Music Rhythm Analysis

Zhiyao Duan
Associate Professor of ECE and CS
University of Rochester

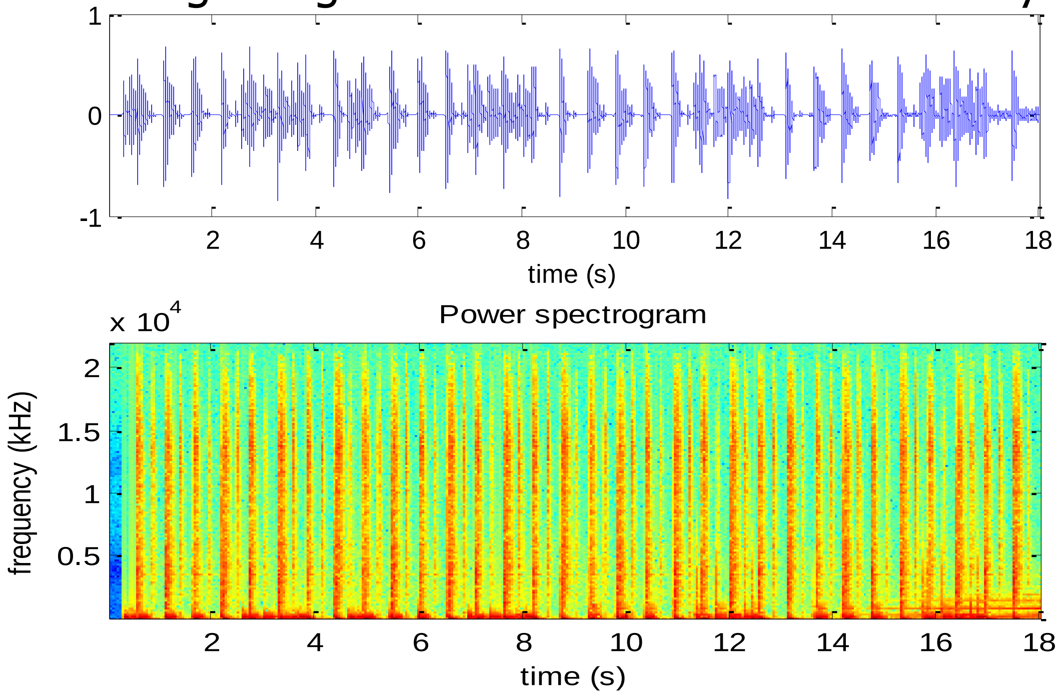
Presentation at WiSSAP 2023, IIT Kanpur, December 18-21, 2023

What is rhythm?

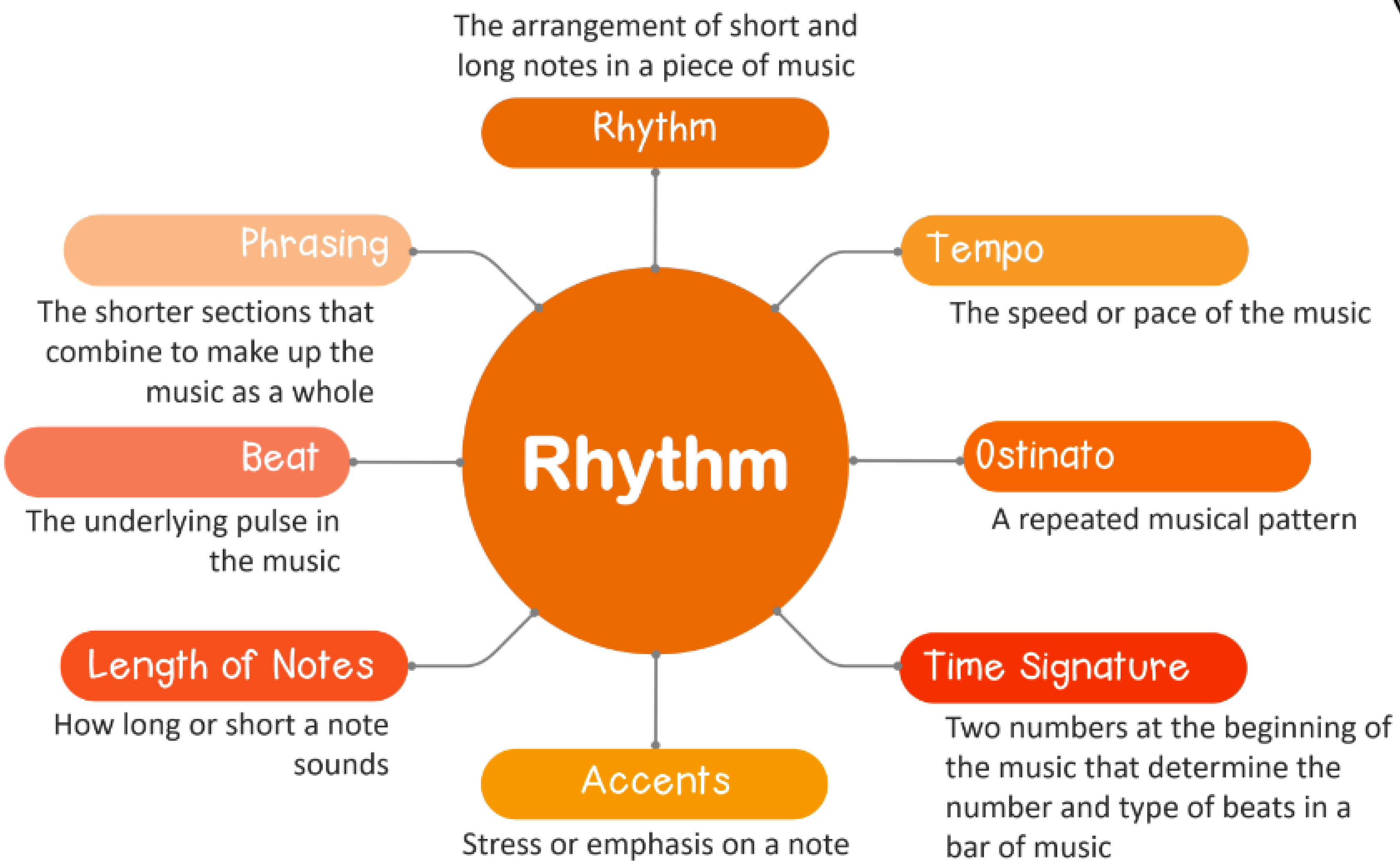
- “Movement marked by the regulated succession of strong and weak elements, or of opposite or different conditions.”

