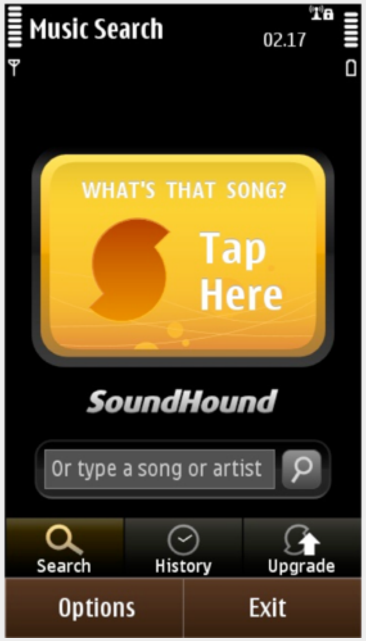
- MIREX Query By Singing/Humming (QBSH)
- run since 2006, led by Roger Jang
- several datasets, including Jang's MIR-SBSH corpus
  - 4431 audio queries
  - 48 MIDI ground truths + 2000 noise from Essen Folk Songs
  - evaluation: Mean Reciprocal Rank
  - scores up to 0.9595 (overfitting?)

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}.$$

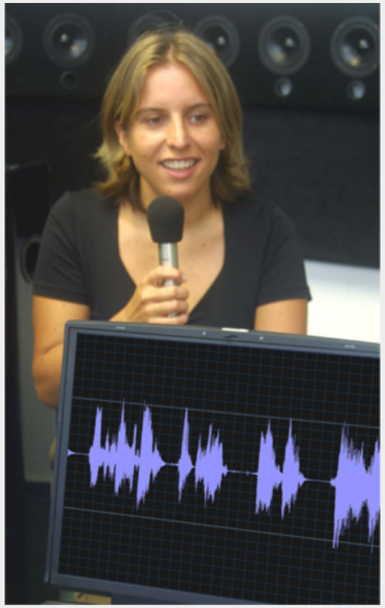
SoundHound

- currently, the most visible commercial application of QBH
- https://www.youtube.com/watch?v=8guE5sveSbM#t=1m27s

- SoundHound documentation doesn’t explain much
  - vague ‘compact and flexible Crystal representation’
  - ‘matches multiple aspects of the user's rendition (including melody, rhythm, and lyrics) with millions of user recordings from midomi.com)
- Midomi database consists of monophonic melodies
  - linked to metadata
  - unclear whether and to what extent melodies are transcribed
  - matching is monophonic
  - http://www.midomi.com/



Human performance in QBH



- constraints on QBH input
  - sound production (i.e. singing)
  - melody
  - accuracy
  - usually, beginning
- PhD research Micheline Lesaffre (Ghent U.)
  - *Music Information Retrieval: Conceptual Framework, Annotation and User Behavior* (2005)
  - experiment: what musical queries do people create when given complete freedom?
  - range is *large*

Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

F0 estimation

- determining the perceptual pitch from the acoustic signal

- the monophonic case is often considered solved
- see e.g. high scores for QBSH
  - query by singing humming
- no separate MIREX task

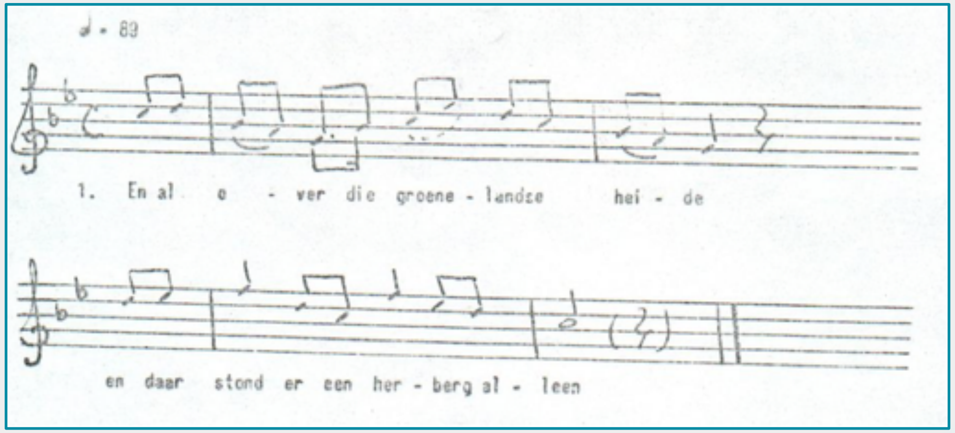
- in practice, pitch transcription is far from easy
  - signal vs. perception, octave errors
  - recording quality
  - musical competence of performers

Example challenge: Dutch folk songs

- lots of songs were just too hard to be transcribed reliably
  - song OGL 19401 is just an average case

beginning end

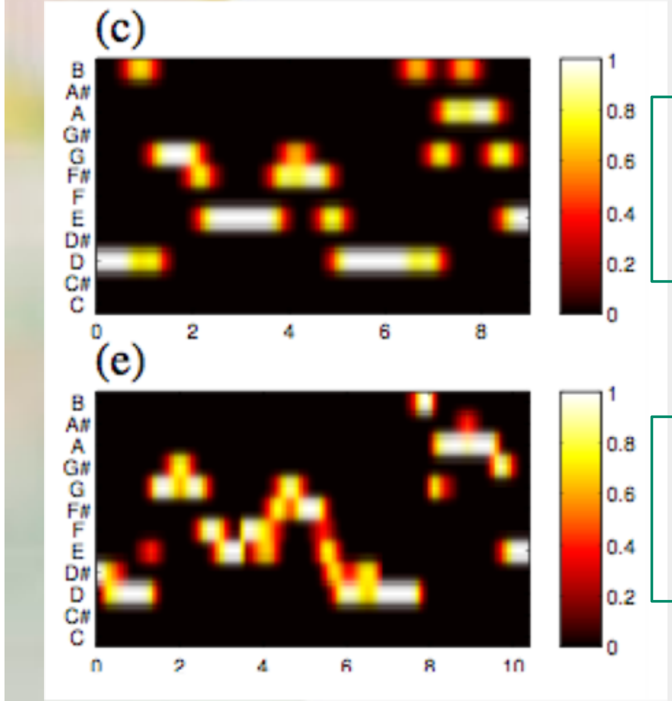
- that’s why we chose to use the existing paper transcriptions and encode these instead


♩ = 89
1. En al o - ver die groene - landse hei - de
en daar stond er een her - berg al - leen

symbolic-audio alignment

- Meertens Tune Collections
  - audio AND encoding AND notation AND metadata AND annotations
  - useful e.g. as ground truth for F0 estimation

- example: research Müller, Grosche & Wiering 2009
  - MIDI is 1 strophe, audio contains many strophes
  - matching might produce segmentation



chroma from MIDI

chroma from audio

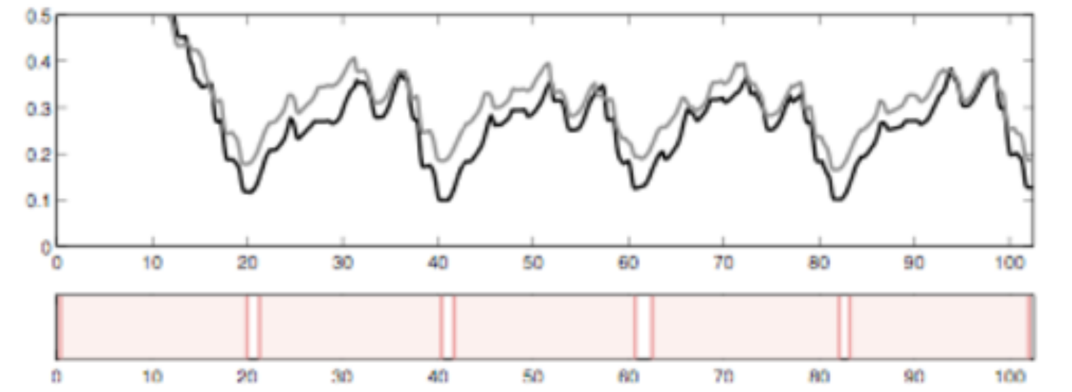


Figure 3. Top: Distance function Δ for NLB73626 using original chroma features (gray) and F0-enhanced chroma features (black). **Bottom:** Resulting segmentation.

