

Topic 11

Score-Informed Source Separation
(chroma slides adapted from Meinard Mueller)

Why Score-informed Source Separation?

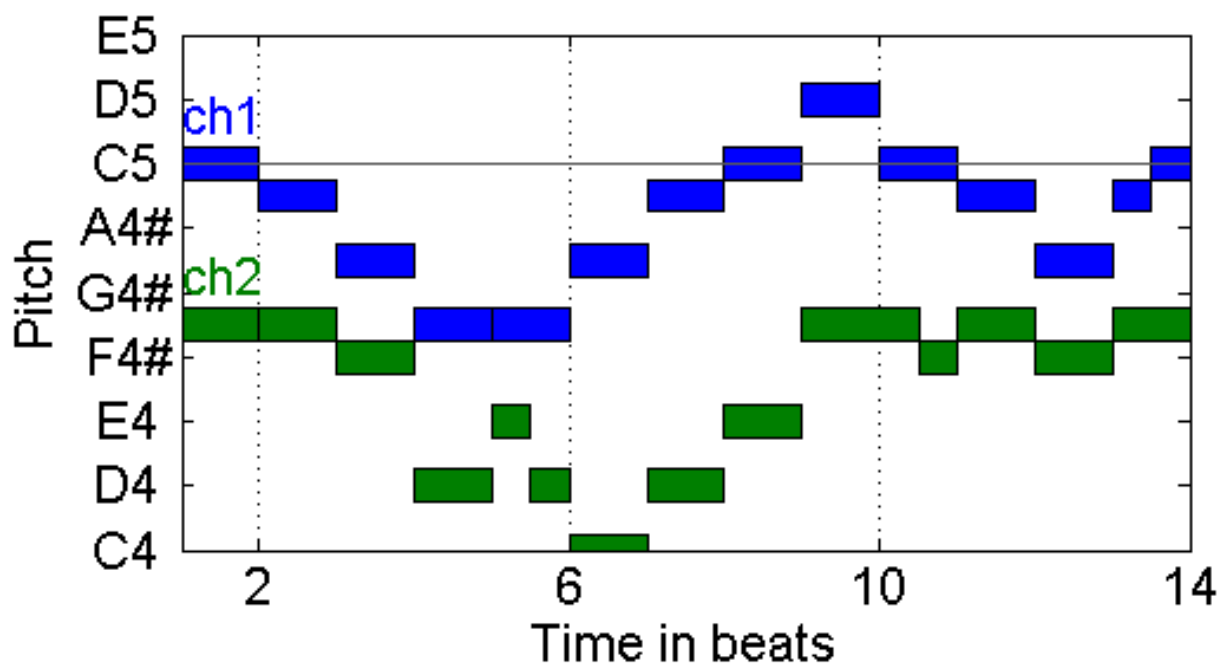
- Audio source separation is useful
 - Music transcription, remixing, search
- Non-satisfying results if only using audio
- Score provides some info that one can use
 - E.g., conductor, learn to sing in a choir
- Lots of scores are out there

Musical Score in MIDI

score
sheet

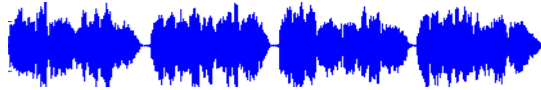


MIDI
score



Would it be trivial?

instantiation



abstraction



- Is map-informed tourism trivial (for machine)?

Remaining Tasks

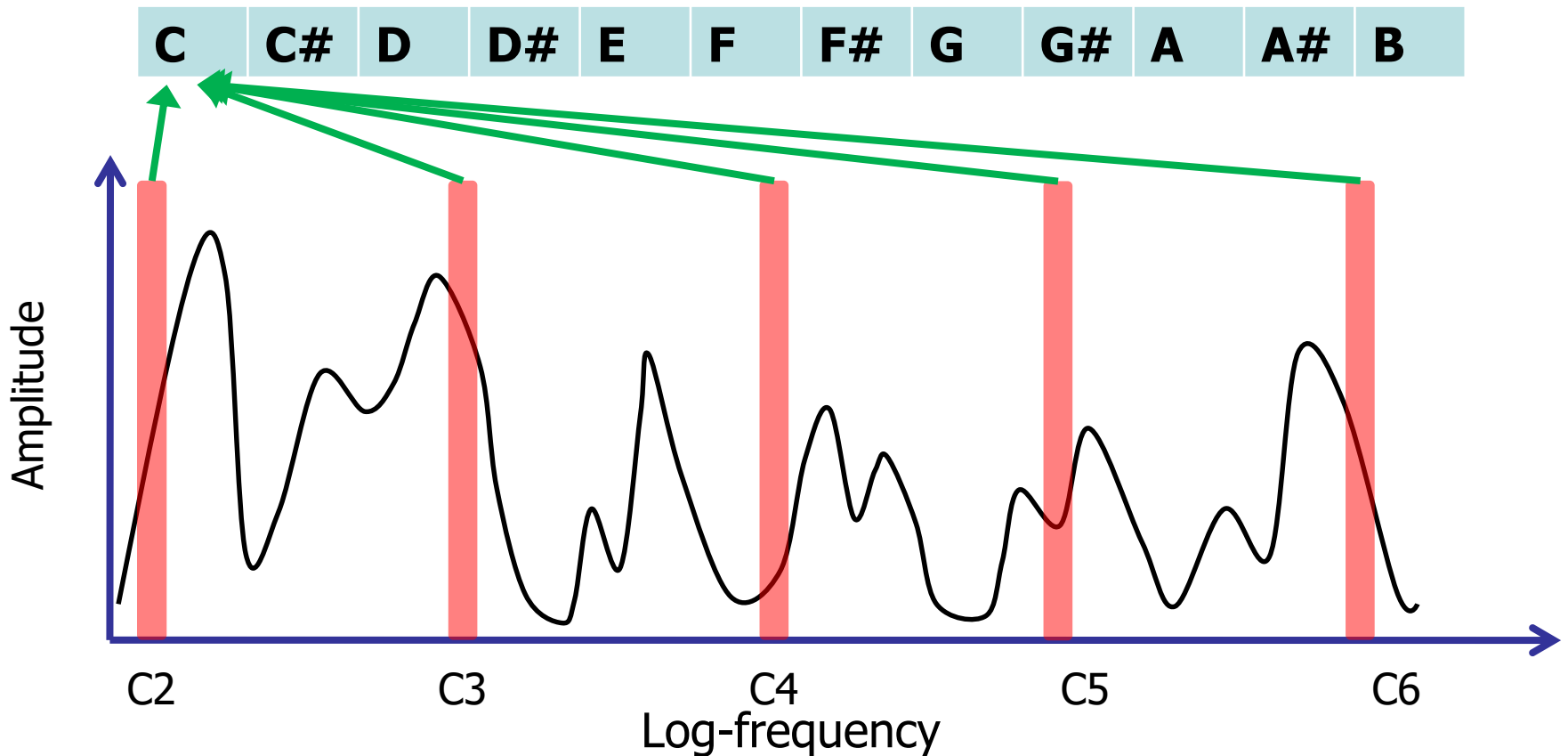
- Score tells us **what** musical objects to look for, but not **where to look** nor **what they sound like**.
- Problems
 - How to align audio with score?
 - How to represent them?
 - How to separate the signal?

Audio/Score representations for alignment

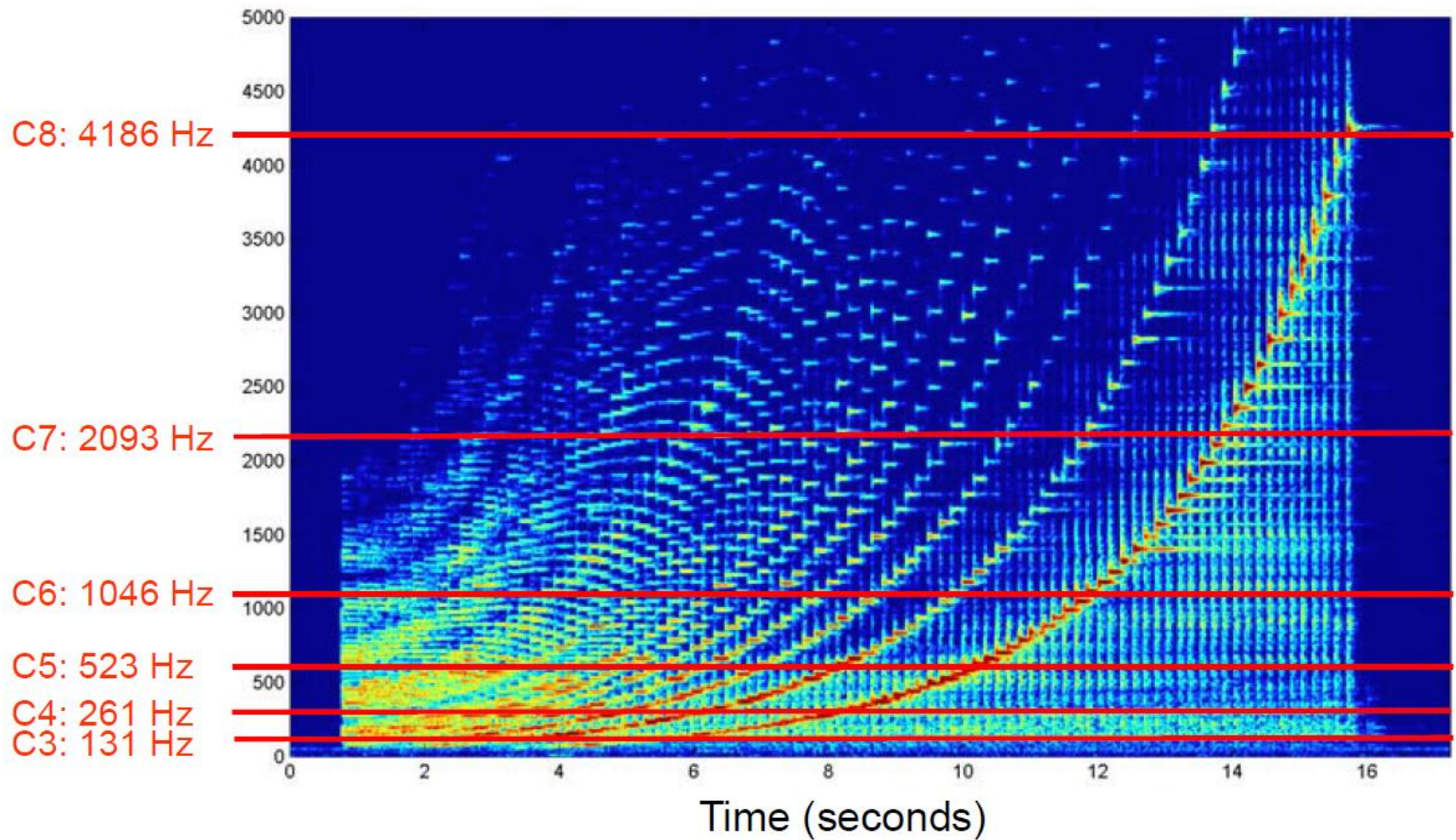
- Represent in the same way
 - Spectrum
 - Only good for monophonic music
 - Chroma feature
 - Good for polyphonic music
 - Pitch info
 - Ideal for both monophonic and polyphonic music
 - Relies on good **multi-pitch estimation** techniques