---- Oxford English Dictionary

Beginning of *Another one bites the dust* by



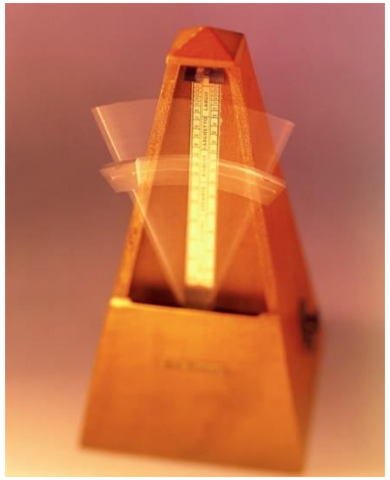
Rhythm in Music



From https://juliajooya.com/2020/12/21/what-is-rhythm-in-music/

Definitions

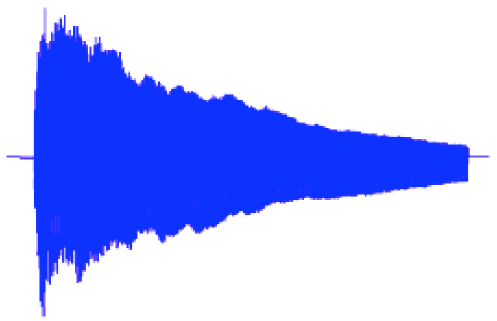
- **Beat:** basic unit of time in music
- **Tempo:** speed or pace of a given piece, typically measured in beats per minute (BPM)

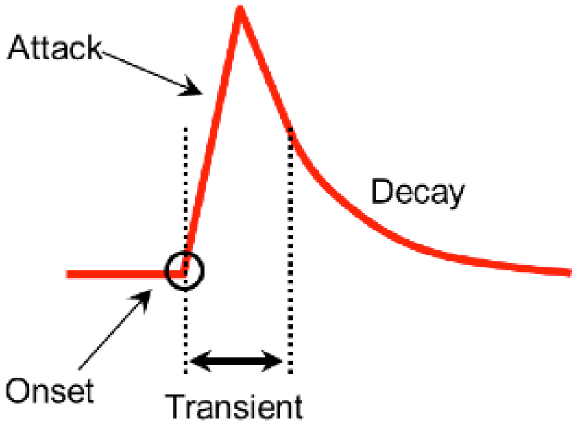


More Definitions

- **Attack**: sharp increase of energy
- **Transient**: a short duration with high amplitude within which signal evolves quickly
- **Onset**: single instant marking the beginning of transient
  - Onsets often occur on beats.

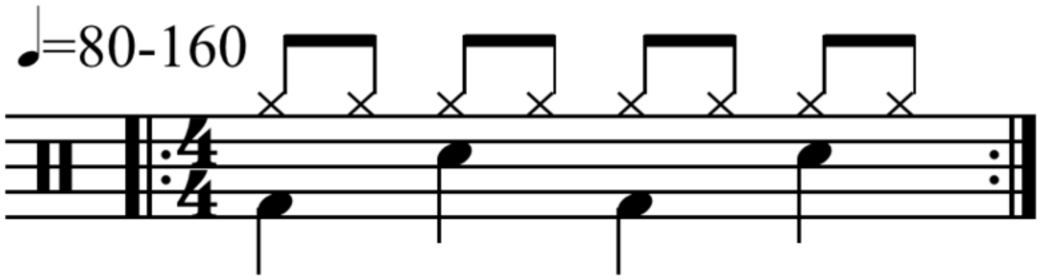
Waveform of one piano note





More Definitions

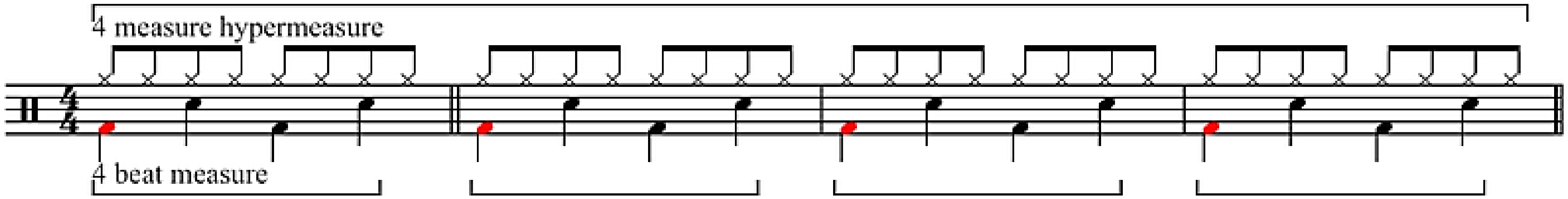
- **Measure (or bar):** segment of time defined by a given number of beats



A 4-beat measure drum pattern.
[http://en.wikipedia.org/wiki/Metre_(music)]

More Definitions

- **Meter:** organization of music into regularly recurring measures of stressed and unstressed beats



Hypermeter: 4-beat measure and 4-measure hypermeasure. Hyperbeats in red.
[http://en.wikipedia.org/wiki/Metre_(music)]

Rhythm Analysis Tasks

- Onset detection
- Tempo estimation
- Beat tracking
- Down beat tracking
- Meter estimation
- Hierarchical rhythm analysis

- Offline: algorithm accesses the entire song
- Online: algorithm only accesses past signals, good for real-time applications

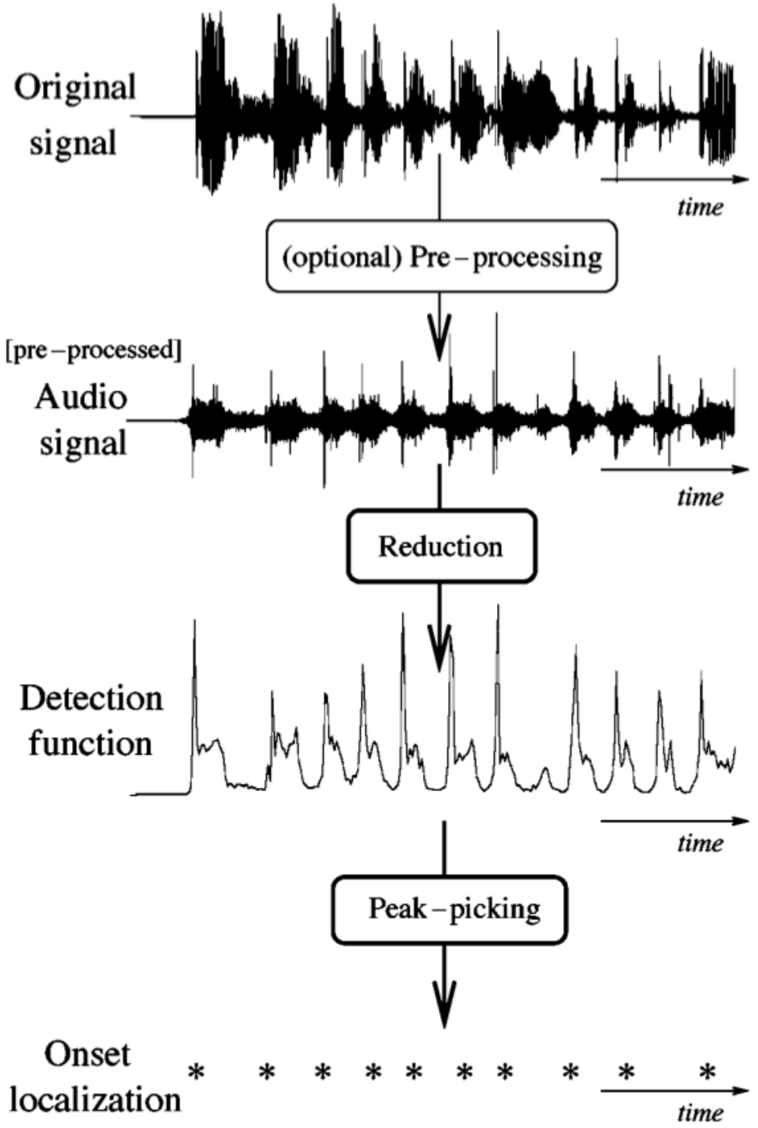
Why is it important?