Melody transcription

2 steps:

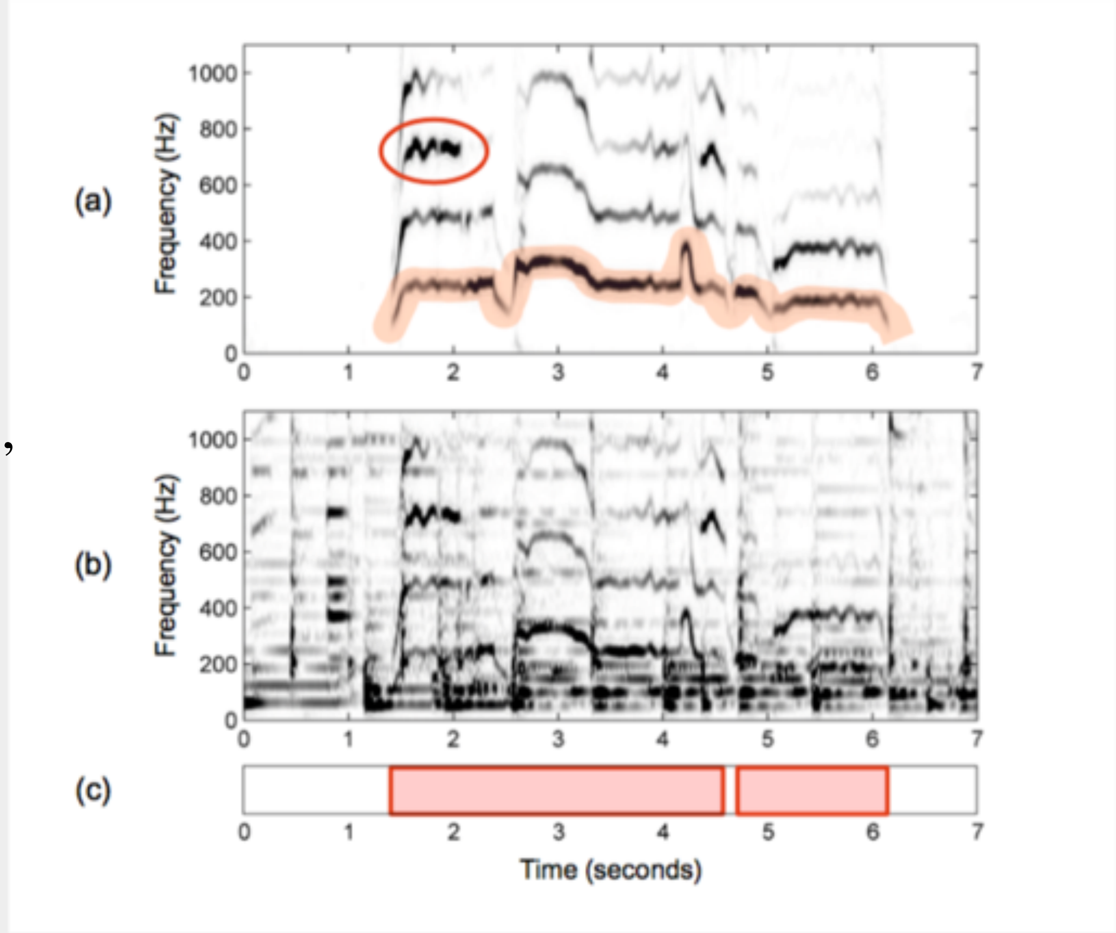
(1) estimate when the melody is present and when it is not
(2) estimate the correct pitch of the melody when it is present

- Considered as a still unsolved problem
  - Issue 1: separating components of polyphonic mixture of complex sounds very difficult
  - Issue 2: estimating pitch from the separated stream is still a challenge in itself
  - Issue 3: determining note boundaries challenging for several instruments, including voice (onset and transition blurred)

Melody transcription

- Example

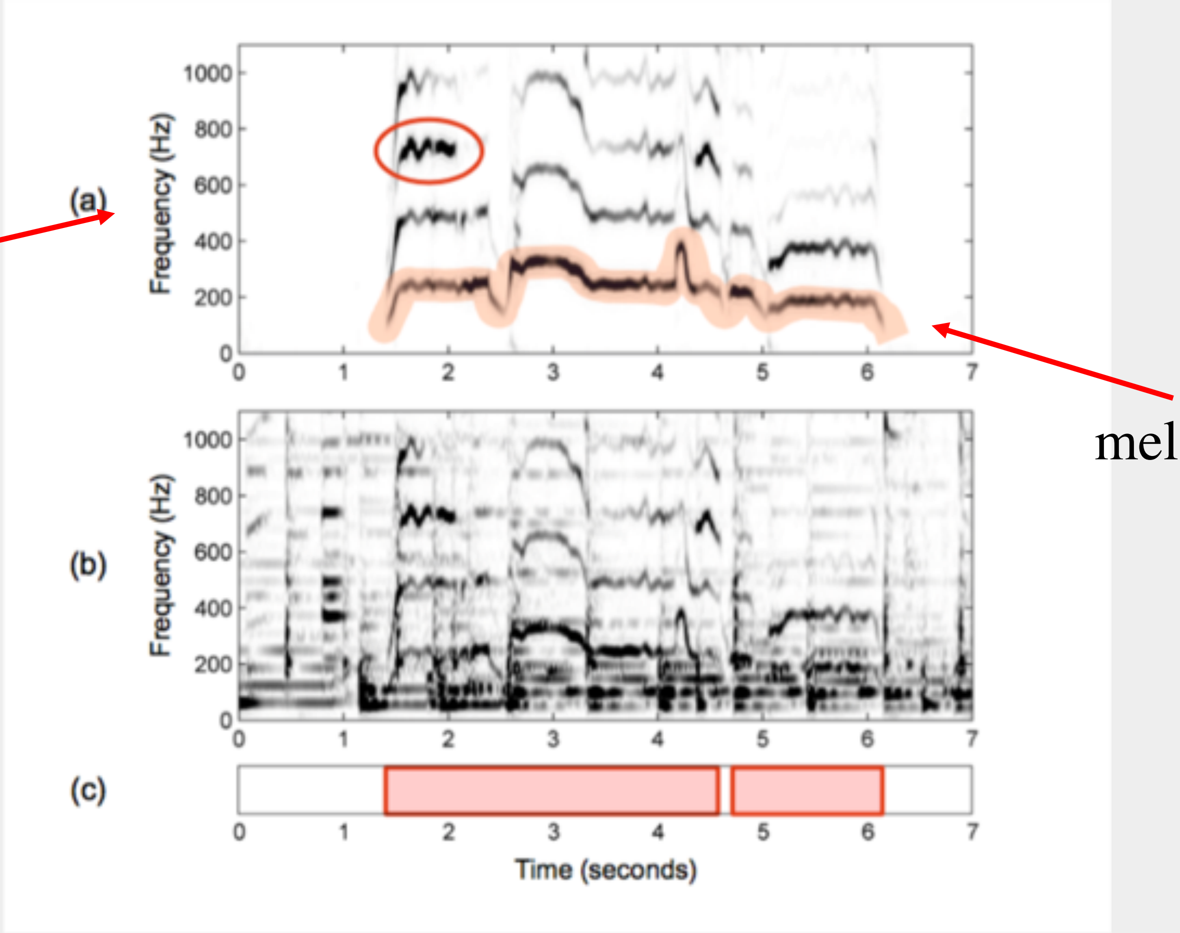
a male singer (melody)
accompaniment (drums, piano, guitar)



M. Mueller: Fundamentals of Music Processing P. 434

Melody transcription

- Example



Spectrogram isolated singing voice

melody

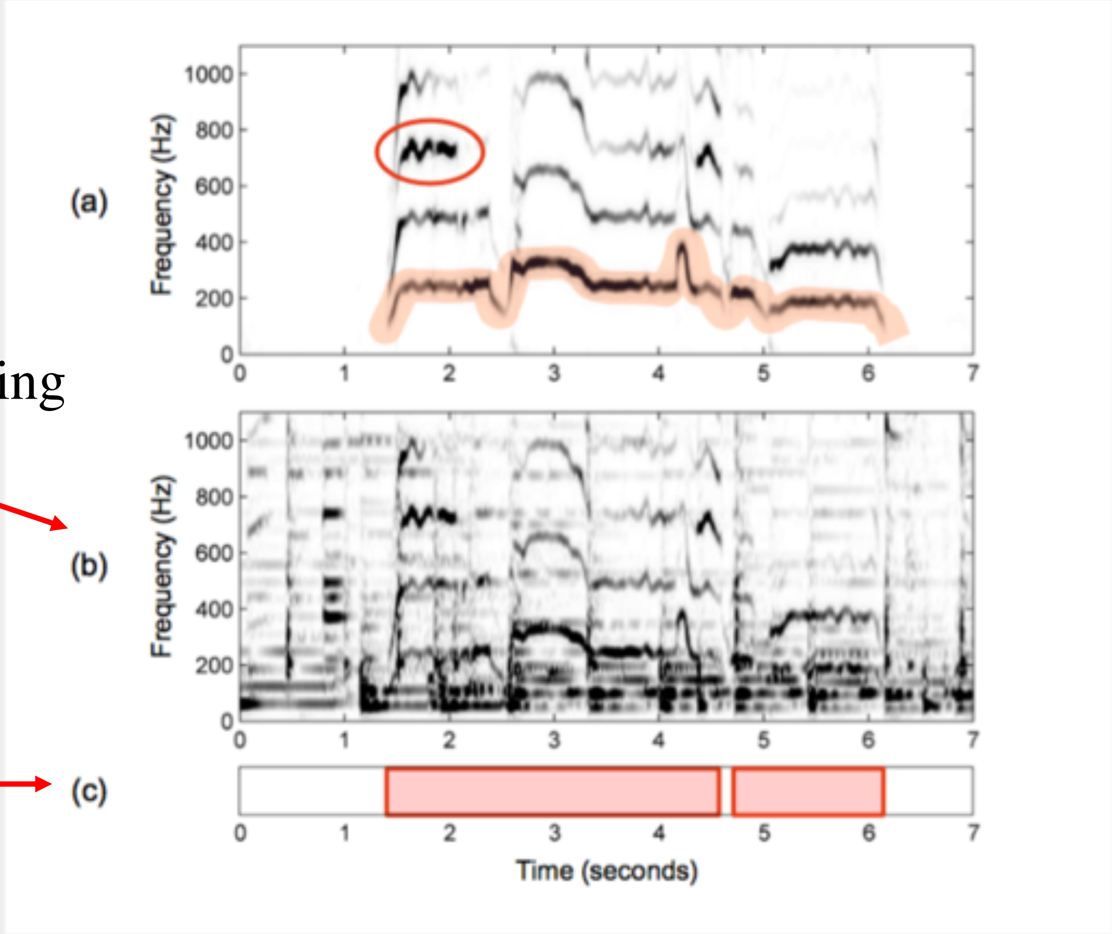
M. Mueller: Fundamentals of Music Processing P. 434