Chroma Feature

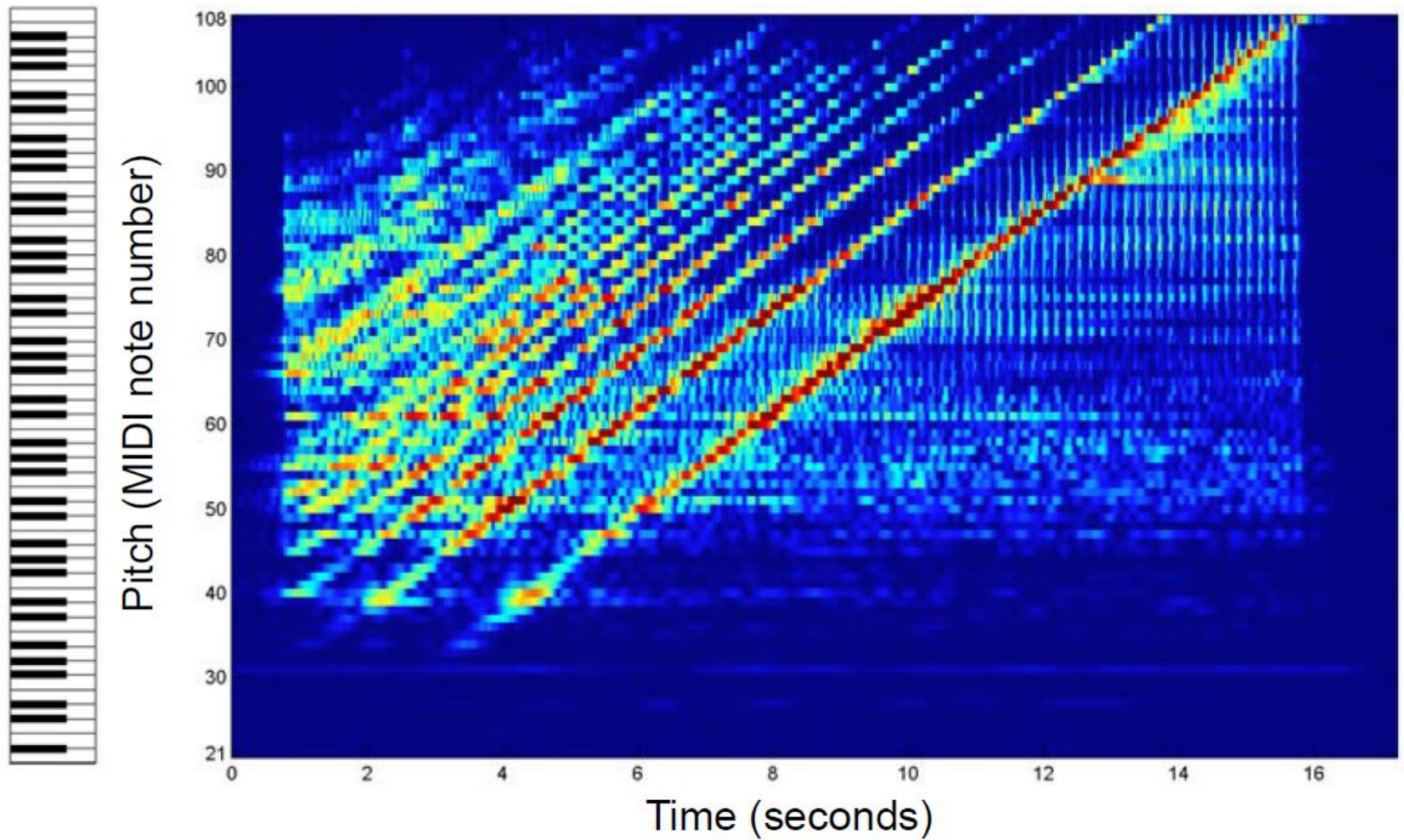
- Spectral energy of the 12 pitch classes
 - 12-d vector