- Intellectual merit
  - Important component of music understanding
  - Music cognition research
- Broad applications
  - Identify/classify/retrieve music by rhythmic similarity
  - Music segmentation/summarization
  - Audio/video synchronization
  - Source separation

Onset Detection

- Signal processing: define a detection function
  - Energy-based
  - Spectral-based
  - Phase-based

- Machine Learning: learn patterns from labeled data
  - Probabilistic models
  - Neural networks



Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M.B., "A tutorial on onset detection in music signals," *IEEE TSAP*, 2005.

Energy-based Onset Detection

Waveform

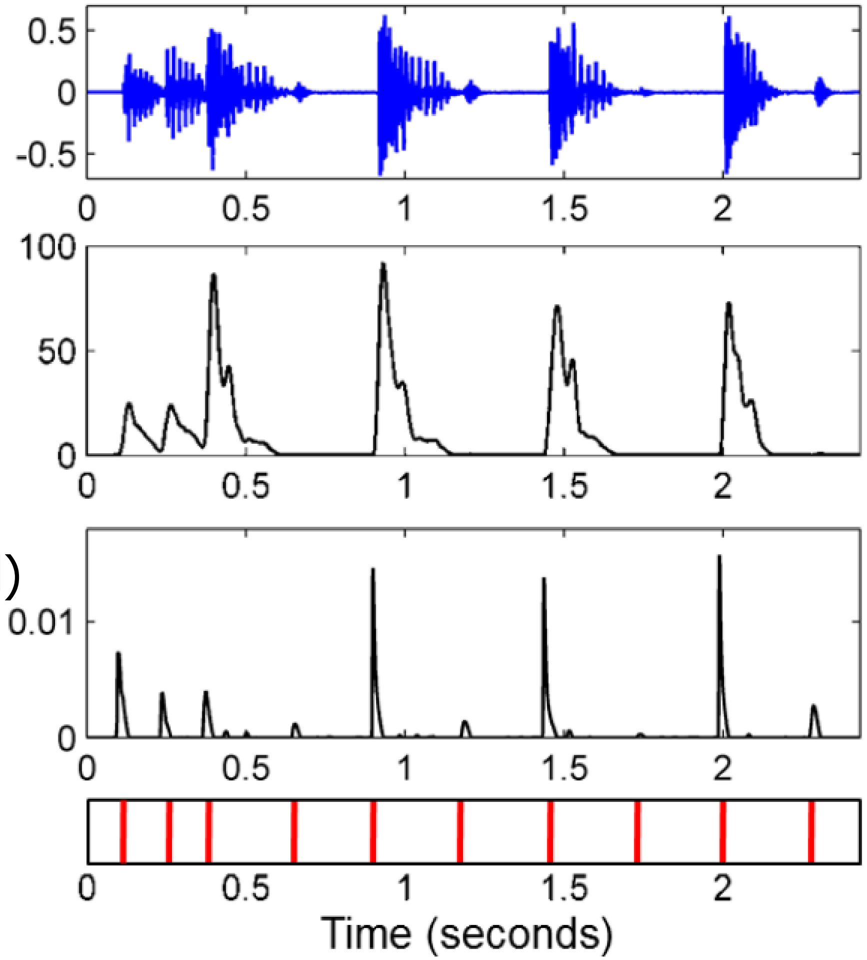
Signal Envelope (local energy)

$$E_w^x(n) := \sum_{m=-M}^{M} |x(n+m)w(m)|^2$$

Envelope Derivative (half-wave rectified)

$$\Delta_{\text{Energy}}(n) := |E_w^x(n+1) - E_w^x(n)|_{\geq 0}$$

Thresholding → Onsets



Note: the local energy is typically computed at a lower time resolution than audio samples, e.g., time frames.

Müller, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*, Springer, 2015.

Energy-based Onset Detection

- Pros and Cons
  - Simple
  - Works well for percussive sounds
  - Soft onsets by string/wind instruments are hard to detect
  - Tremolo/vibrato can cause false detections

- How to improve
  - Use logarithmic-energy to replace linear energy
  - Perform analysis in different frequency bands, then summarize