Melody transcription

- Example

Spectrogram full recording

Singer is active



M. Mueller: Fundamentals of Music Processing P. 434

Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

Melody transcription

- YIN
  - Cheveigne & Kawahara (2001)
  - pitch-estimation algorithm operating in the time domain (no Fourier transformation)
  - Works with autocorrelation function (ACF) of the signal

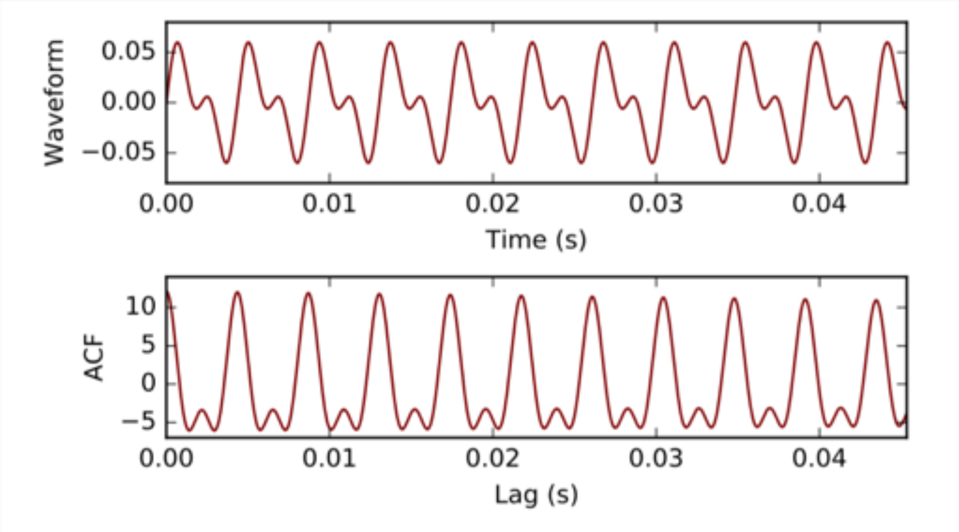
$$\mathrm{ACF}(Y)(\tau) = \sum_t y(t)\, y(t-\tau)$$

$$y(t) = y(t+T) \quad \forall T$$

$$\mathrm{ACF}(Y)(T) = \sum_t y(t)\, y(t-T) = \sum_t y(t)^2.$$

Melody transcription

- YIN
  - Cheveigne & Kawahara (2001)
  - pitch-estimation algorithm operating in the time domain (no Fourier transformation)
  - Works with autocorrelation function (ACF) of the signal

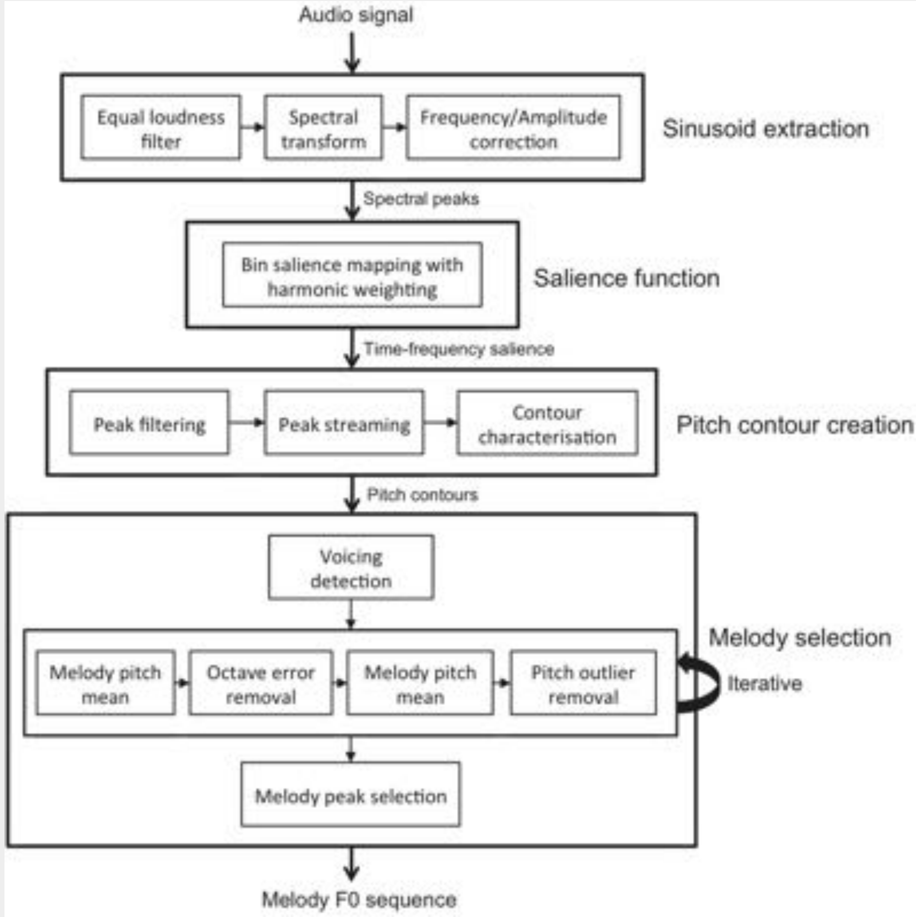


Melody transcription

- YIN
  - Cheveigne & Kawahara (2001)
  - pitch-estimation algorithm operating in the time domain (no Fourier transformation)
  - Works with autocorrelation function (ACF) of the signal
  - the signals on which YIN is typically used are **largely monophonic**, with only one, or one very dominant pitch present

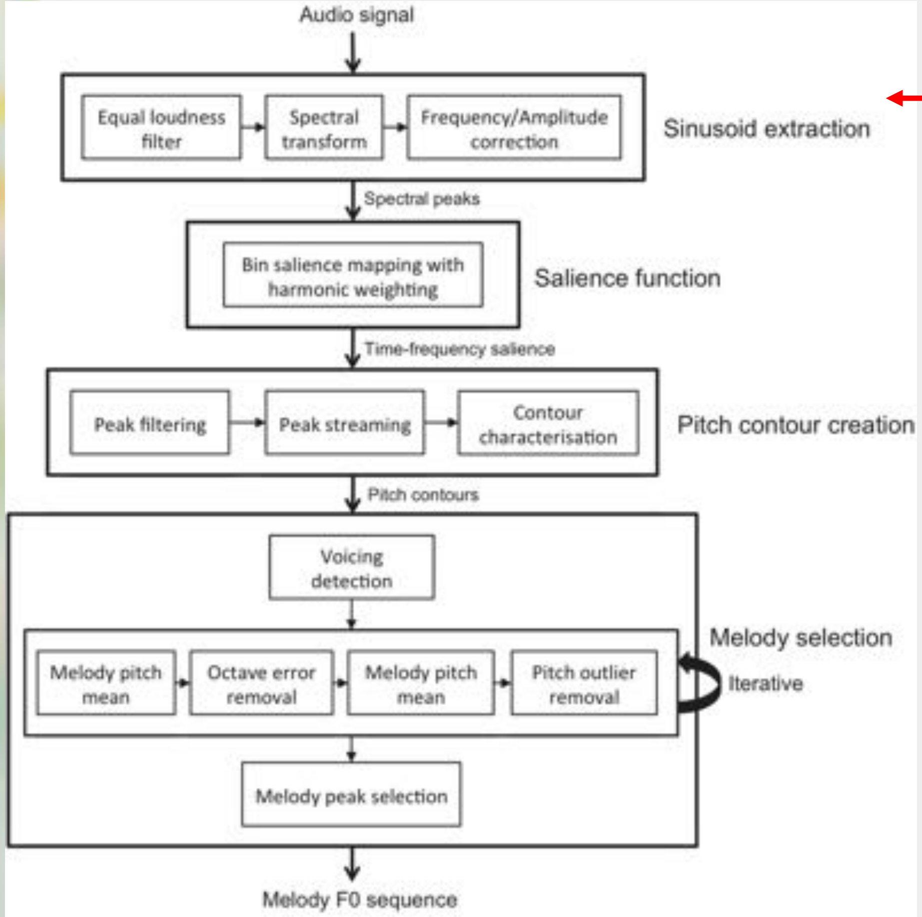
Melody transcription

- Melodia:
  - Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.