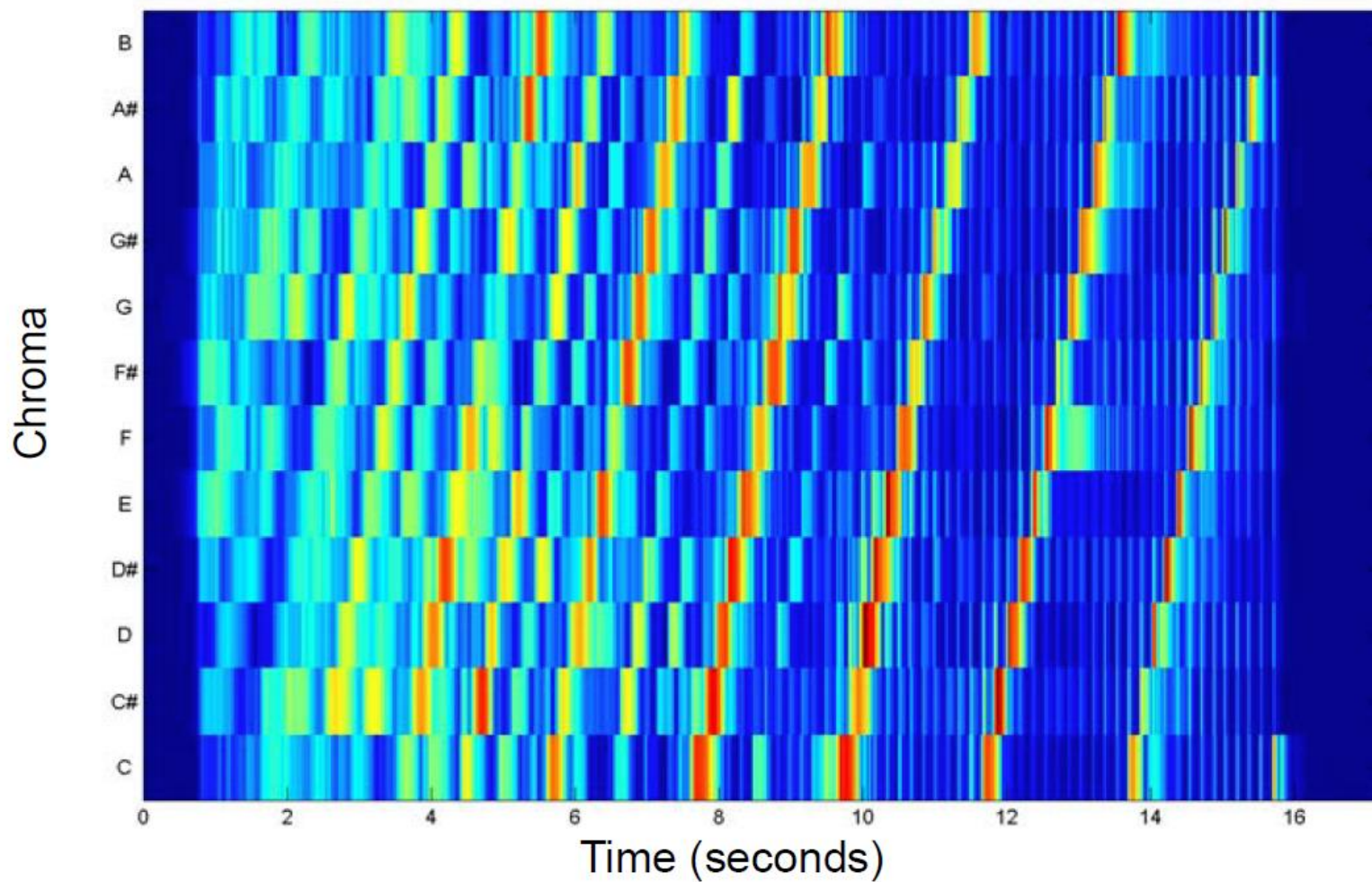
Spectrogram



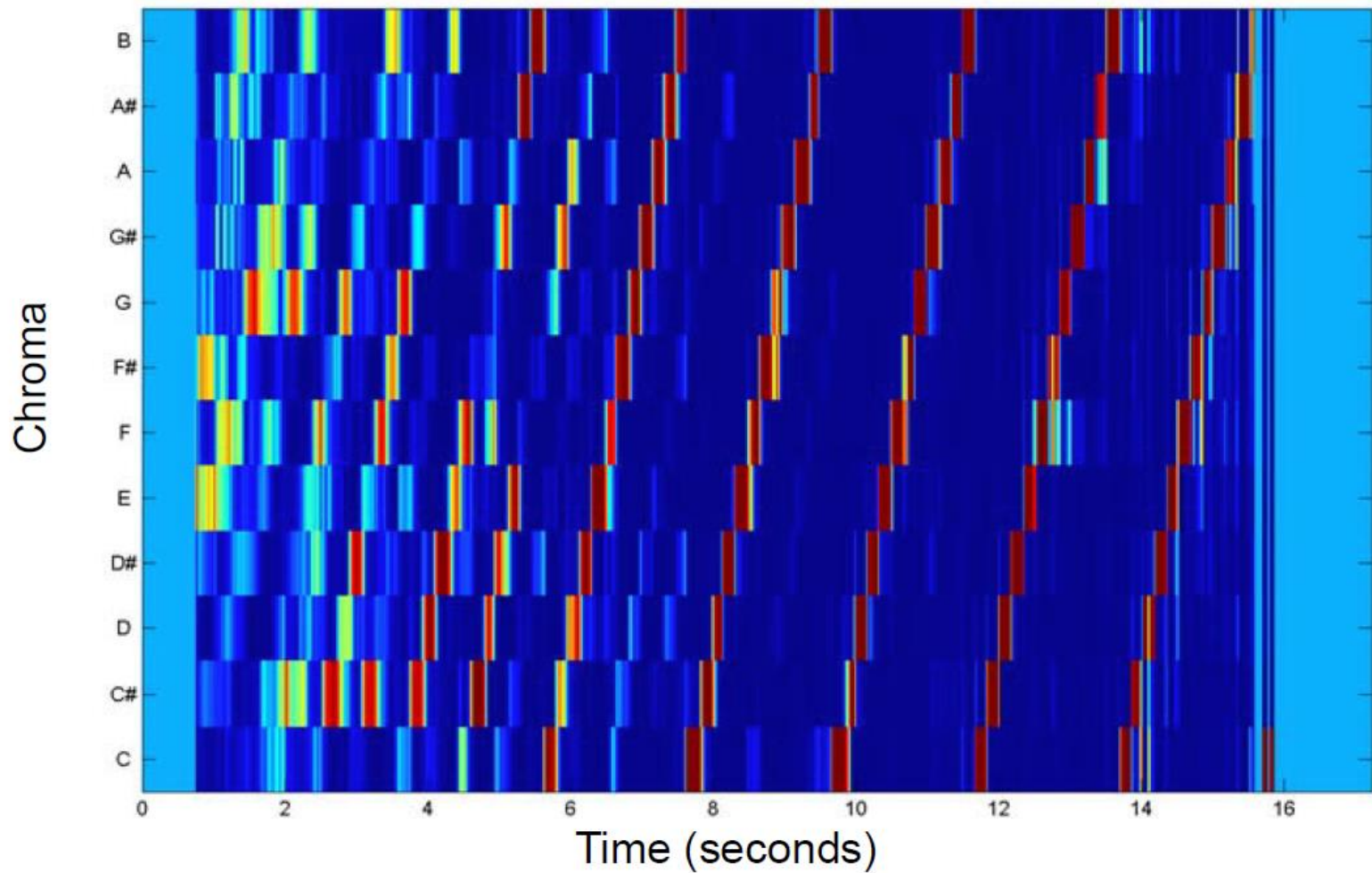
Log-frequency Spectrogram



Chromagram

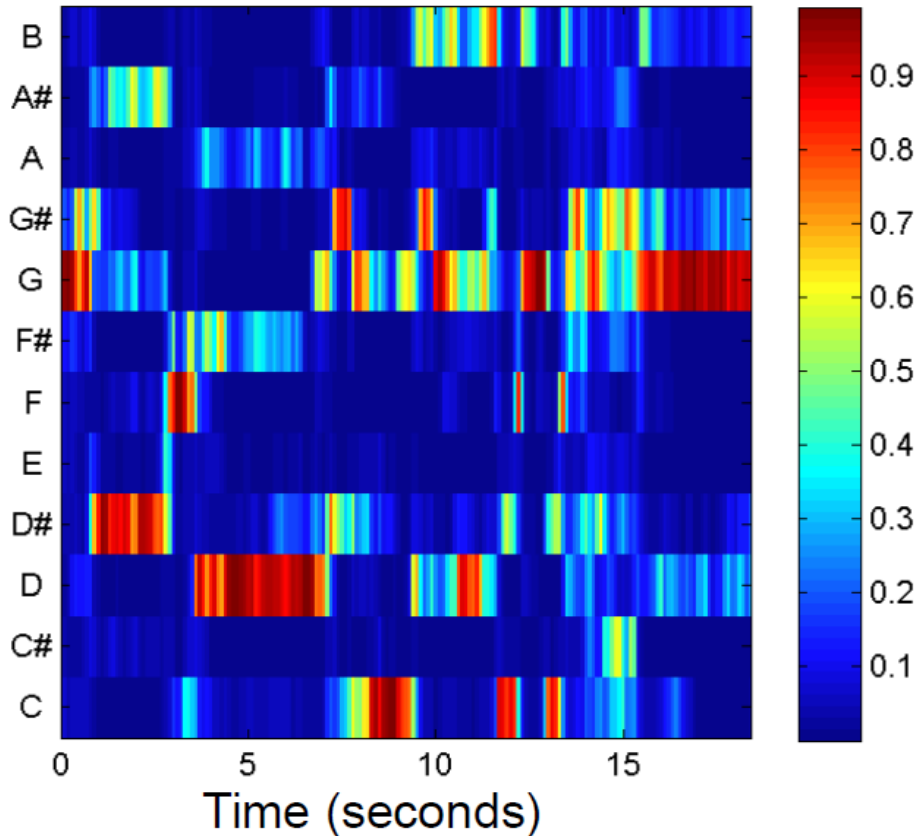


Normalized Chromagram

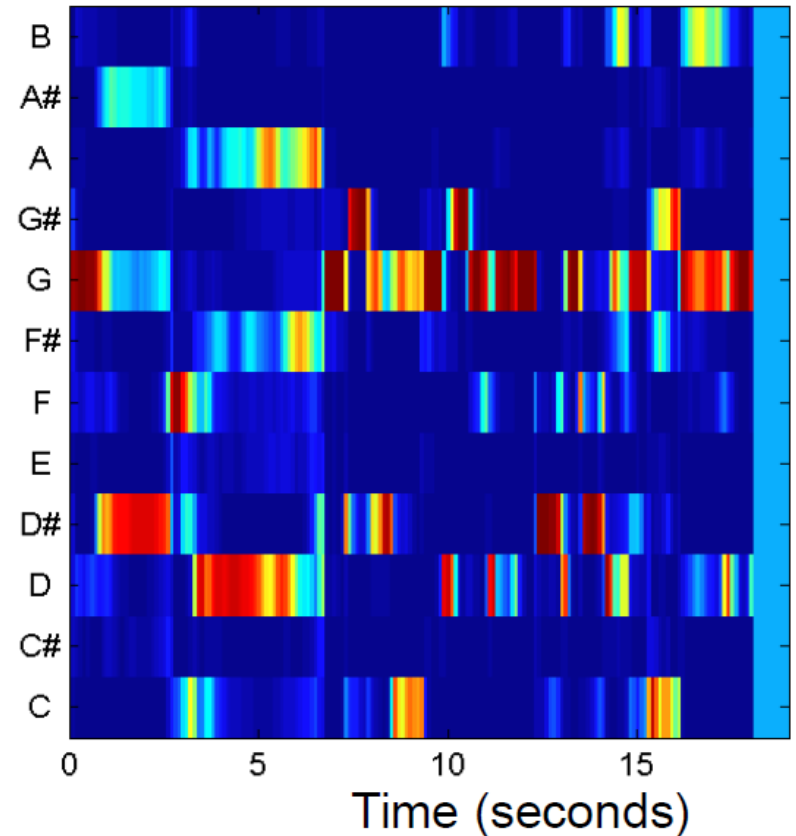


Chromagram of Polyphonic Music

Karajan

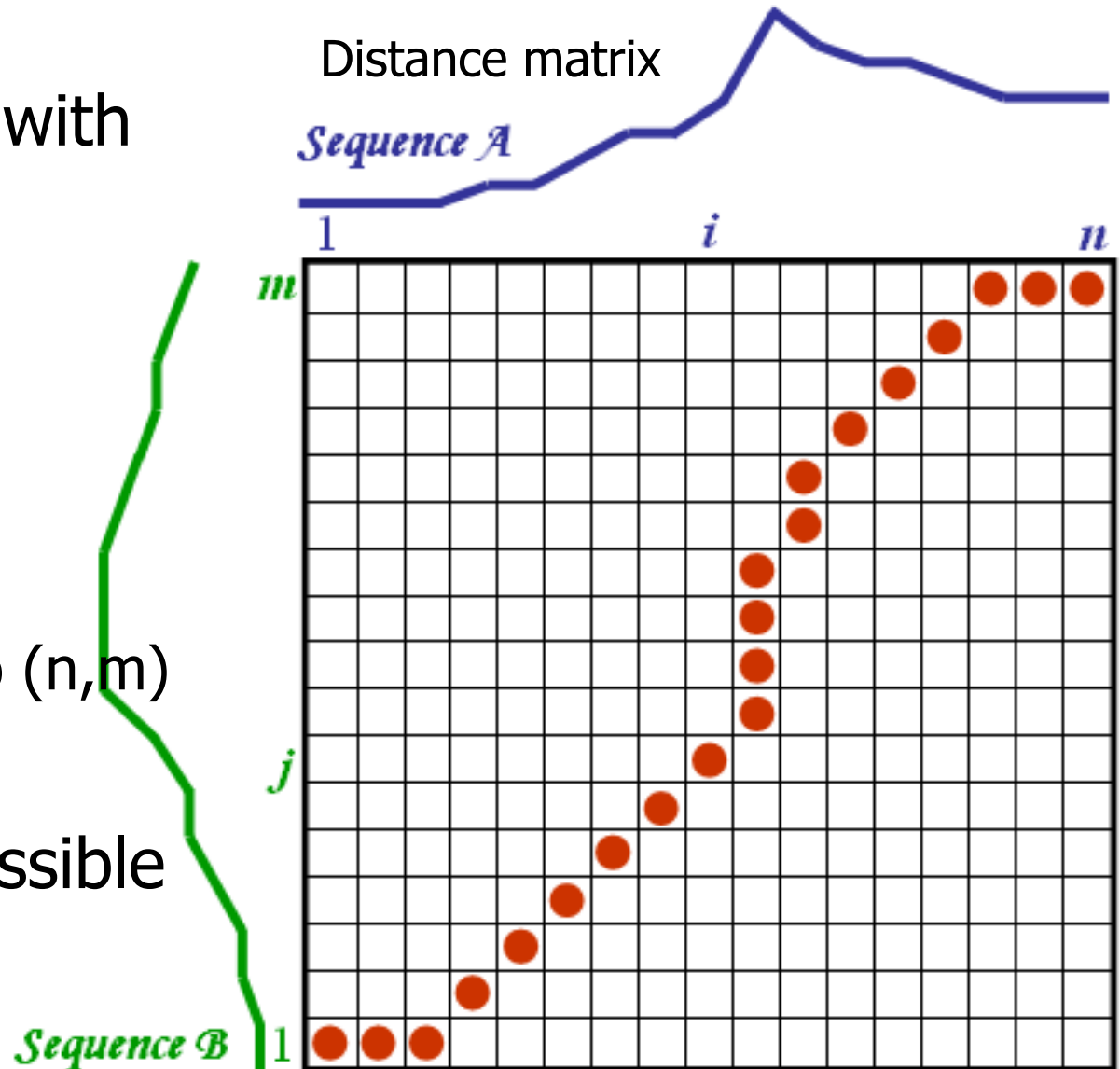


Scherbakov



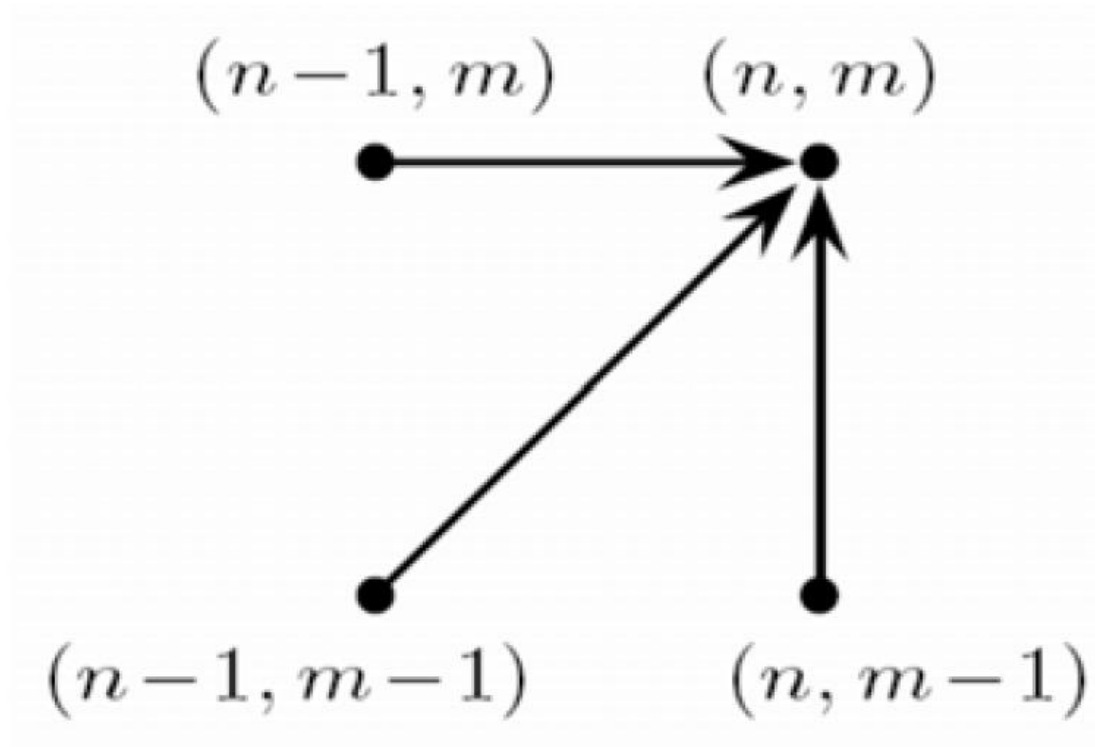
Dynamic Time Warping

- Find the path with lowest cost
- A path should
 - Monotonic
 - Step size 1
 - From $(1,1)$ to (n,m)
- How many possible paths?