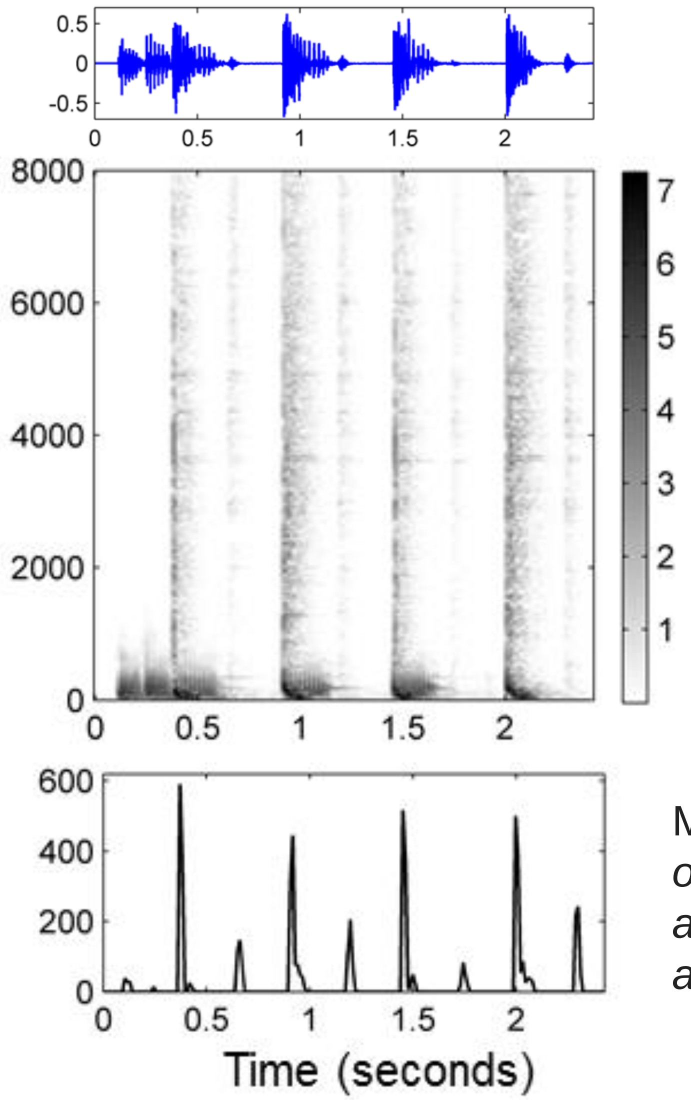
Spectral-based Onset Detection

- STFT to get magnitude spectrogram $|\chi|$
- (optional) compression

$$\mathcal{Y} := \Gamma_\gamma(|\mathcal{X}|) = \log(1 + \gamma \cdot |\mathcal{X}|)$$

- Spectral flux:
  - Take derivative w.r.t. time (half-wave rectified)

$$\Delta_{\text{Spectral}}(n) := \sum_{k=0}^{K} |\mathcal{Y}(n+1,k) - \mathcal{Y}(n,k)|_{\geq 0}$$



Müller, Meinard, *Fundamentals of music processing: Audio, analysis, algorithms, applications*, Springer, 2015.

Spectral-based Onset Detection

- Pros and Cons
  - More complex than energy-based
  - Can weigh different frequencies differently
  - Works better for soft onsets (e.g., legato notes) and polyphonic music
  - Still doesn’t work very well for vibrato

Examples

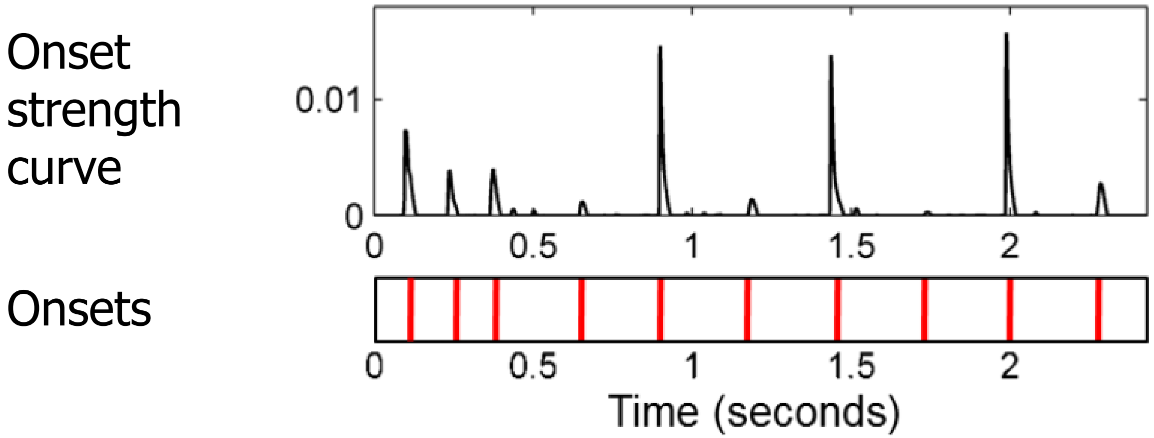
	Original audio	Original audio + energy-based onsets	Original audio + spectral-based onsets
Pop			
Schumann			
Hayden			

Tempo Estimation

- Tempo = beats / minutes
- Beat tracking is a sufficient but not a necessary condition for tempo estimation
- How to estimate tempo without tracking beats?

- Idea: look at the regularity of onsets
- Assumptions
  - Onsets mostly occur on beats
  - Tempo is constant within a period of time

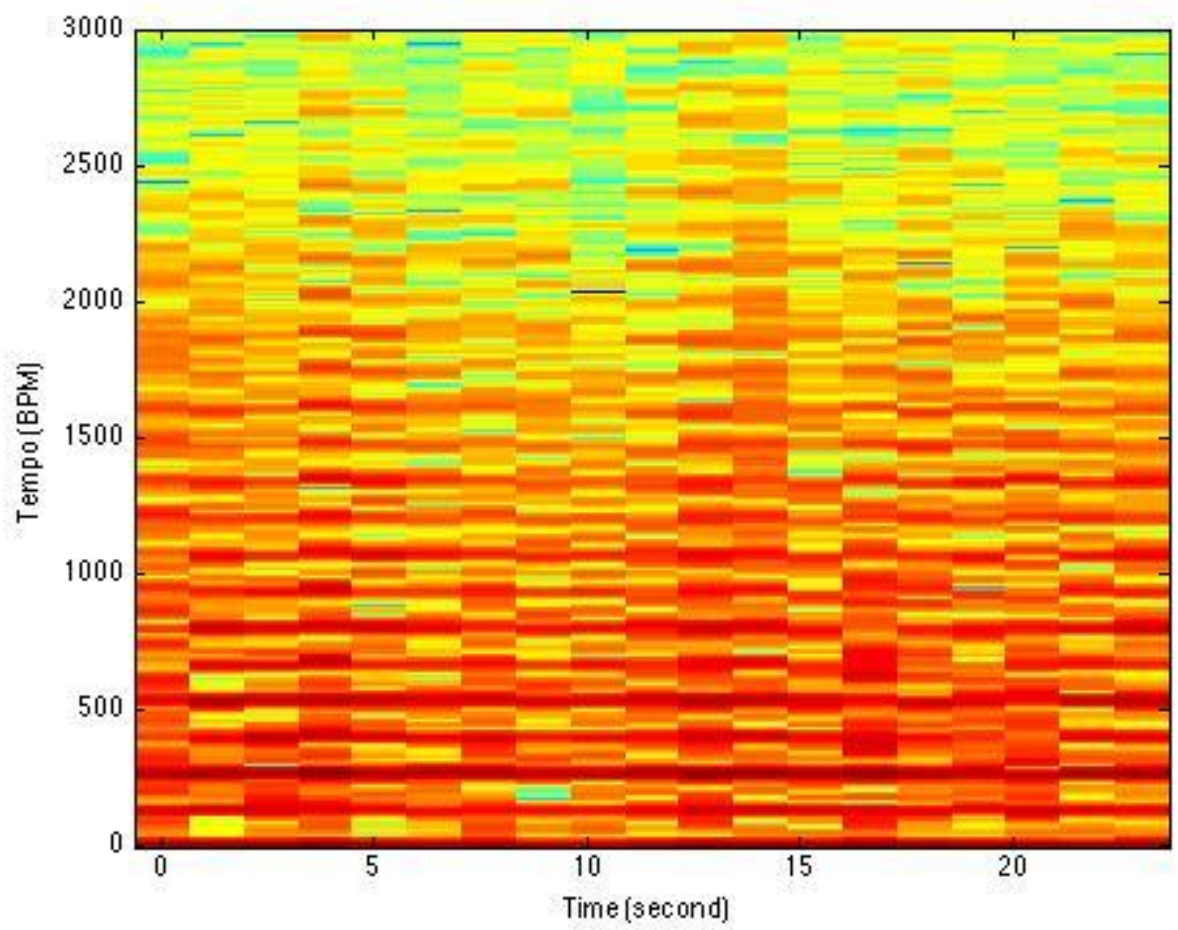
Tempo Estimation



- Take the onset strength curve and analyze its periodicity
  - Autocorrelation
  - STFT

Tempogram

- Perform STFT on the onset strength curve
- What is the fundamental frequency?