Melody transcription

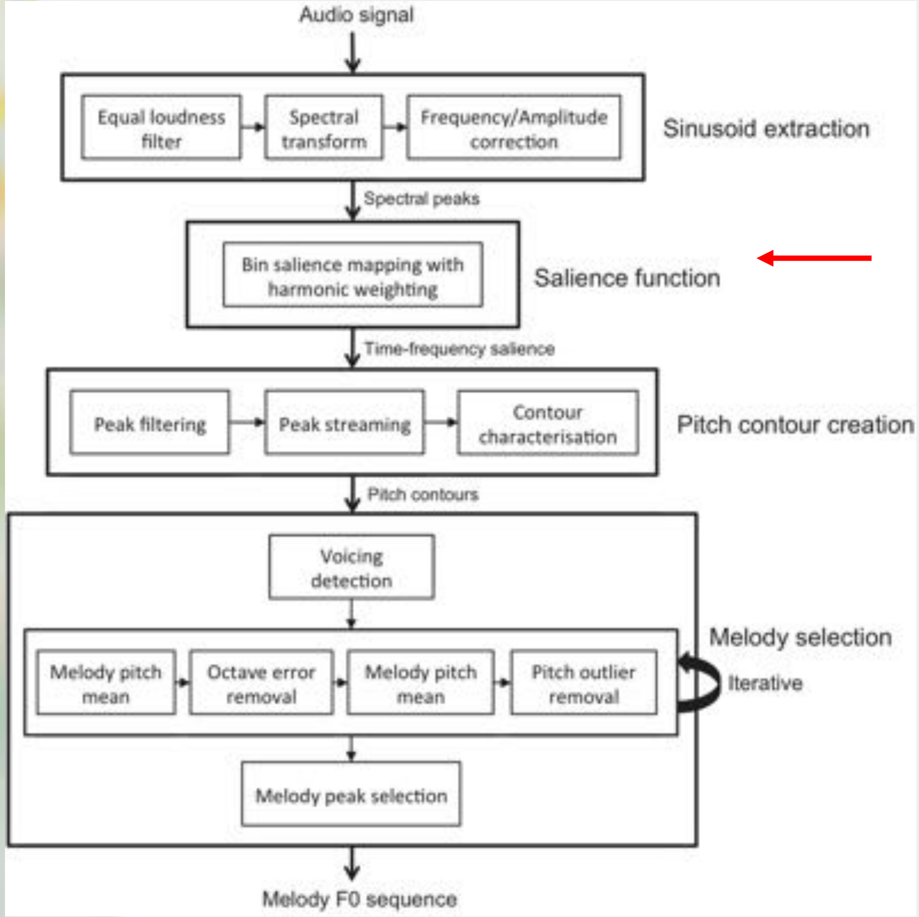
- Melodia:
  - Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.



Which frequencies are present in audio signal at every point in time?

Melody transcription

- Melodia:
  - Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.

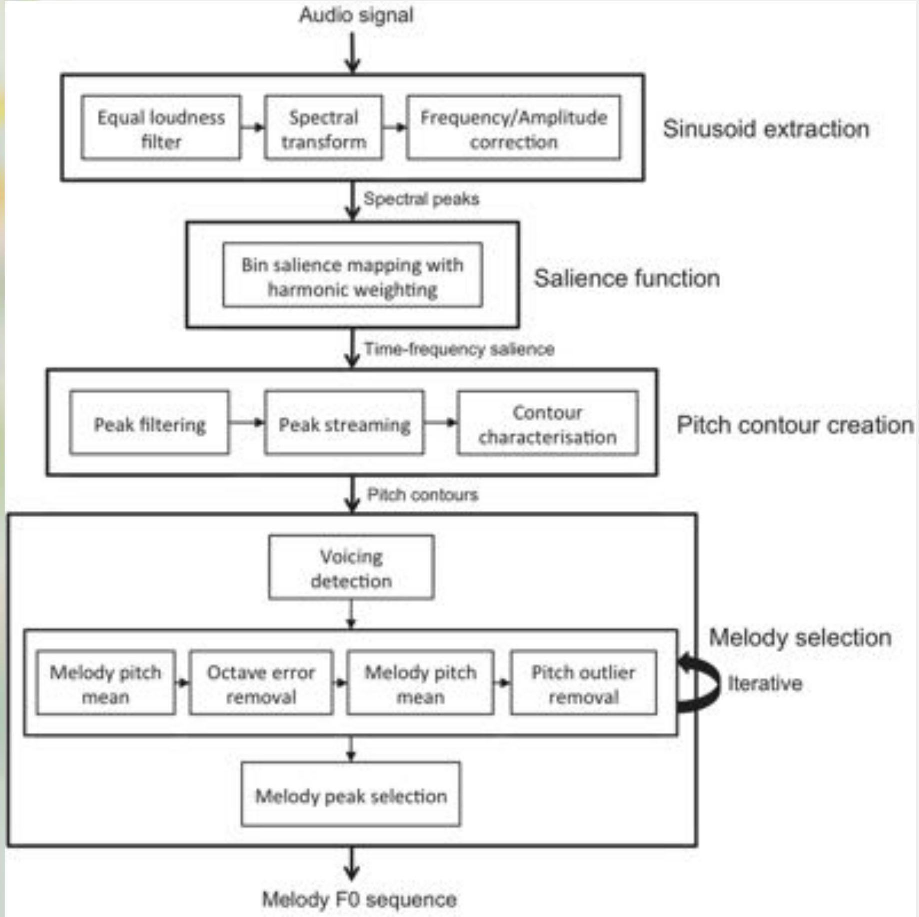


search for a harmonic series of frequencies that would contribute to our perception of this pitch

Salience = (weighted) sum of energies of these harmonic frequencies

Melody transcription

- Melodia:
  - Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.



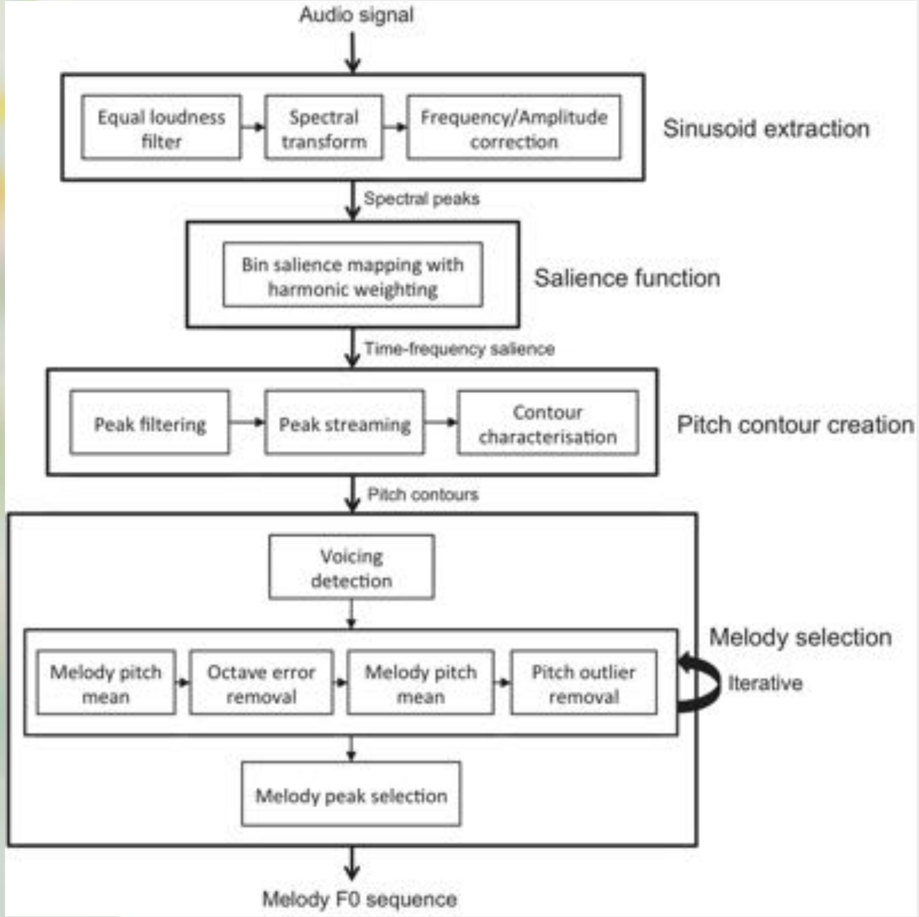
←

pitch contour = a series of consecutive pitch values which are continuous in both time and frequency

[Faculty of Science
Information and Computing Sciences]

Melody transcription

- Melodia:
  - Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.



By studying the distribution of these characteristics for contours that belong to melodies and contours that belong to accompaniments, we were able to devise a set of rules for filtering out non-melodic contours!

← calculation of contour characteristics.

Melody transcription

- Melodia
  - main melody extraction from polyphonic signals
  - start from a frequency representation of the signal, and assess, the salience of all possible candidate pitches
  - based on a high-resolution STFT, from which peaks are found
  - Estimate of exact location: instantaneous frequency is found by considering the magnitudes of the spectrum in each frame, and interpolating between the phases of peaks in consecutive frames of the STFT

Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteris- tics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.

Melody transcription

- Melodia
  - Then: harmonic summarization on the set of peaks
  - peak candidates are grouped in time-varying melodic contours using a set of heuristics based on perceptual streaming cues
  - candidate contours are then given a score based on their total salience and shape, and post-processed
  - selects the set of contours that most likely constitutes the melody.

Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteris- tics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.