Possible Progression

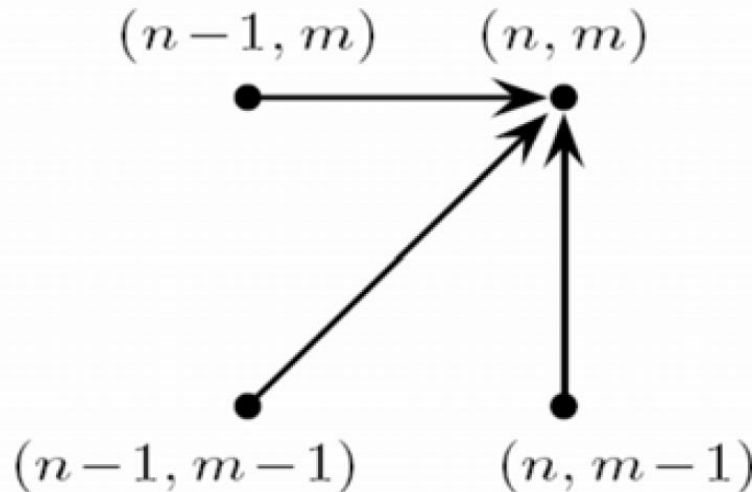
- Three ways for a path to get to (n, m) in one step



A Nice Property

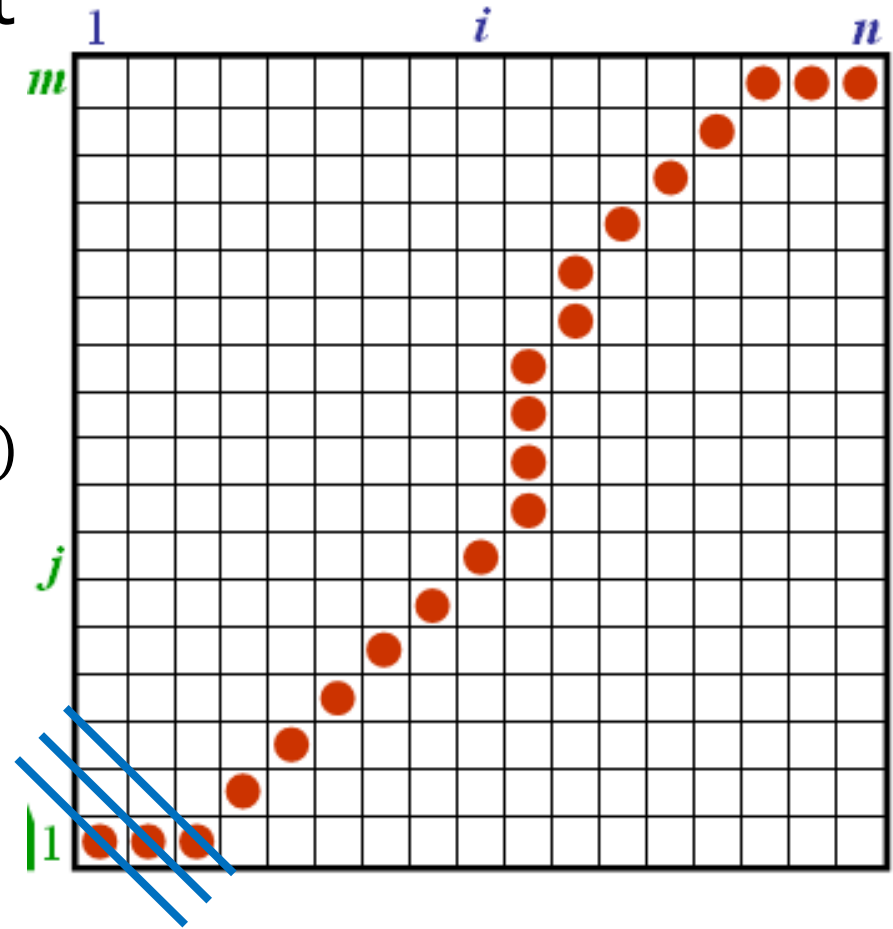
- Let $d(i, j)$ be the distance matrix
- Let $C(n, m)$ be the **lowest** cost from $(1, 1)$ to (n, m)
 - Then $C(1, 1) = d(1, 1)$

$$C(n, m) = \min \begin{cases} C(n-1, m) + d(n, m) \\ C(n-1, m-1) + d(n, m) \\ C(n, m-1) + d(n, m) \end{cases}$$



Dynamic Programming!

- Calculate the lowest cost matrix $C(i, j)$
 - Starting from $C(1,1)$
 - Then calculate $C(1,2), C(2,1)$
 - Then $C(1,3), C(2,2), C(3,1)$
 -
 - Finally, calculate $C(n, m)$
- Remember how you calculated, and trace back to get the path



Two SISS Systems for Polyphonic Music

- Score-informed NMF

[Ewert et al., 2009]

[Ewert & Muller, 2012]

- Chroma feature to represent audio
- Dynamic time warping for alignment
- NMF-based separation
- Offline

- Soundprism

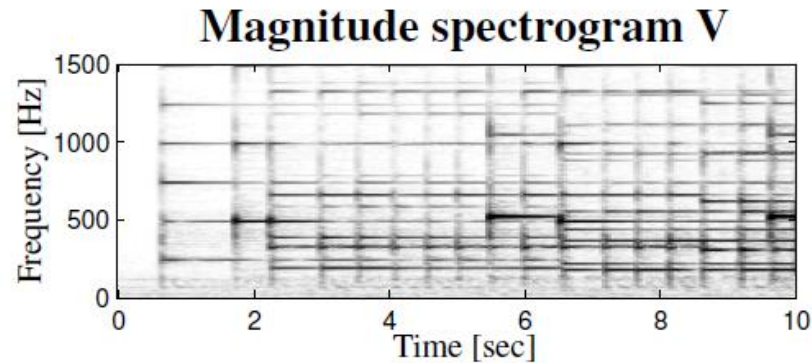
[Duan & Pardo, 2011]

- Multi-pitch info of audio
- Particle filtering for alignment
- Pitch-based separation
- Online

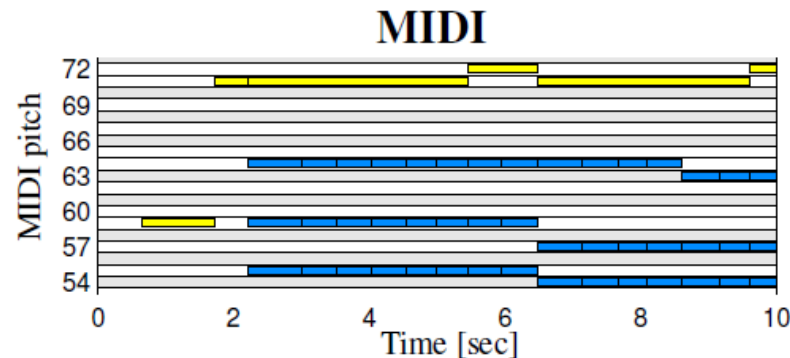
Score-informed NMF

[Ewert & Muller, 2012]

Polyphonic
audio



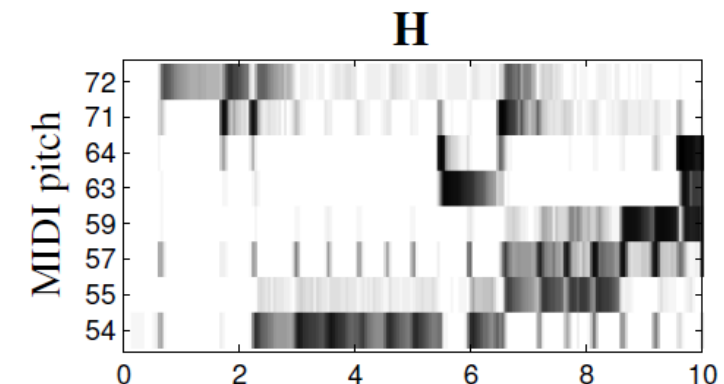
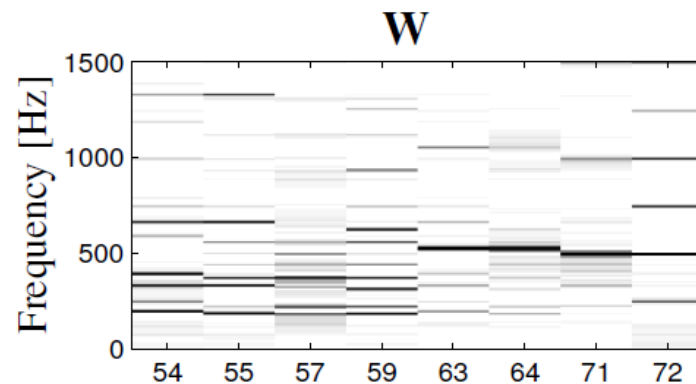
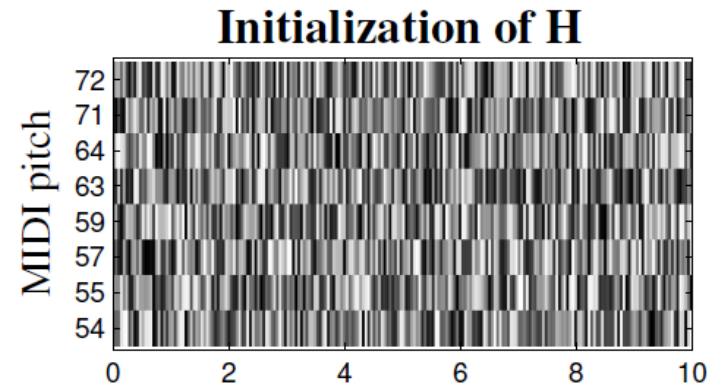
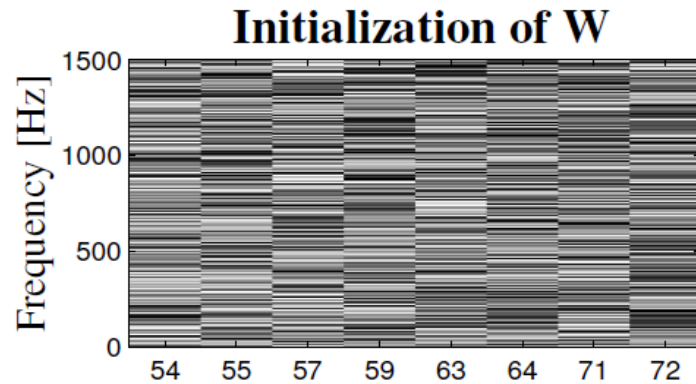
Aligned
MIDI score