Beat Tracking

- Identify the beat times, i.e., the times to which we tap our feet
  - Detected onsets provide useful but noisy information, since not all onsets are on beats.
  - Estimated tempo tells us the space between two beats, but not the exact locations (i.e., phase).

- To simply the problem, we assume
  - Onsets, especially strong ones, are mostly on beats.
  - Tempo is constant.

Beat Tracking

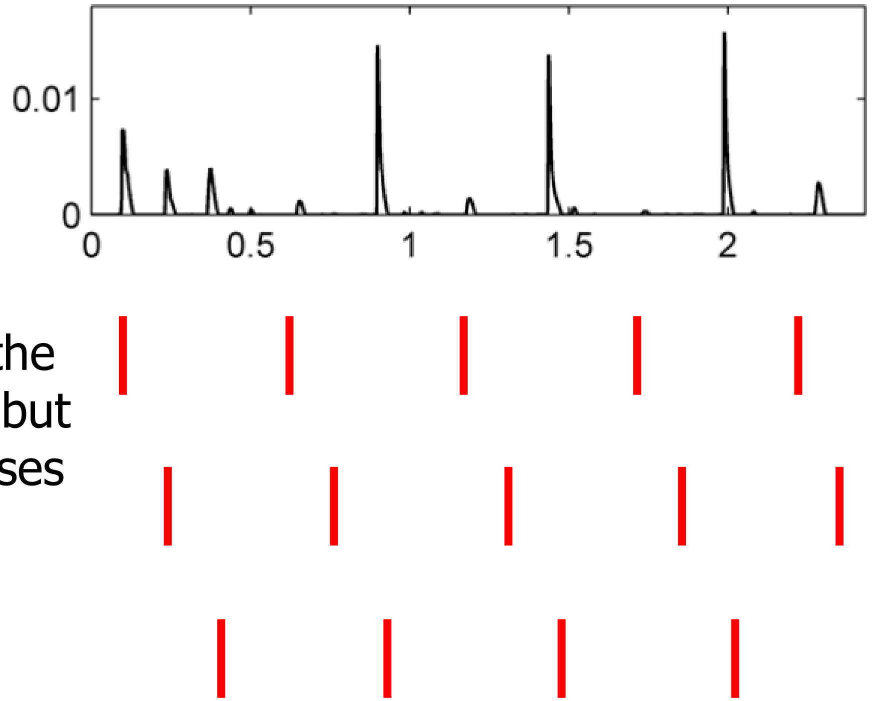
- A 2-step approach
  - Step 1: Tempo estimation
  - Step 2: Identify beats from onsets using the tempo
    - Create an impulse train (i.e., “comb”) with the tempo
    - Cross-correlate the “comb” with the onset strength curve.
    - The lag that gives us the highest cross-correlation value tells us the beat phase.

Beat Tracking

- A 2-step approach, illustration

Onset strength curve

Combs with the same tempo but different phases



- Problem: too rigid about beat spacing

Beat Tracking by Dynamic Programming

- Beat tracking: finding a sequence of beat locations such that they
  - 1) are well aligned with strong onsets
  - 2) mostly regularly spaced

Rough estimate of beat spacing

Score function

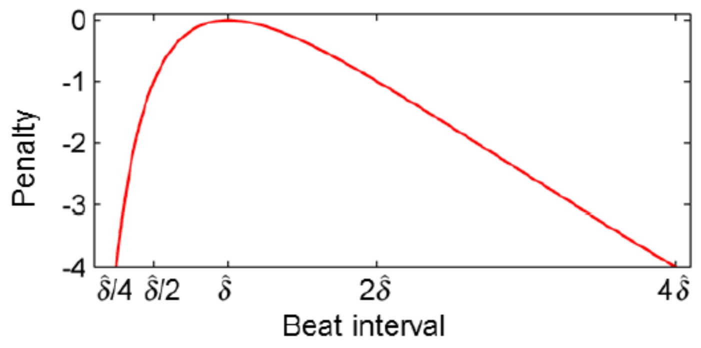
$$\mathbf{S}(B) := \sum_{\ell=1}^{L} \Delta(b_\ell) + \lambda \sum_{\ell=2}^{L} P_{\hat{\delta}}(b_\ell - b_{\ell-1})$$

Beat sequence

Onset strength

Regularity penalty function

$$P_{\hat{\delta}}(\delta) := -\left(\log_2(\delta/\hat{\delta})\right)^2$$



- Find $B = (b_1, b_2, \cdots, b_L)$ that maximizes $\boldsymbol{S}(B)$

Daniel PW Ellis, "Beat tracking by dynamic programming." *Journal of New Music Research* 36.1 (2007): 51-60.

Beat Tracking by Dynamic Programming

- Suppose beat locations are precise to audio frames, and suppose there are N frames, then how many possible sequences?
  - 2^N (although many are bad ones!)
  - Can't enumerate all!

- Key idea: reuse calculations by recursion!

Beat Tracking by Dynamic Programming

- Consider a beat sequence $B_{n,L} = (b_1, b_2, \cdots, b_L)$ where $b_L = n$.
- Let $D(n)$ be the maximal score among all sequences that end at n with various lengths.
- Assume B_n^* is the optimal sequence, then it can be constructed from optimal subsequences $B_{n-1,L-1}^*$ or $B_{n-2,L-1}^*$ or $B_{n-3,L-1}^*$ …, by appending one more beat, and choosing the best one.
- Then

if $L = 1$

$$\mathbf{D}(n) = \Delta(n)$$

recursion

if $L > 1$

$$\mathbf{D}(n) = \Delta(n) + \lambda P_{\hat{\delta}}(n - b_{L-1}) + \boxed{\mathbf{D}(b_{L-1})}$$

Beat Tracking by Dynamic Programming

- Considering the two cases, we have

$$\mathbf{D}(n) = \Delta(n) + \max \begin{cases} 0, \\ \max_{m \in [1:n-1]} \{\mathbf{D}(m) + \lambda P_{\hat{\delta}}(n-m)\} \end{cases}$$

  – We can calculate $D(n)$ from $D(1) = \Delta(1)$.