Melody transcription

- Melodia
  - Demo:

Justin Salamon and Emilia Gomez. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteris- tics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2010.

http://www.justinsalamon.com/melody-extraction.html#demo

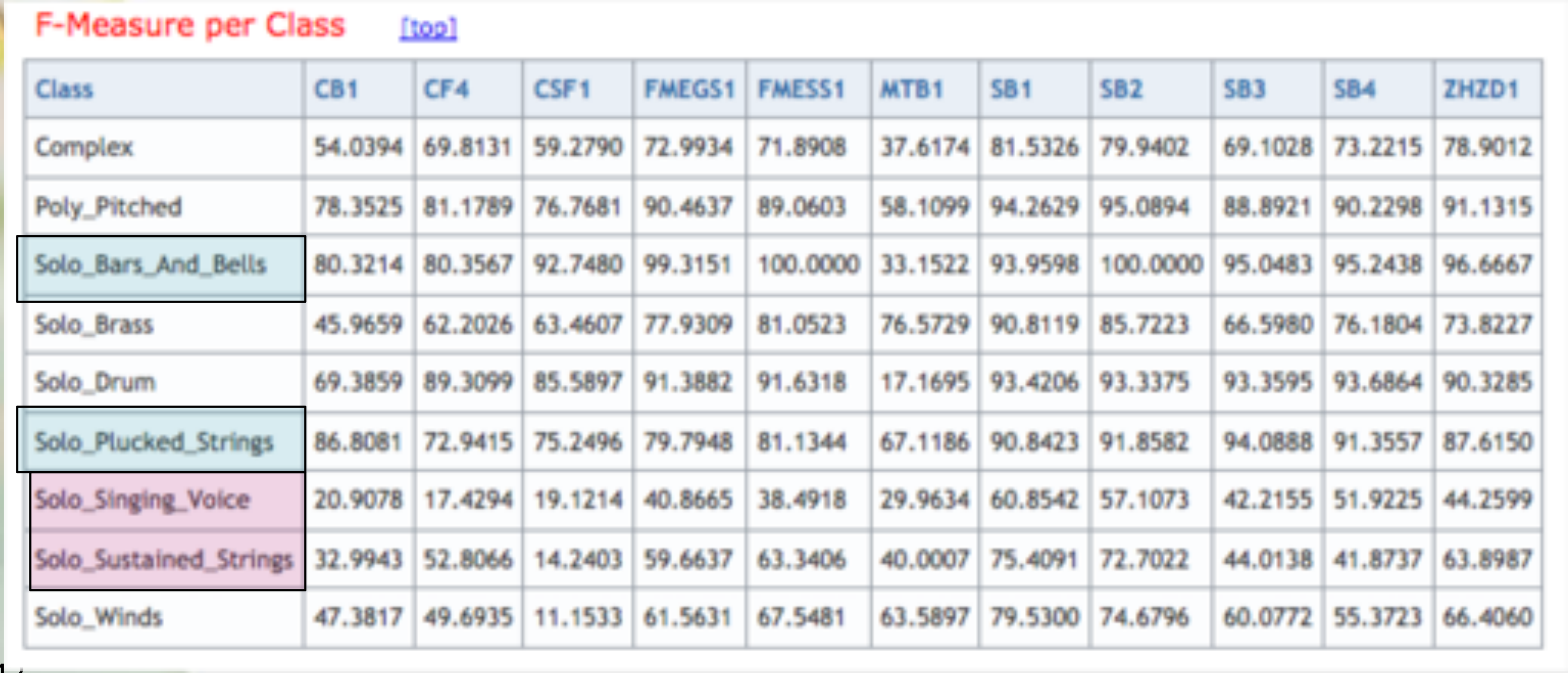
Melody transcription

- Data-driven and source separation-based systems
  - extract the melody by separating it from the rest of the mix
  - E.g.: use a trained timbre model to describe each of two sources (melody and accompaniment)
    - timbre models can be Gaussian mixture models (GMM's), in which each source is seen as a weighted sum of a finite set of multidimensional Gaussians, each describing a particular spectral shape, or hidden Markov models (HMM), a generalisation of GMM's
  - Models are trained on on source-separated ground truth data using expectation maximization

Graham E. Poliner and Dan Ellis. A Classification Approach to Music Transcription. *Proc. 6th International Society for Music Information Retrieval Conference*, 2005.

The matter of onset detection

- most transcriptions so far present a pitch contour
  - onset-offset detection missing
- Audio onset detection is (again) MIREX task
  - difficulty of subtasks is very different

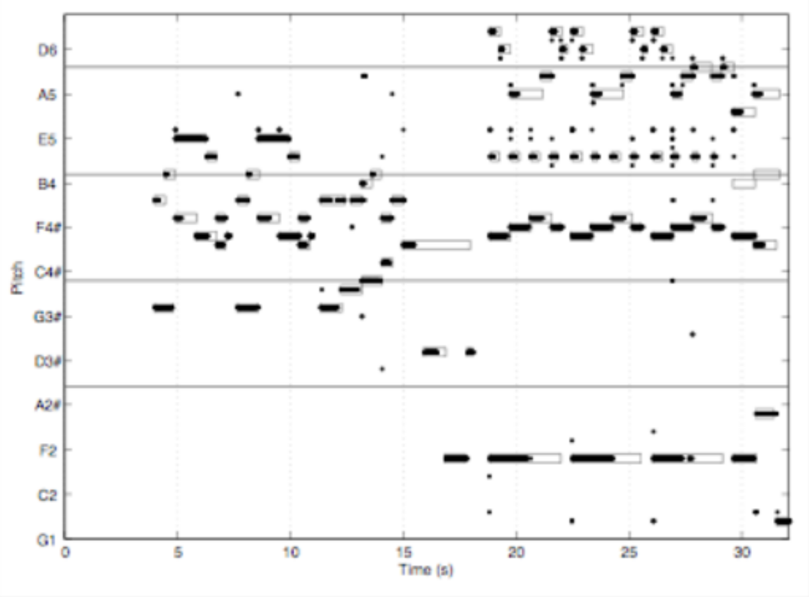


F-Measure per Class [top]

Class	CB1	CF4	CSF1	FMEGS1	FMESS1	MTB1	SB1	SB2	SB3	SB4	ZHZD1
Complex	54.0394	69.8131	59.2790	72.9934	71.8908	37.6174	81.5326	79.9402	69.1028	73.2215	78.9012
Poly_Pitched	78.3525	81.1789	76.7681	90.4637	89.0603	58.1099	94.2629	95.0894	88.8921	90.2298	91.1315
Solo_Bars_And_Bells	80.3214	80.3567	92.7480	99.3151	100.0000	33.1522	93.9598	100.0000	95.0483	95.2438	96.6667
Solo_Brass	45.9659	62.2026	63.4607	77.9309	81.0523	76.5729	90.8119	85.7223	66.5980	76.1804	73.8227
Solo_Drum	69.3859	89.3099	85.5897	91.3882	91.6318	17.1695	93.4206	93.3375	93.3595	93.6864	90.3285
Solo_Plucked_Strings	86.8081	72.9415	75.2496	79.7948	81.1344	67.1186	90.8423	91.8582	94.0888	91.3557	87.6150
Solo_Singing_Voice	20.9078	17.4294	19.1214	40.8665	38.4918	29.9634	60.8542	57.1073	42.2155	51.9225	44.2599
Solo_Sustained_Strings	32.9943	52.8066	14.2403	59.6637	63.3406	40.0007	75.4091	72.7022	44.0138	41.8737	63.8987
Solo_Winds	47.3817	49.6935	11.1533	61.5631	67.5481	63.5897	79.5300	74.6796	60.0772	55.3723	66.4060

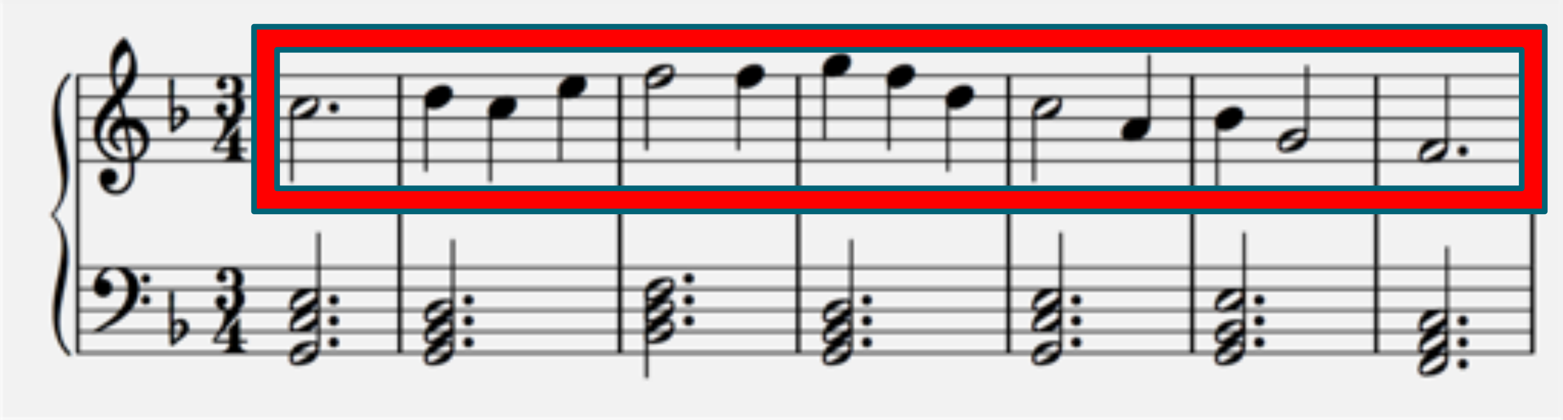
Multiple F0 estimation

- many approaches tried
  - Deep Learning (neural networks)
  - Non-negative Matrix Factorization (NMF)
  - …
- for example, in NMF the audio signal is rendered as a matrix, which is the a product of
  - a matrix of activations (sound events)
  - a matrix of (learned) note templates
- aim is to calculate the matrix of activations that has the best match with the audio signal
- activations correspond to notes