Score sheet

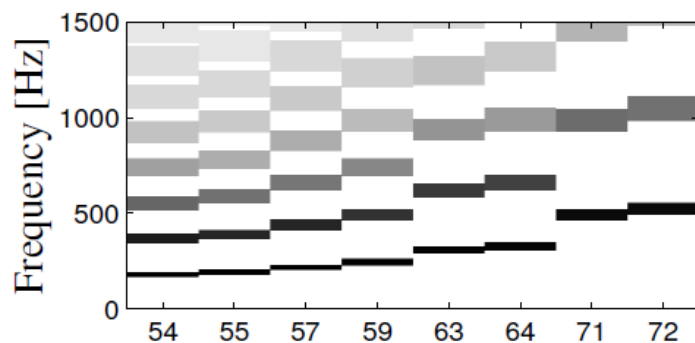


When score info is not used

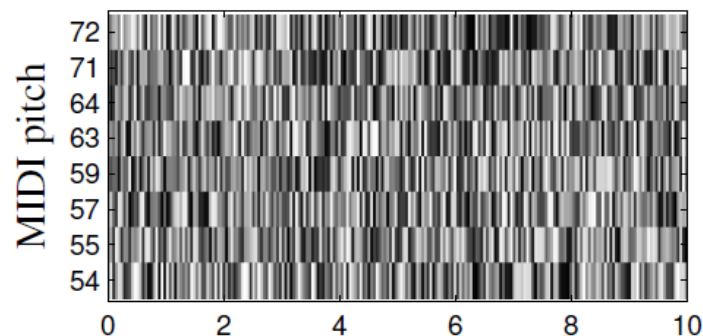


When **dictionary** is initialized by score notes

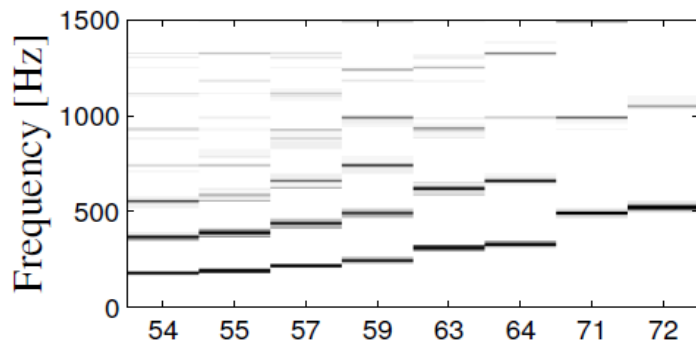
Initial W



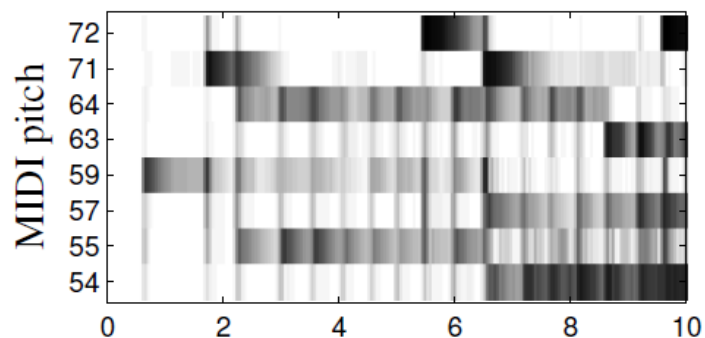
Initial H



Final W

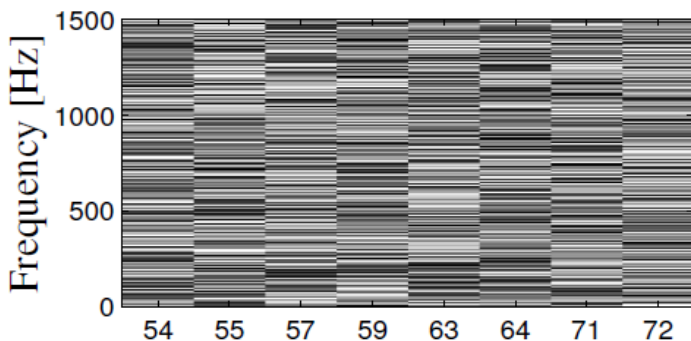


Final H

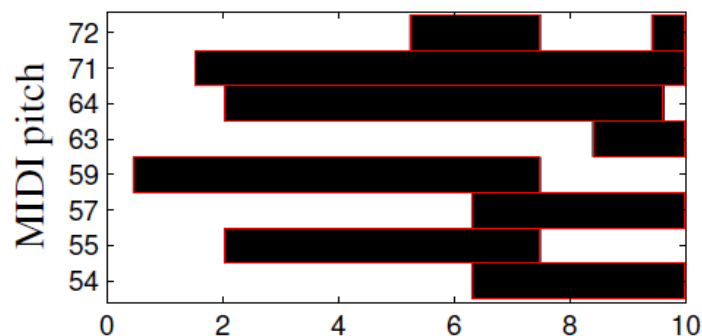


When **activation** is initialized by score notes

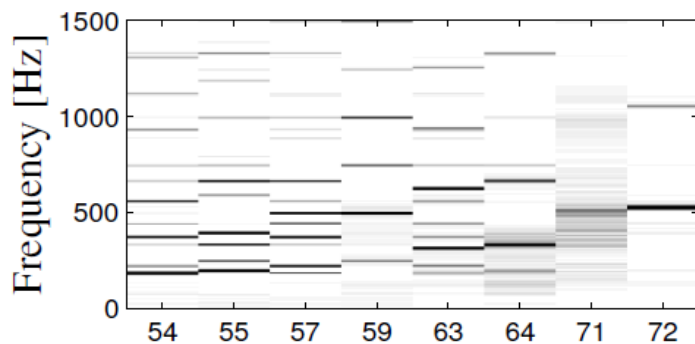
Initial W



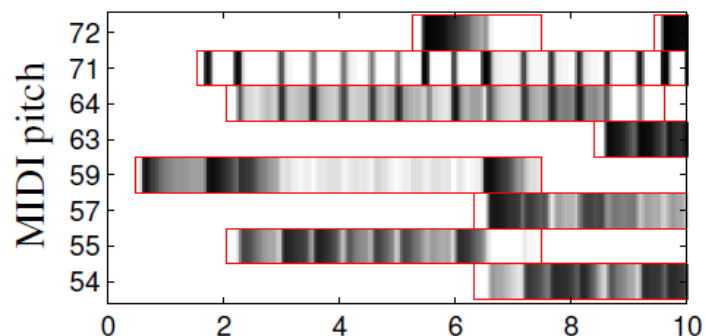
Initial H



Final W

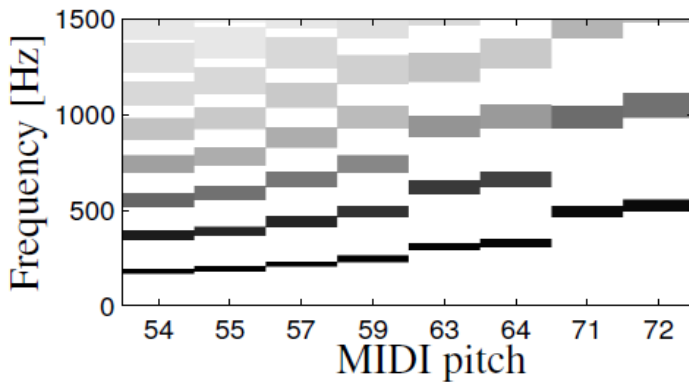


Final H

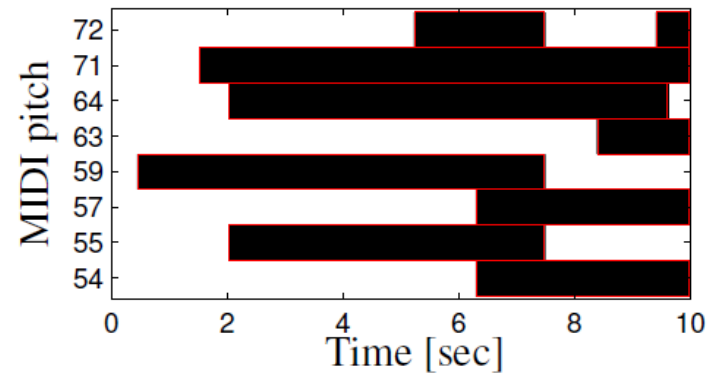


When both W and H are initialized

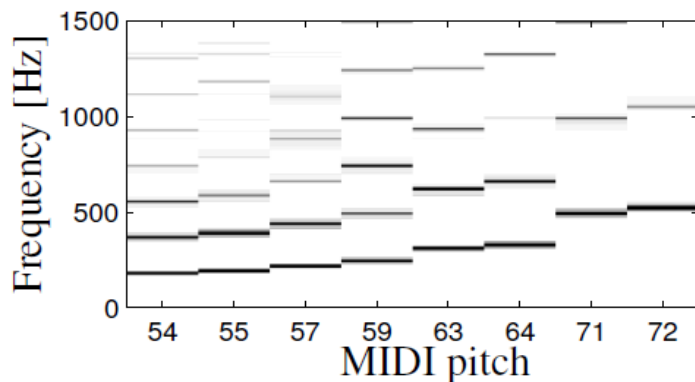
Initial W



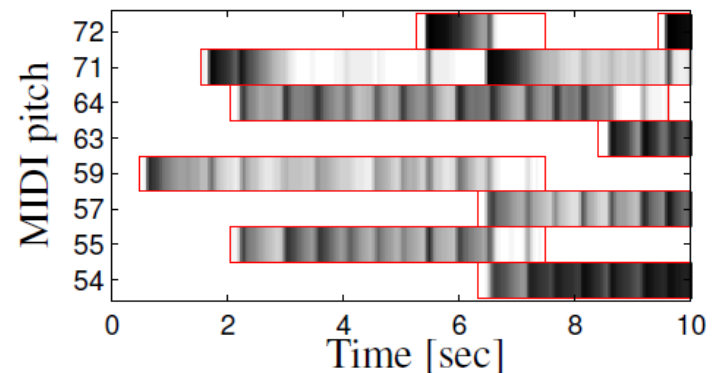