- Record the preceding beat

$$\mathbf{P}(n) := \underset{m \in [1:n-1]}{\operatorname{argmax}} \{\mathbf{D}(m) + \lambda P_{\hat{\delta}}(n-m)\}$$

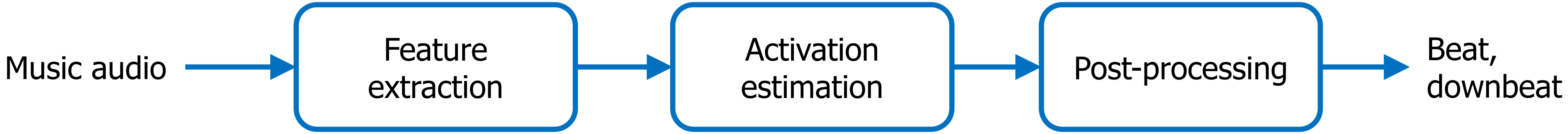
- Best score

$$\mathbf{S}(B^*) = \max_{n \in [0:N]} \mathbf{D}(n)$$

Pop + beats

- Trace back from $b_L = n^*$ to get the best sequence

Deep Learning Based



- Feature extraction: log-mel spectrogram, CQT spectrogram, chromagram, onset strength curve, etc.

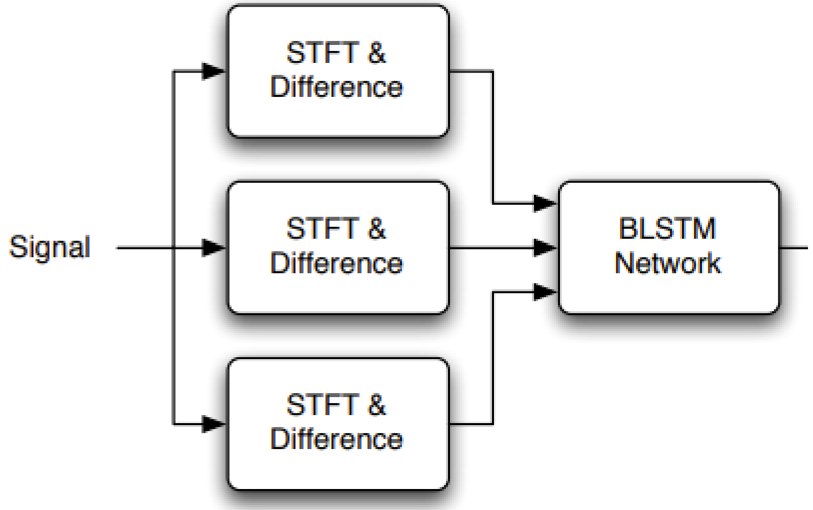
- Activation estimation: using neural networks (e.g., RNN, CRNN, Transformers) to estimate the salience of beats or downbeats

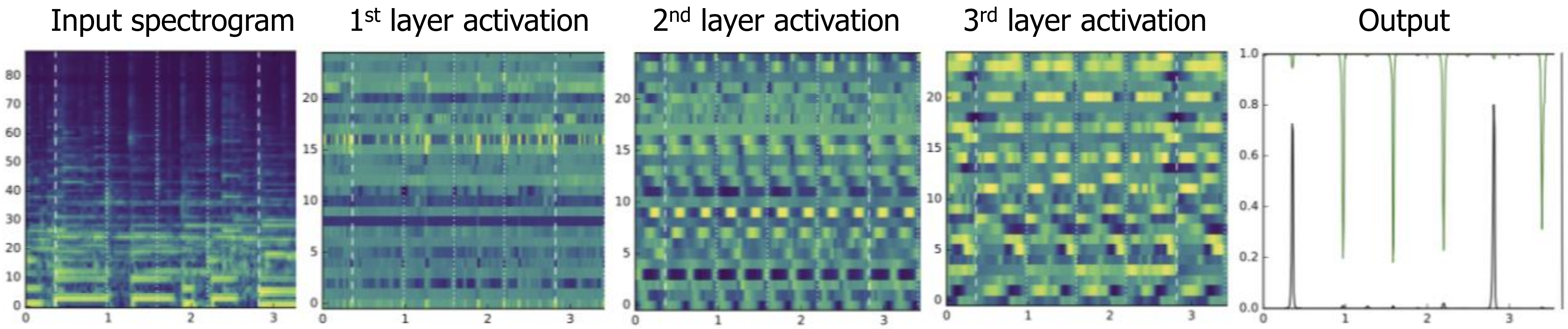
- Post-processing: make the final (binarized) decision for beats or downbeats

Joint Beat/Downbeat Tracking (offline)

S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *Proc. ISMIR,* 2016.

- Feature extraction: STFT and temporal difference at 3 temporal resolutions
- Activation estimation: 3 layers of Bidirectional LSTM with 3 outputs through softmax (beat, downbeat, non-beat) for each audio frame