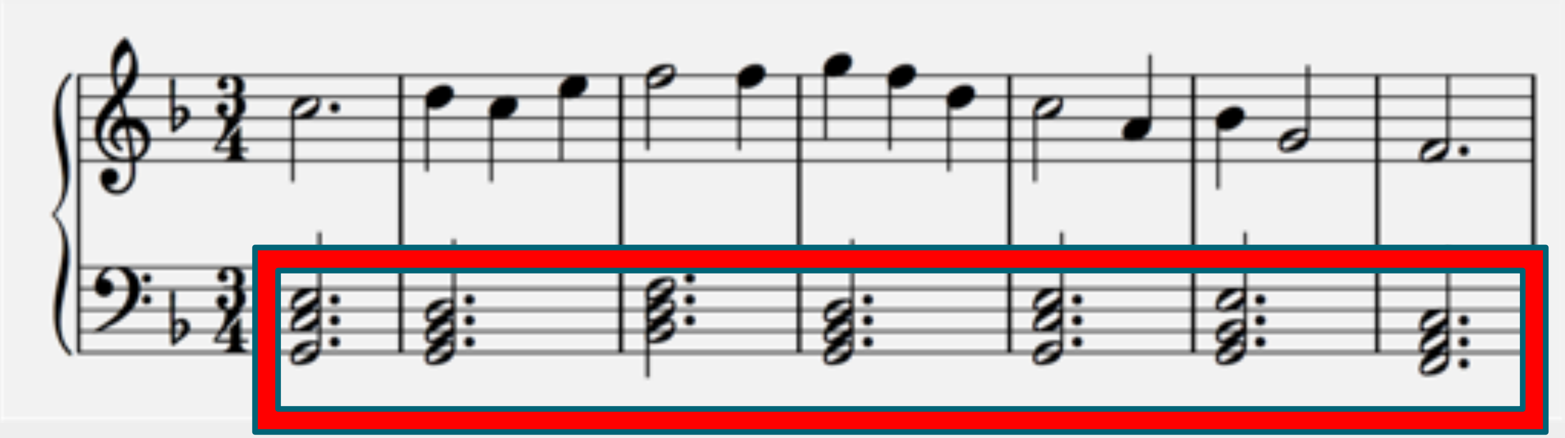
automatic transcription of piano piece (black) and ground truth (white)

Chord recognition



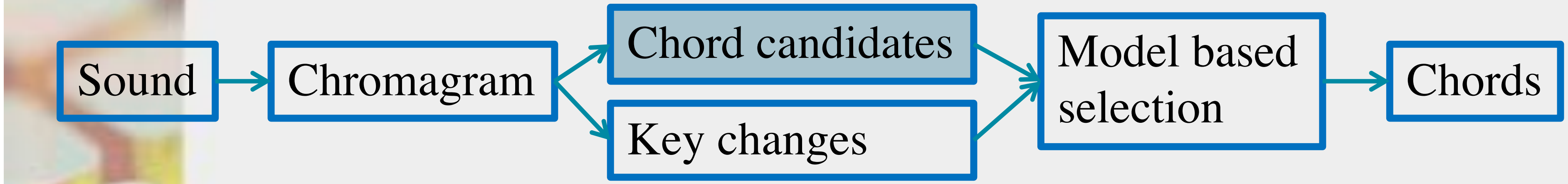
- Melody vs Chords:
  - *Note*: a single sounding tone with a pitch (height)
    - *Melody*: a sequence of monophonic notes

Chord recognition



- Melody vs Chords:
  - *Note*: a single sounding tone with a pitch (height)
    - *Melody*: a sequence of monophonic notes
  - *Chords*: sounding of simultaneous (at least two) notes
    - *Chord sequence*: a sequence of chords

Pipeline: chord recognition


Sound

Chromagram

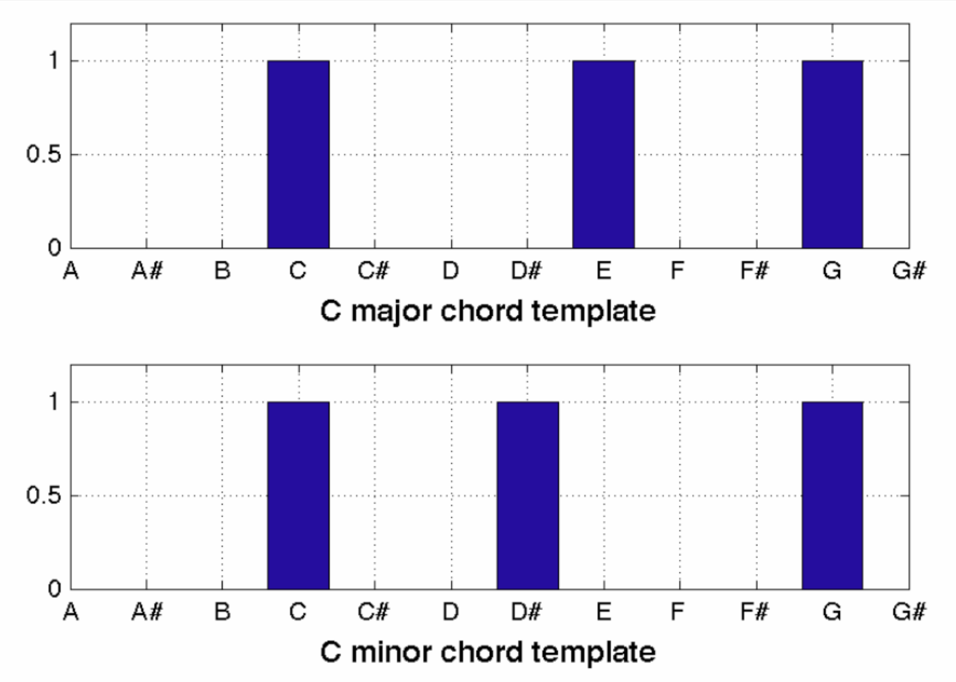
Chord candidates

Key changes

Model based selection

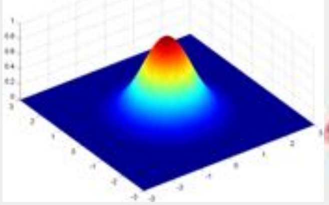
Chords

- Match a the beat synchronised chroma features with chord templates
- Different approaches:
  - Use knowledge-based templates
  - Match by Euclidean distance
  - Learn an average profile from the data
  - Learn 12 dimensional Gaussian distribution for all chords