Initial H



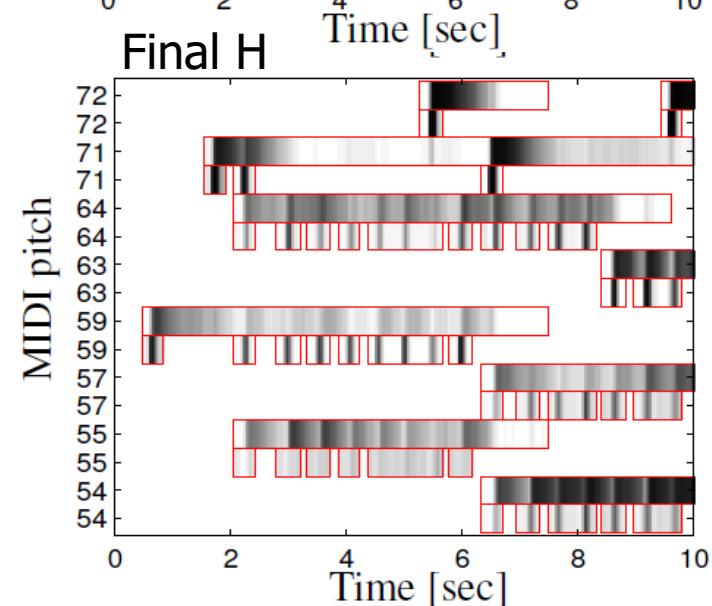
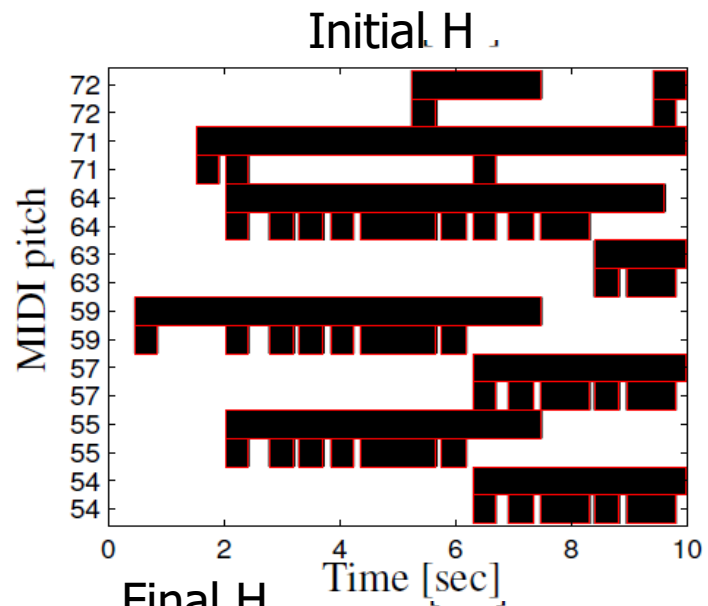
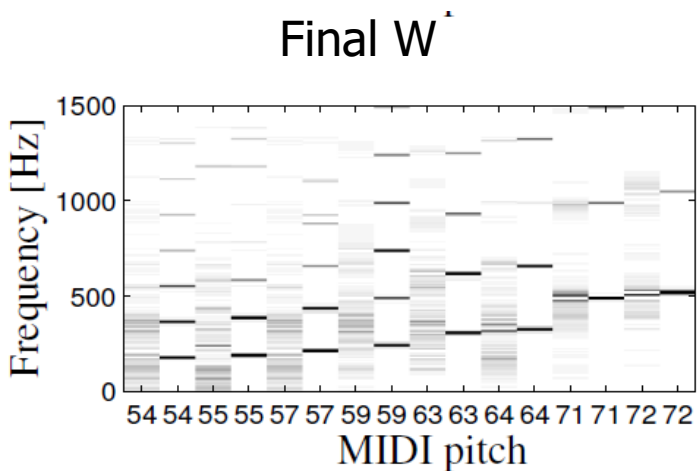
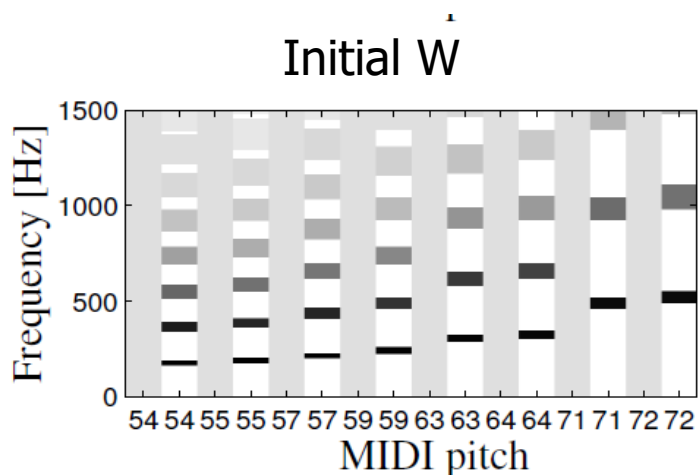
Final W



Final H

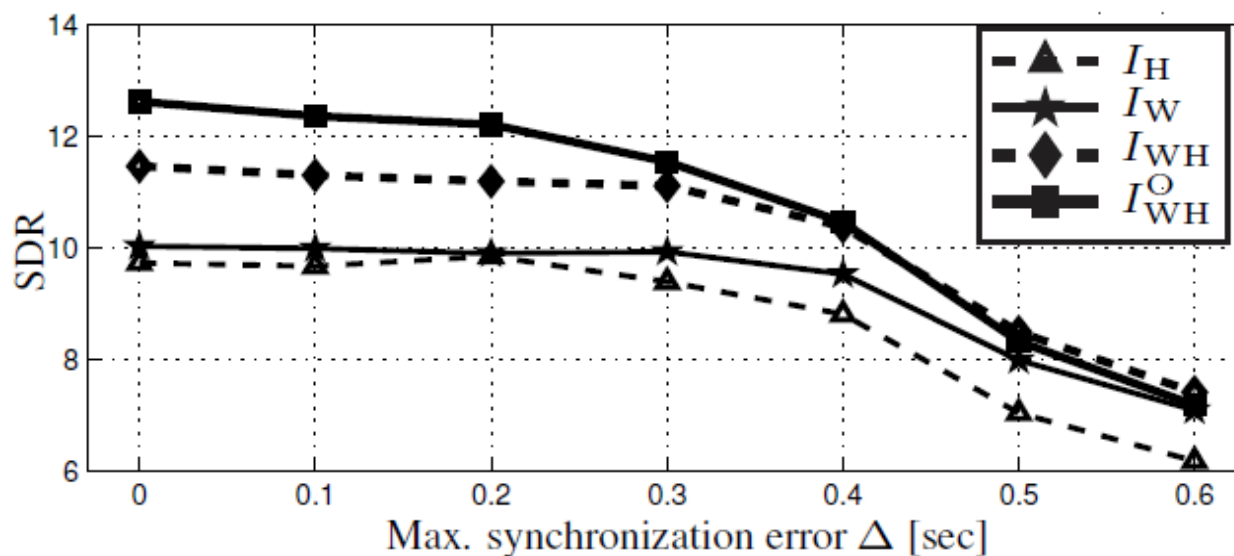


Also Considering Onset Models



Experiments

- MIDI-synthesized piano music with randomly imposed alignment errors
 - Audio has accurate pitch, simple timbre
- Separate left/right hand notes



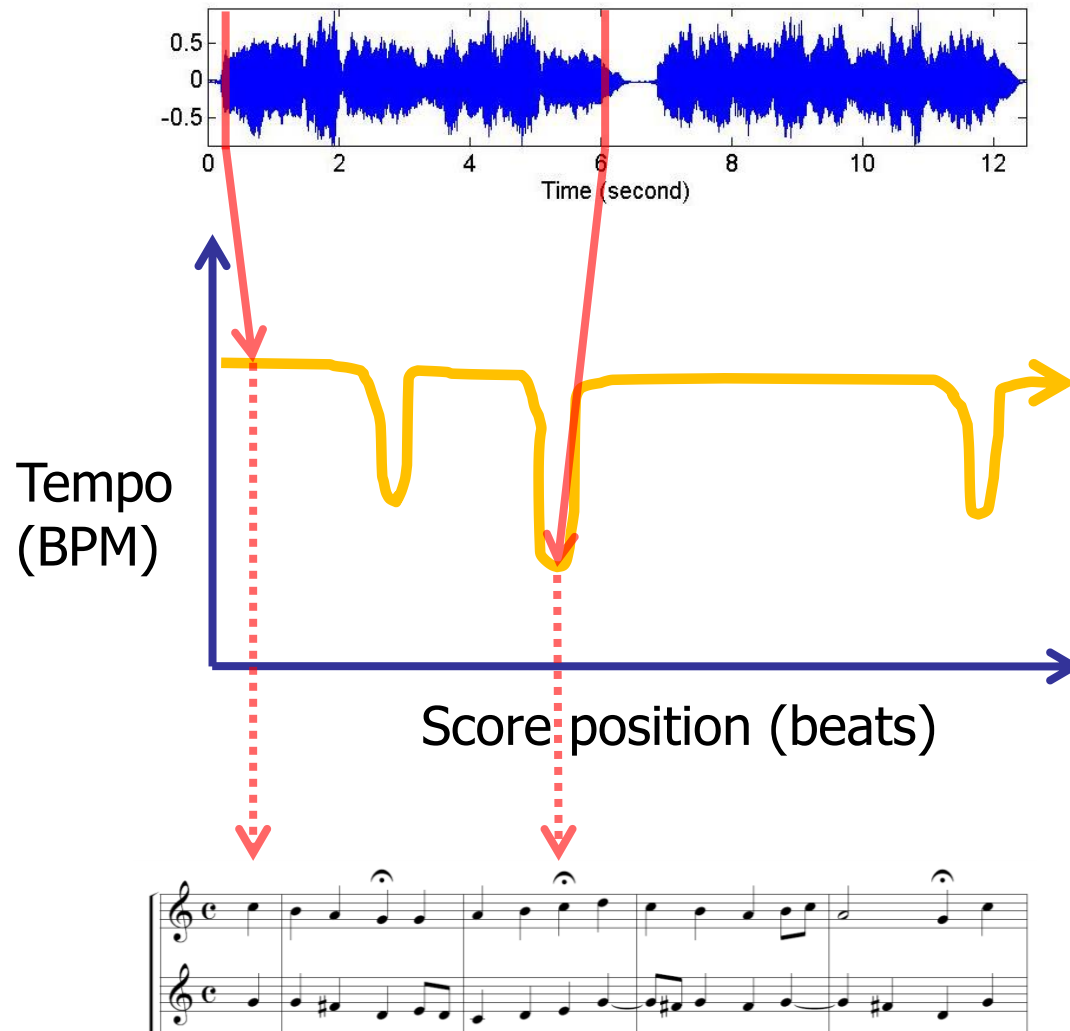
Discussions

- Advantages
 - “smart initialization” of W and H
 - Detailed timbre model using NMF
 - Onset modeling
- Disadvantages
 - May be hard to deal with multi-instrument polyphonic audio
 - The same note can have different pitch and timbre
 - How many dictionary elements do we need then?