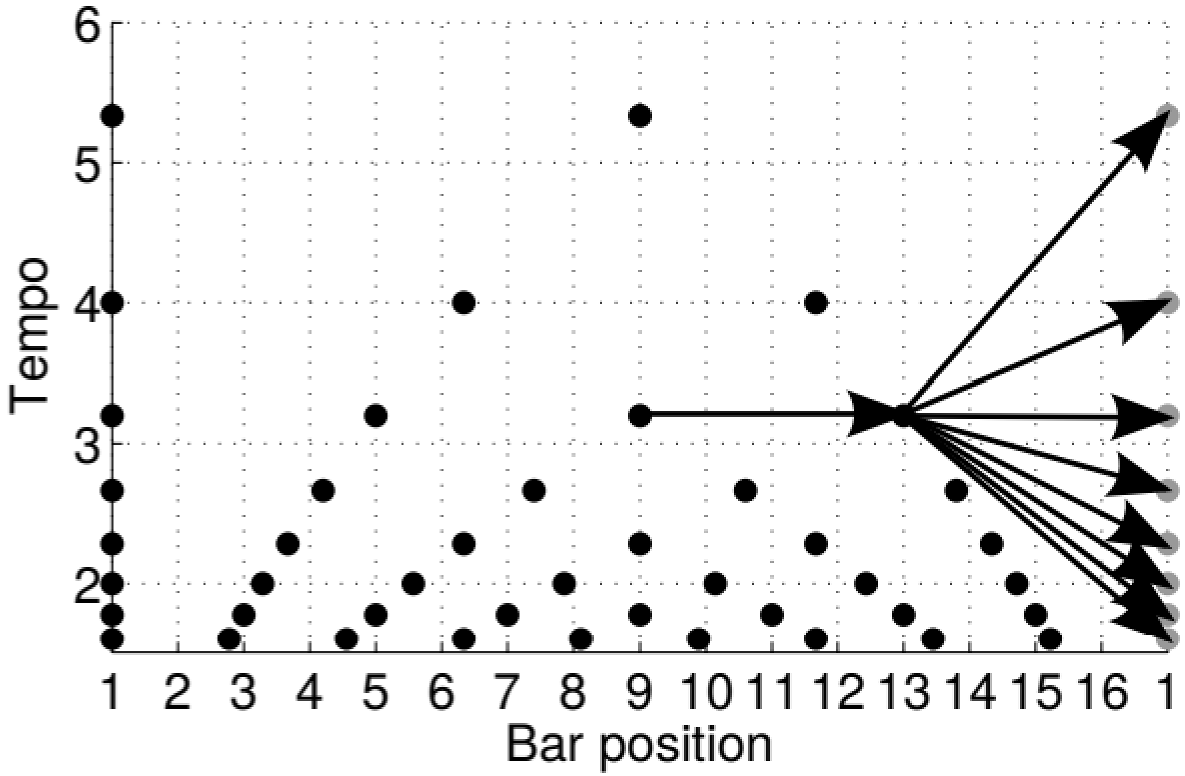




Joint Beat/Downbeat Tracking (offline)

Implemented in Madmom:
https://github.com/CPJKU/madmom

- Post-processing: dynamic Bayesian network
- State space
  - Transition model shown in figure
  - Observation model using activation values estimated from neural network

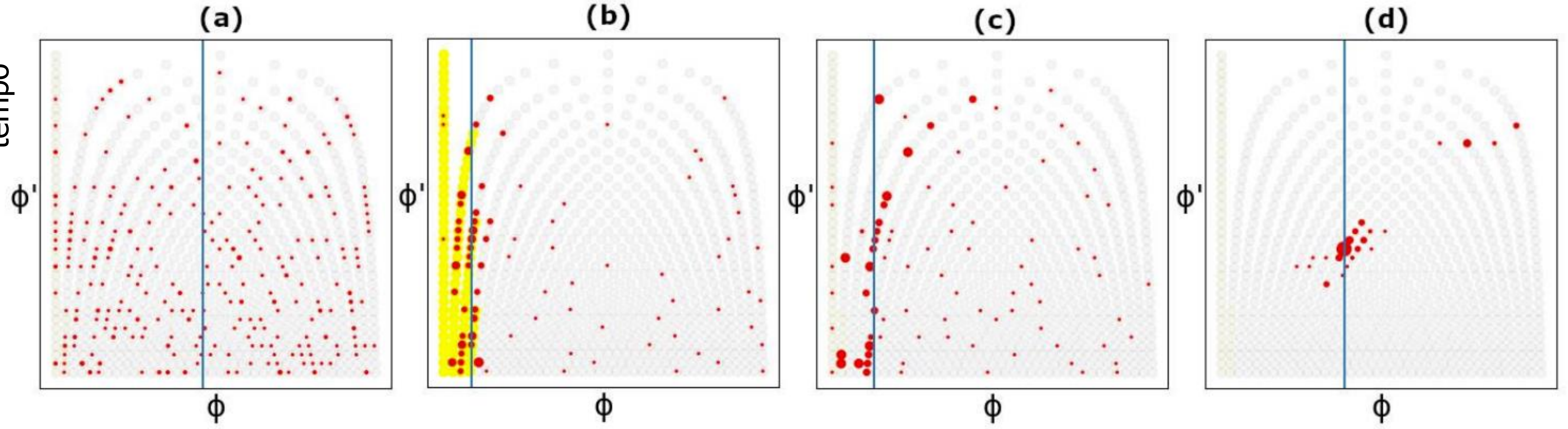


F. Krebs, S. Bock, and G. Widmer. An Efficient State Space Model for Joint Tempo and Meter Tracking. In *Proc. ISMIR*, 2015.

Beat Tracking (online)

- Key idea: using particle filtering to replace dynamic Bayesian network inference

tempo

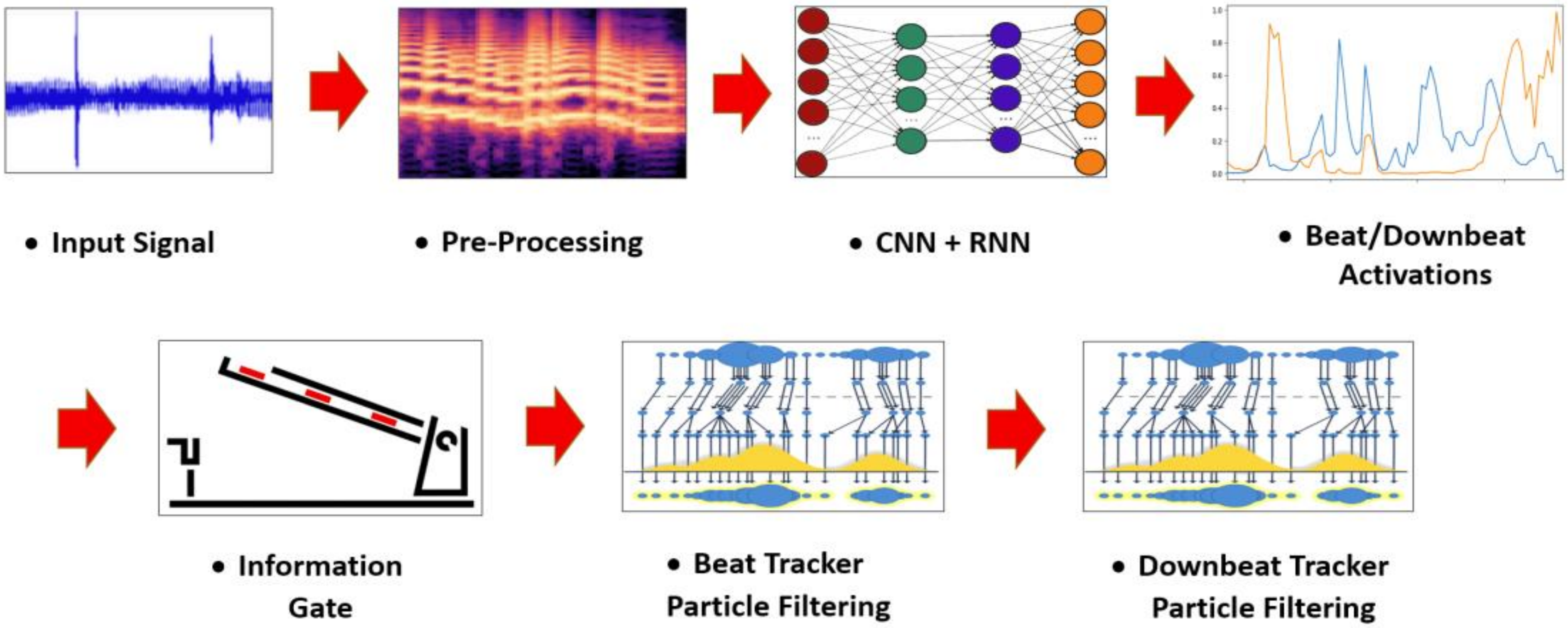


Beat phase

Mojtaba Heydari and Zhiyao Duan, ”Don't look back: An online beat tracking method using RNN and enhanced particle filtering,” in by *Proc. ICASSP 2021.*

Joint Beat & Downbeat Tracking (online)

- First real-time joint beat/downbeat/meter tracking model



Mojtaba Heydari, Frank Cwitkowitz, and Zhiyao Duan, “BeatNet: A real-time music integrated beat and downbeat tracker,” in *Proc. ISMIR*, 2021.