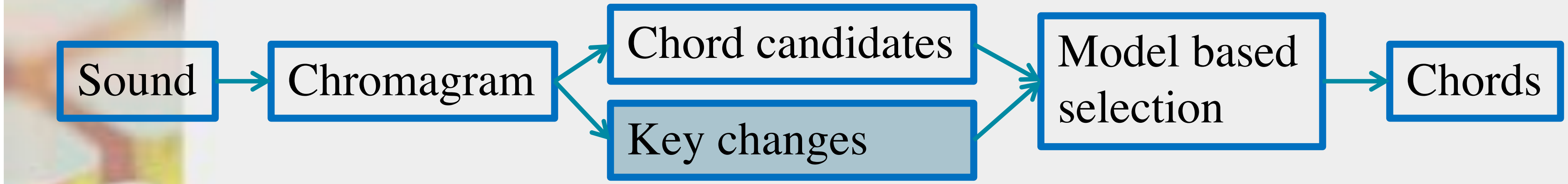


C major chord template

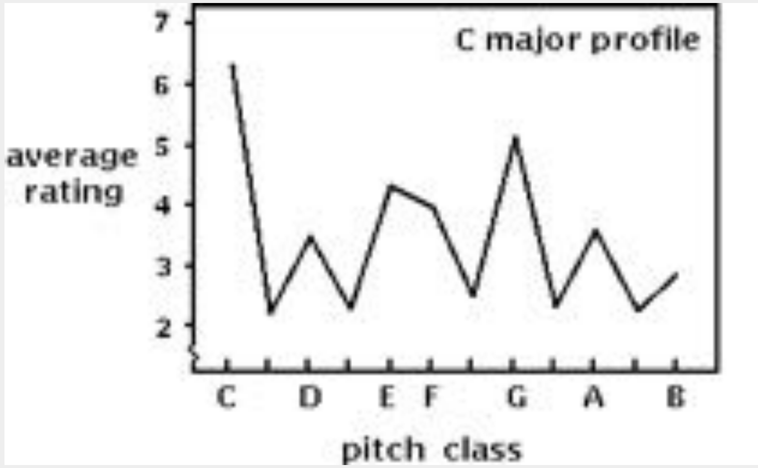
C minor chord template

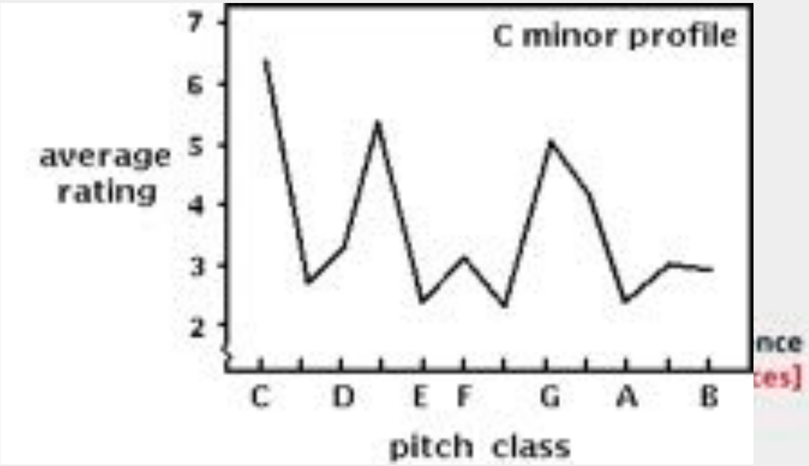


Case study: chord recognition



- Some models require information about the Key
- General approach
  - Krumhansl-Kessler profiles
  - Match by Pearson correlation with chroma feature
  - Variations exist
  - Approach is similar to chord candidate selection





nce
es]

Universiteit Utrecht

Short recap: What is a key?

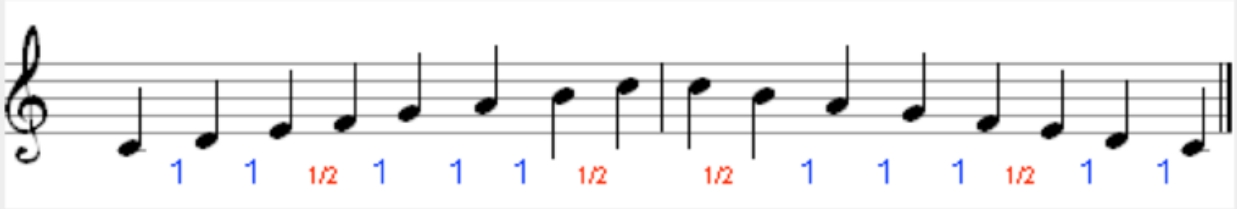
- pitch is generally not equally distributed within a piece of music
- if it is, you get ‘atonal’ music
  - e.g. Webern’s piano variations
- when we use only a subset, music generally sounds much more structured

Short recap: What is a key?

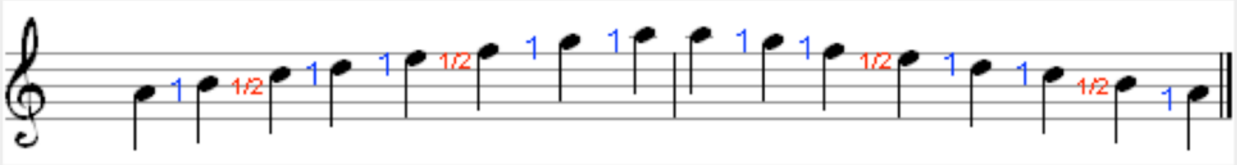
- subsets are often visualised as musical scales
  - perceptually, they help us identify ‘tonality’
  - music hovers around certain pitch, the ‘final’ or ‘tonic’
- most common scales: major and minor
  - 7 different pitches within octave
  - most audible difference: third note of scale
  - change can be quite dramatic

http://virtualpiano.net/

major

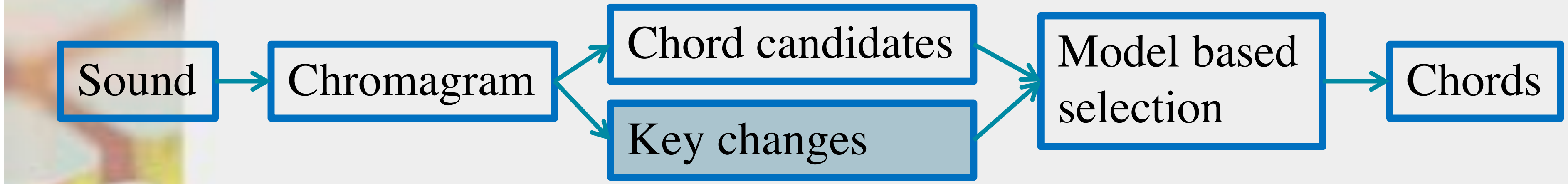


minor



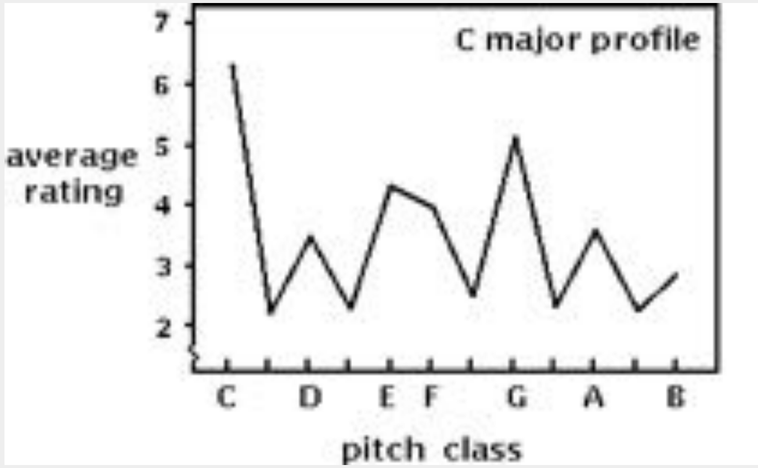
ex. Gustav Mahler, 1st symphony, 3rd mvt

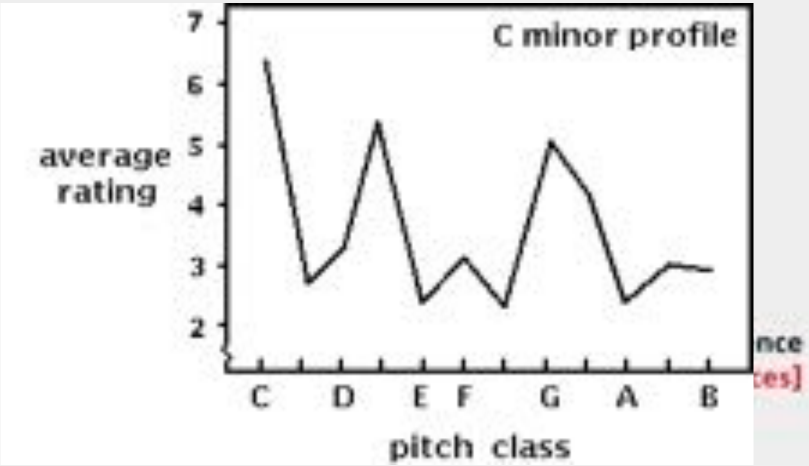
Case study: chord recognition



- Some models require information about the Key

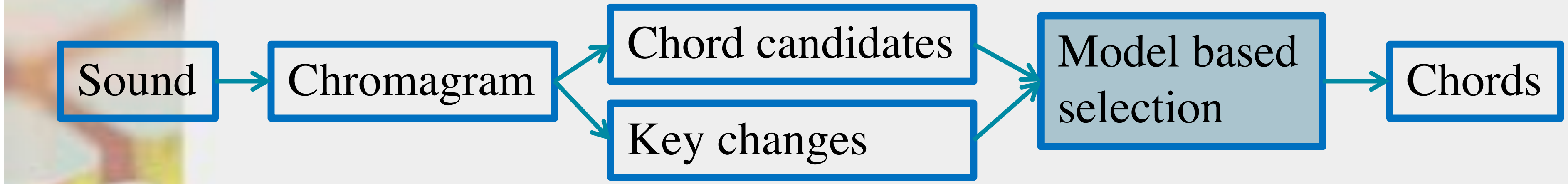
- General approach
  - Krumhansl-Kessler profiles
  - Match by Pearson correlation with chroma feature
  - Variations exist
  - Approach is similar to chord candidate selection





nce
es]

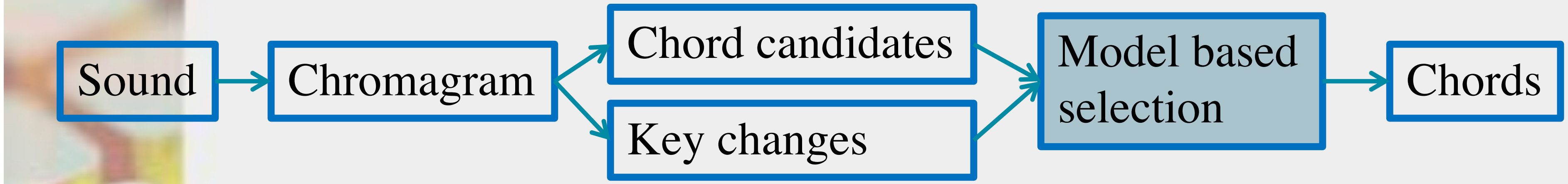
Case study: chord recognition



Two kind of approaches:

- Knowledge driven:
  - Musical knowledge is modelled and used to select a plausible chord sequence
- Data driven:
  - The Transition probabilities between chords are learned from a large corpus of chords

Case study: chord recognition

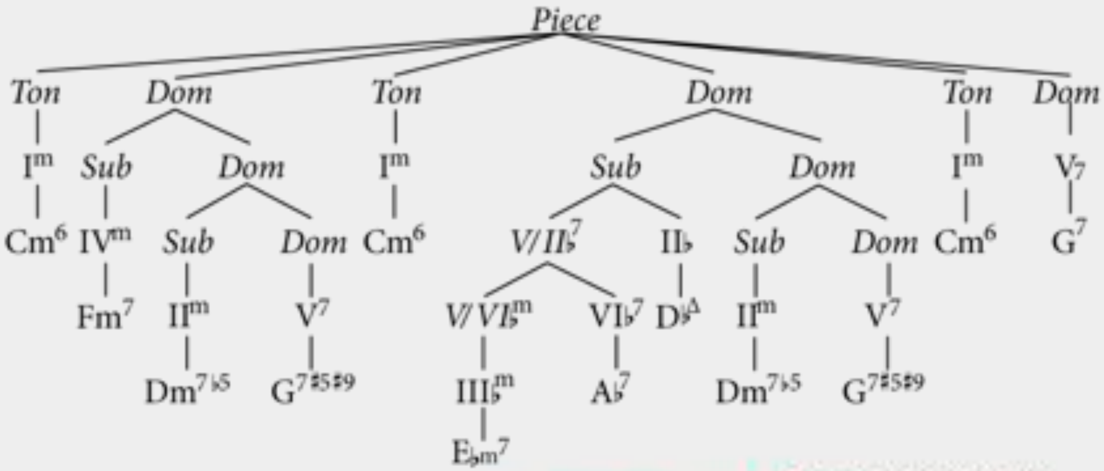


Sound → Chromagram → Chord candidates / Key changes → Model based selection → Chords

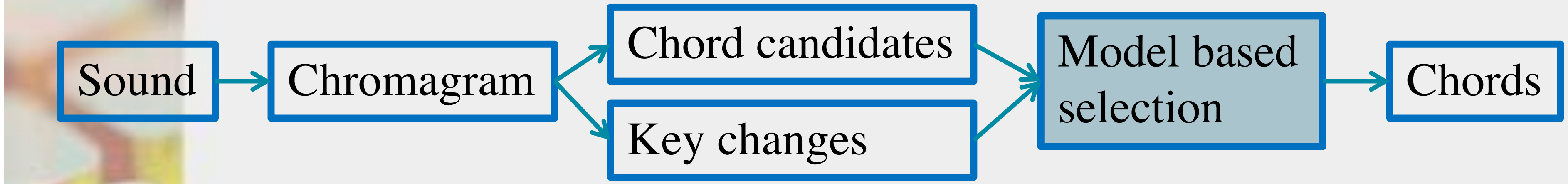
Knowledge driven:

HarmTrace model (de Haas):

- Analyses the function of chord
- Needs key information
- Robust against noisy data
- Flexible
- Based on functional programming



Case study: chord recognition



Data Driven:

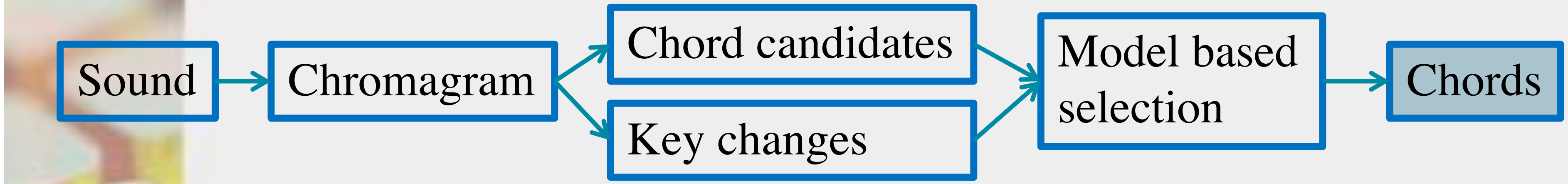
Hidden Markov models:

- Estimate the probability of a chord-transition by counting the chord transitions in a corpus
- Use chroma features to estimate the probability of a chord candidate
- Goal: to find the most likely sequence of chords that results on the currer chromagram:
  - Viterbi algorithm
- Many variants exist

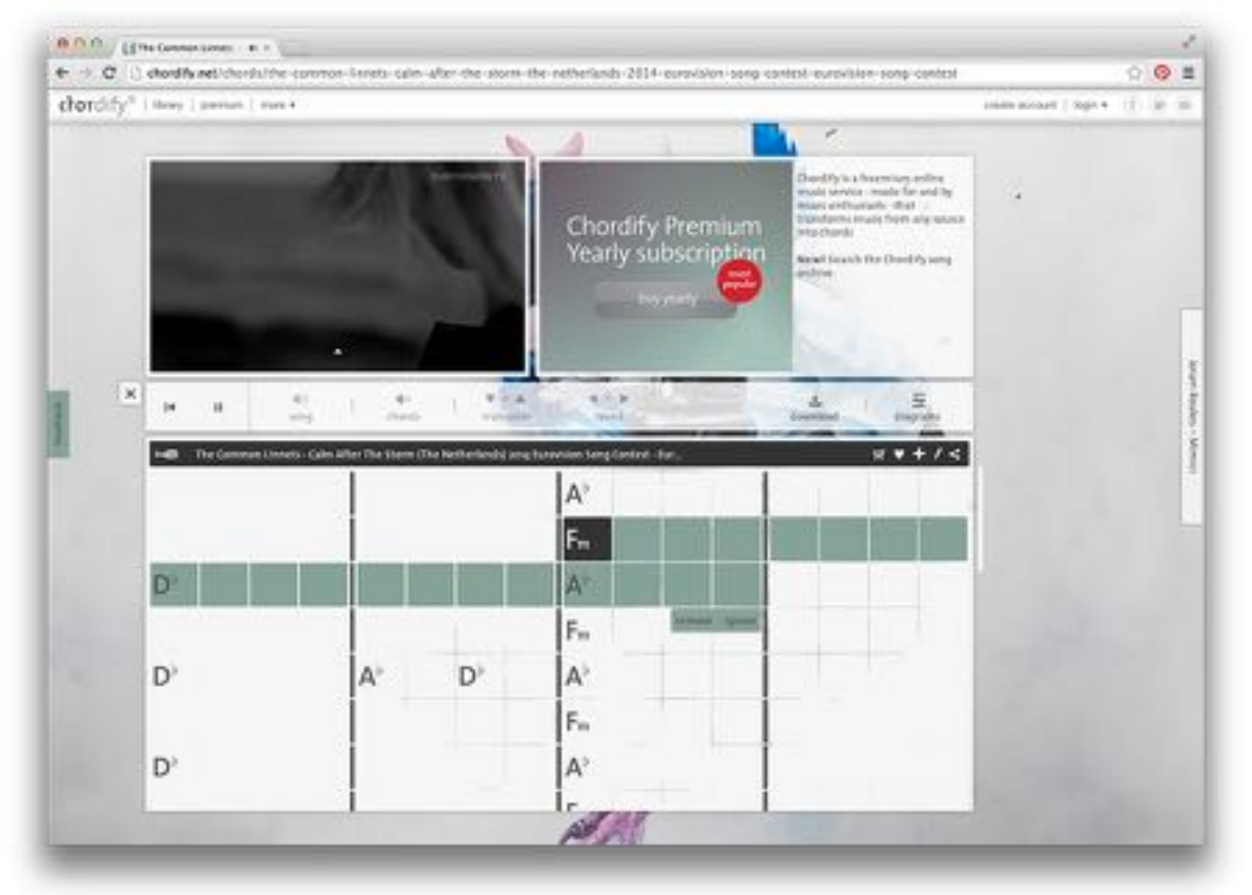
Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

Case study: chord recognition



chordify®



chordify.net/chords/the-common-linnets-calm-after-the-storm-the-netherlands-2014-eurovision-song-contest-eurovision-song-contest

Chordify Premium
Yearly subscription

The Common Linnets - Calm After The Storm (The Netherlands) 2014 Eurovision Song Contest - Eur...

A♭
Fm
D♭ A♭
Fm
D♭ A♭ D♭ A♭
Fm
D♭ A♭

Practical approaches

- based on consideration of (western) predominant musical structures
- melody + accompaniment is common
- complex polyphony is relatively scarce, especially in popular music
  - counterpoint between the ‘inner voices’ does matter in classical works such as fugues (or the ex. of the Shostakovich string quartet)
- pragmatic approaches are thus focussing on
  - melody extraction
  - chord recognition

Summary

- transcription = deriving a symbolic representation from musical audio
- F0 estimation: deriving perceived pitch from audio
- Query by Humming as early MIR paradigm
- F0 estimation for monophony considered solved
  - but in practice…
- multiple F0 estimation: very hard task
- practical restrictions: melody extraction, chord labelling
- onset detection: yet another important task
- no functional systems yet that combine F0 extraction and onset detection