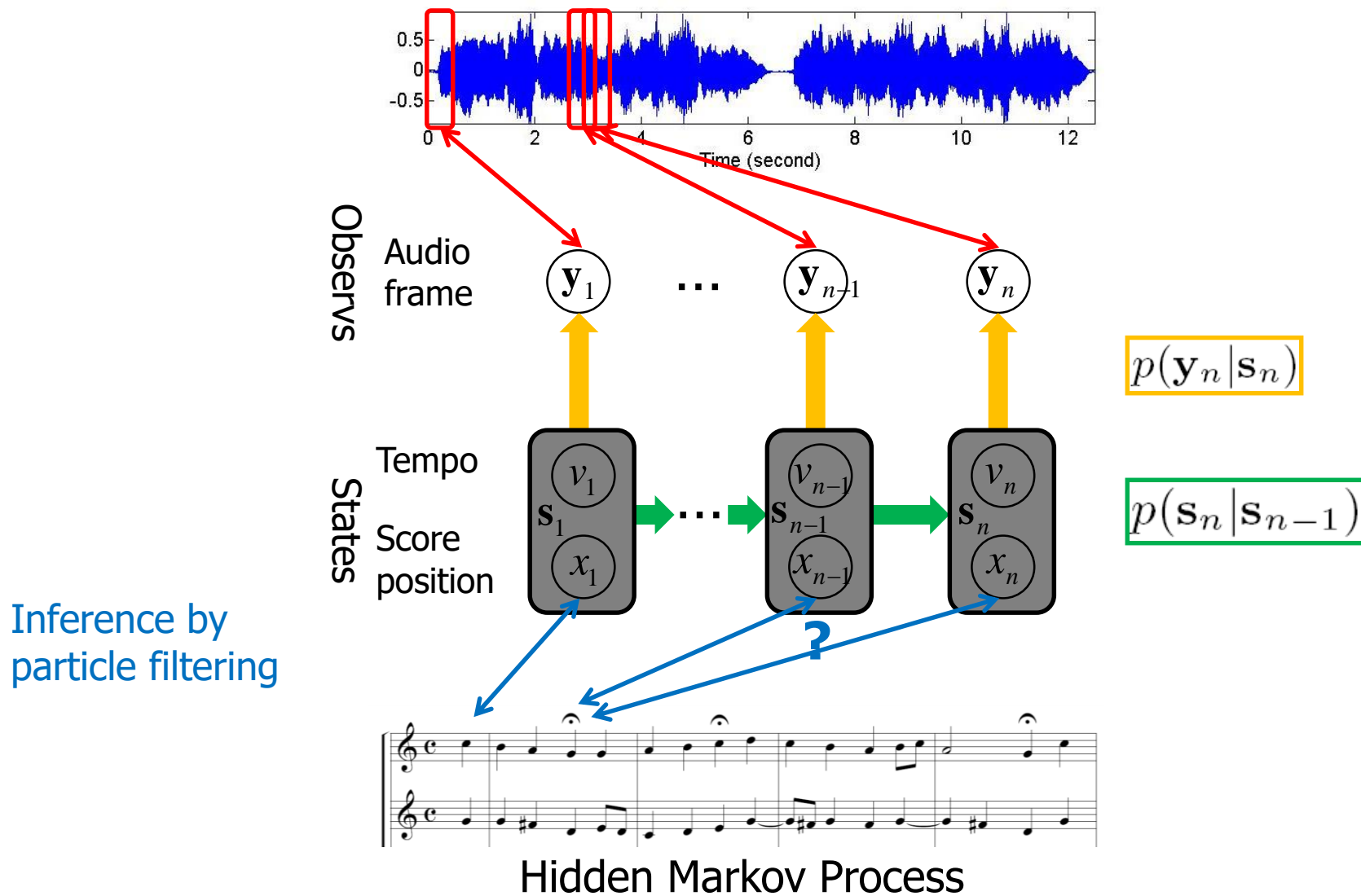
Soundprism

- Multi-pitch info of audio [Duan & Pardo, 2011]
- Particle filtering for alignment
- Pitch-based separation
- Online

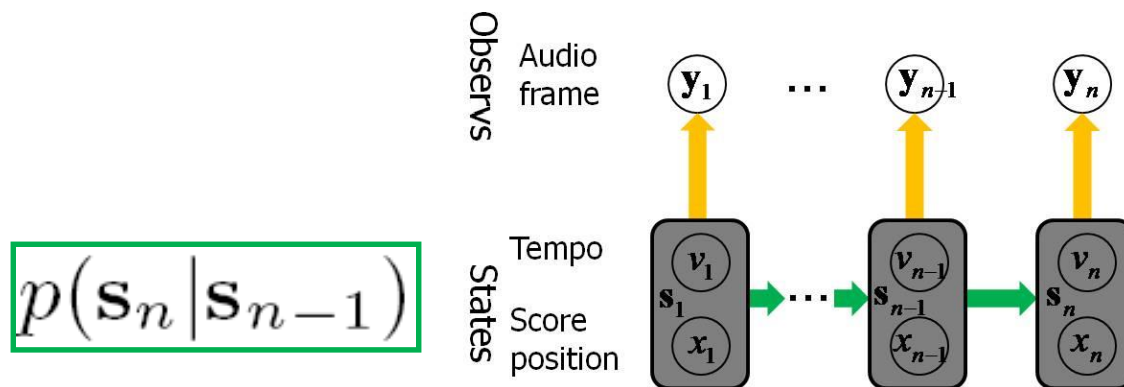
Align Audio with Score



A State Space Model



Transition Model



- Dynamical system

- Position: $x_n = x_{n-1} + l \cdot v_{n-1}$

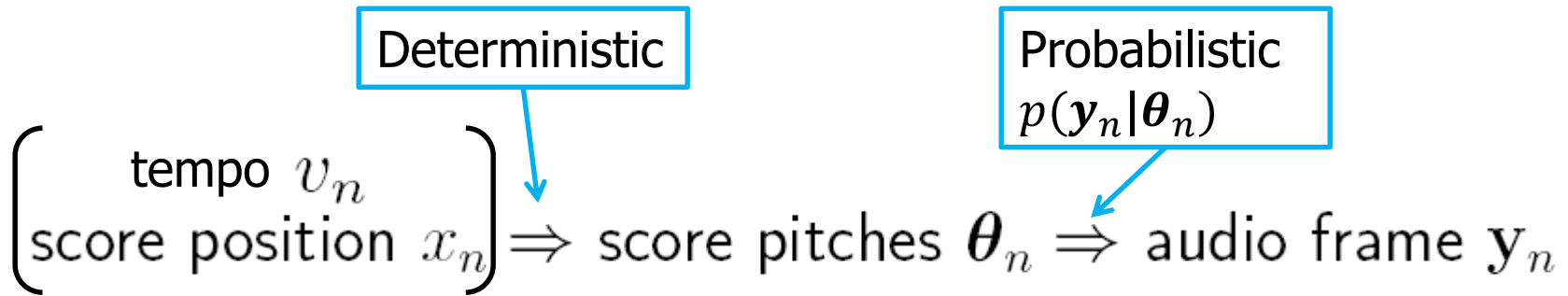
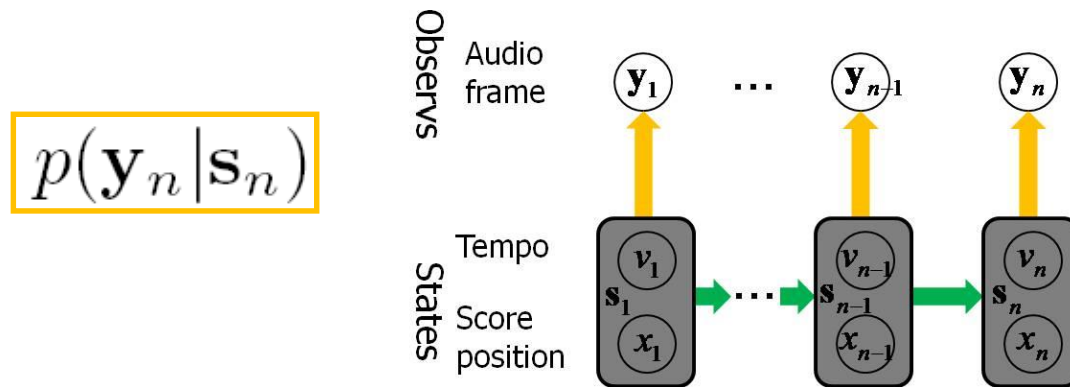
- Tempo: $v_n = \begin{cases} v_{n-1} + n_v & \text{If the score position } x_n \text{ just passed a score note onset} \\ v_{n-1} & \text{otherwise} \end{cases}$

where $n_v \sim \mathcal{N}(0, \sigma_v^2)$

If the score position x_n just passed a score note onset

otherwise

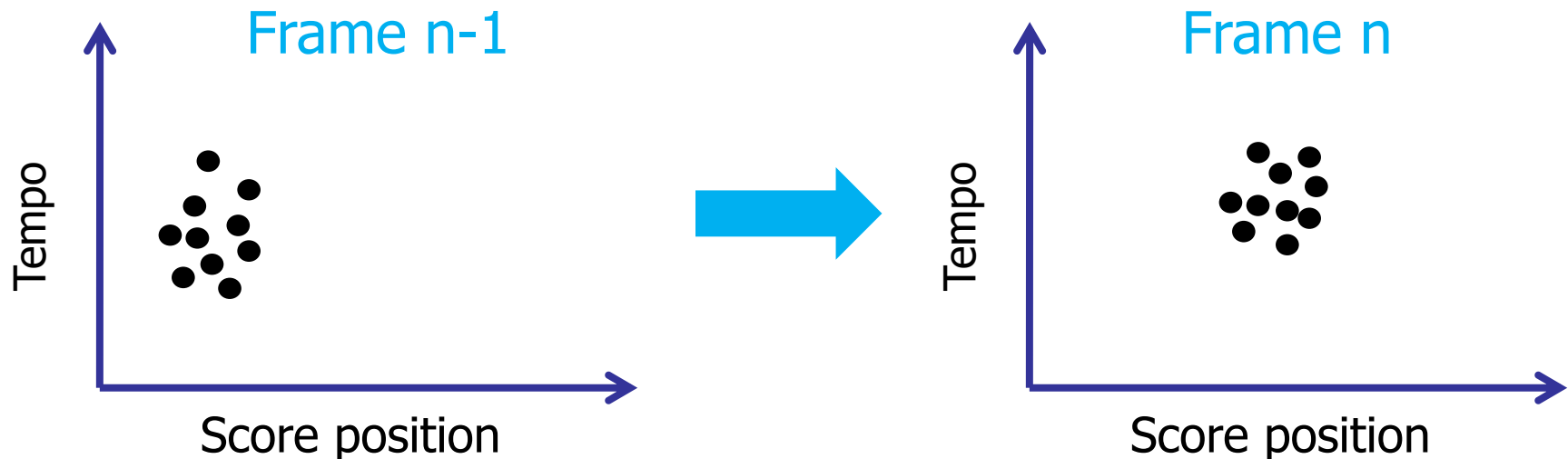
Observation Model



- $p(\mathbf{y}_n | \boldsymbol{\theta}_n)$ is the multi-pitch estimation model trained from thousands of random chords

Online Inference by Particle Filtering

- In n -th frame, estimate posterior $p(\mathbf{s}_n | \mathbf{Y}_{1:n})$ from past observations $\mathbf{Y}_{1:n} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$
- Update $p(\mathbf{s}_n | \mathbf{Y}_{1:n})$ from $p(\mathbf{s}_{n-1} | \mathbf{Y}_{1:n-1})$ with a fixed number of particles
 - Move by $p(\mathbf{s}_n | \mathbf{s}_{n-1})$ (i.e. the dynamic equations), resample by $p(\mathbf{y}_n | \mathbf{s}_n)$

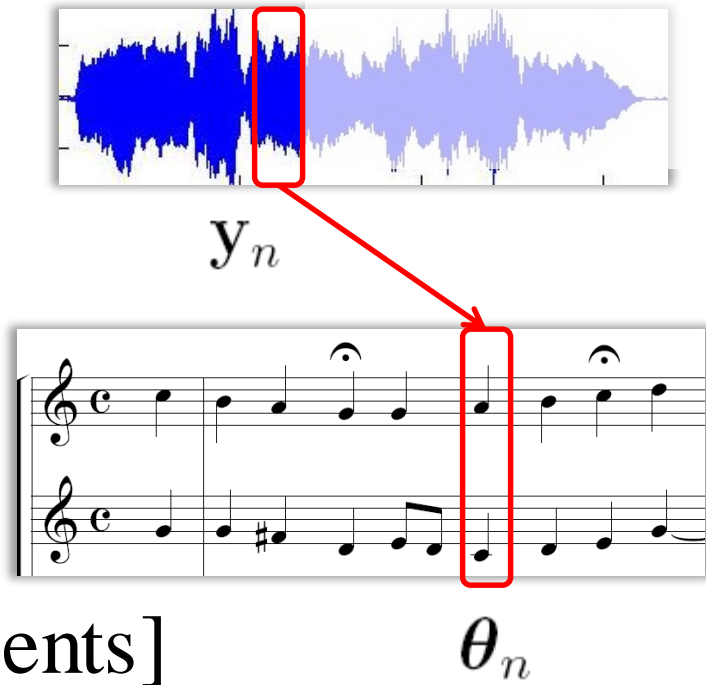


Source Separation

- 1. Accurately estimate performed pitches $\hat{\theta}_n$
 - Around score pitches θ_n

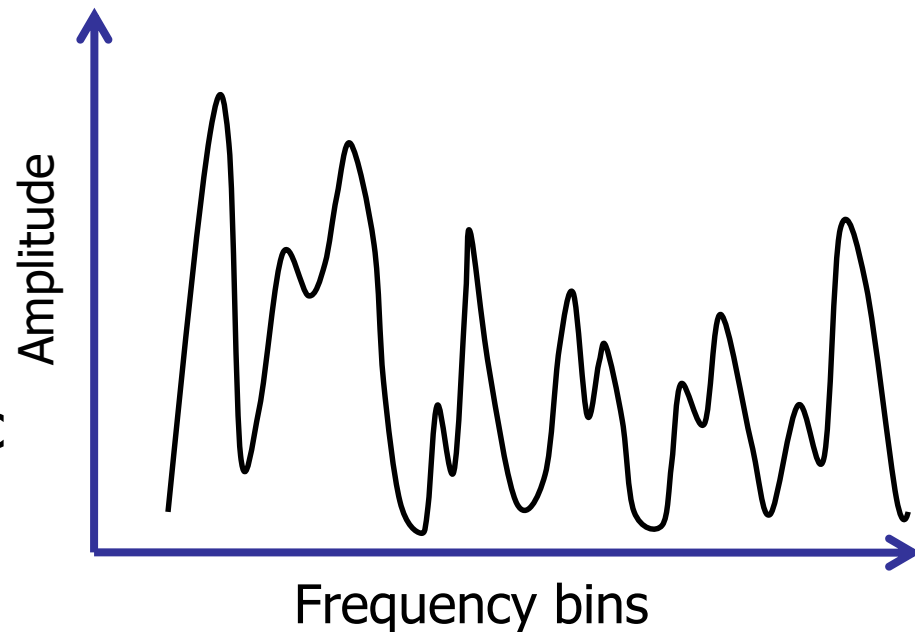
$$\hat{\theta}_n = \arg \max p(\mathbf{y}_n | \theta)$$

$$\text{s.t. } \theta \in [\theta_n - 50\text{cents}, \theta_n + 50\text{cents}]$$



Reconstruct Source Signals

- 2. Allocate mixture's spectral energy
 - Non-harmonic bins
 - To all sources, evenly
 - Non-overlapping harmonic bins
 - To the active source, solely
 - Overlapping harmonic bins
 - To active sources, in inverse proportion to the square of harmonic numbers
- 3. IFFT with mixture's phase to time domain



Harmonic positions for Source 1

0	1	0	1	0	1	0	1	0	1
0	0	1	0	0	1	0	0	1	0

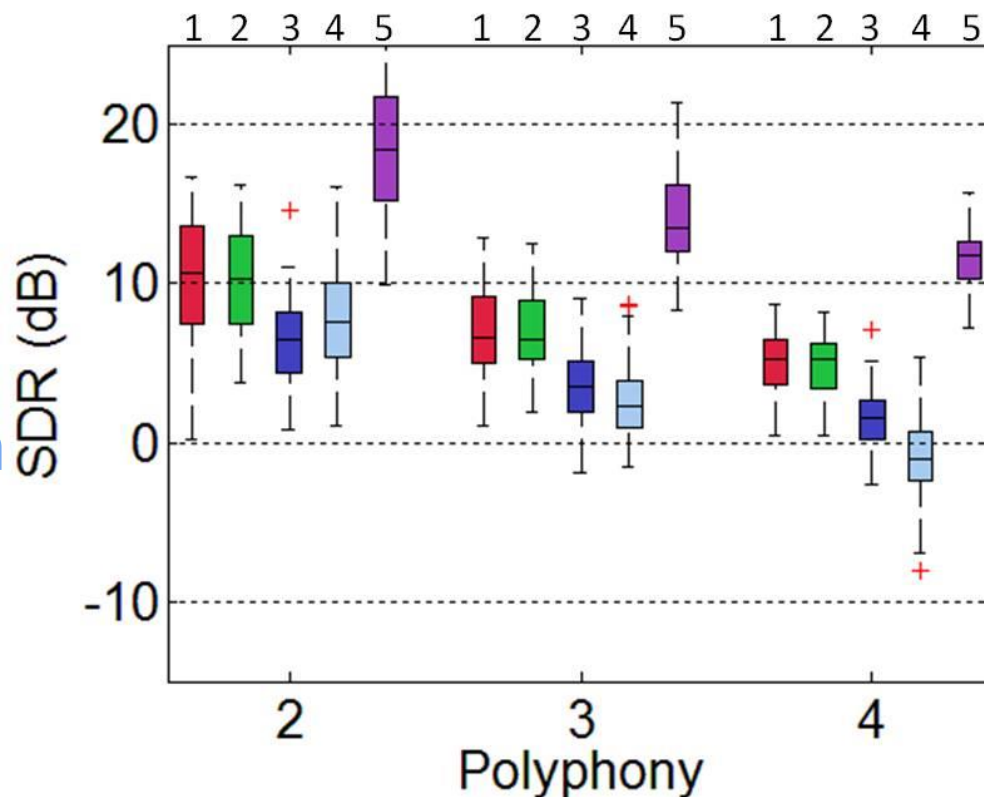
Harmonic positions for Source 2

Experiments

- 10 pieces of J.S. Bach 4-part chorales
- Audio played by violin, clarinet, saxophone and bassoon, separately recorded and then mixed.
- MIDI score downloaded
- Ground-truth alignment manually annotated
- 150 combinations = 40 solos + 60 duets + 40 trios + 10 quartets

Source Separation Results

- 1. Proposed
- 2. Ideally-aligned
- 3. Ganseman et al'10 (offline algorithm)
- 4. Multi-pitch estimation & streaming-based separation (without score)
- 5. Oracle



Average
input SDR

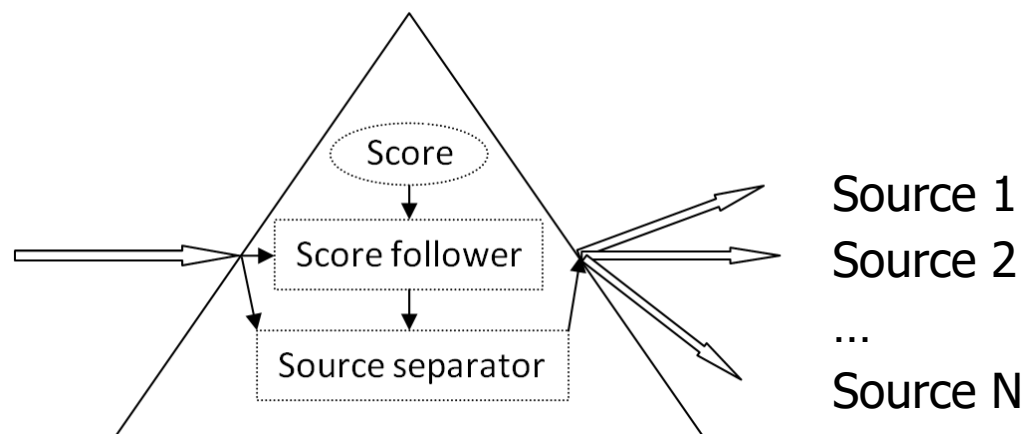
0dB

-3dB

-4.78dB

Soundprism

Single-channel
polyphonic music



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

Andantino.

p semplice

senza sord.

p senza sord.

p

Interactive Music Editing



Demo...

Discussions

- Advantages
 - Online system, potential for real-time applications
 - Can deal with multi-instrument polyphonic audio
 - Multi-pitch info is used
- Disadvantages
 - Multi-pitch model cannot distinguish different parts of a note
 - No onset modeling, alignment not precise
 - No timbre modeling in separation