Evaluation of Online Beat Tracking

Training and test data combined

Dataset	*# Files*	*Total Length*
Ballroom [33, 34]	685	5 h 57 m
Beatles [2]	180	8 h 9 m
Carnatic [35]	176	16 h 38 m
GTZAN [36, 37]	999	8 h 20 m
Rock Corpus [38]	200	12 h 53 m

BeatNet repo: https://github.com/mjhydri/BeatNet

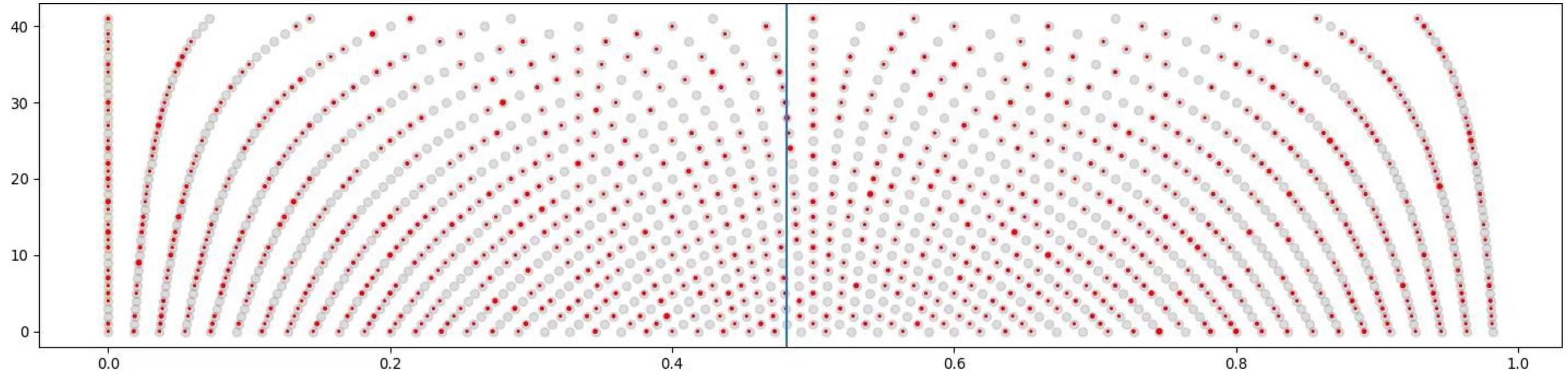
Method	*F-Measure Beats*	*F-Measure Downbeats*
Comparison of Online Methods		
GTZAN Dataset		
Aubio	57.09	—
BeatNet	$\boxed{\underline{75.44}}$	$\boxed{\underline{46.49}}$
Böck ACF	64.63	—
Böck FF	74.18	—
DLB	73.77	—
IBT	68.99	—
Ballroom Dataset		
Aubio	56.73	—
BeatNet	$\boxed{\underline{77.41}}$	$\boxed{\underline{47.45}}$
IBT	70.79	—
Rock Corpus Dataset		
Aubio	59.83	—
BeatNet	$\boxed{\underline{73.13}}$	$\boxed{\underline{44.98}}$
IBT	68.55	—
Comparison of Offline Methods		
GTZAN Dataset		
BeatNet + DBN	<u>80.64</u>	<u>54.07</u>
Böck	79.09	51.36

Singing Voice Beat/Downbeat Tracking (online)

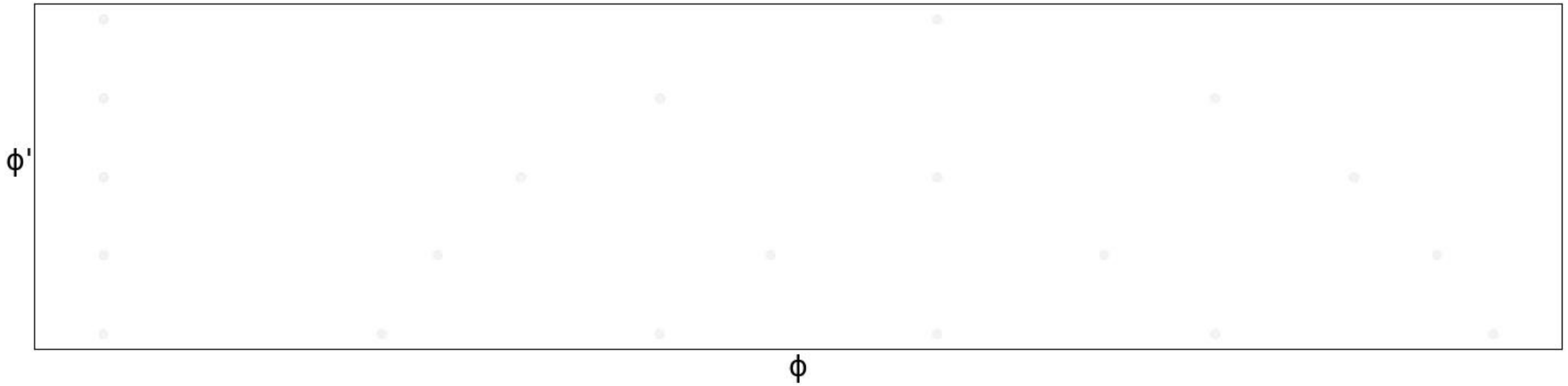
- Similar architecture to BeatNet
  - Using speech self-supervised learning features (wavLM)

beat phase

tempo



#beats/bar



bar phase

Mojtaba Heydari, Ju-Chiang Wang, and Zhiyao Duan, SingNet: a real-time singing voice beat and downbeat tracking system, in *Proc. ICASSP*, 2023.

Summary

- Important tasks
  - Onset detection
  - Beat tracking
  - Downbeat tracking
- Signal processing vs. machine learning
- Online vs. offline

- ISMIR 2019 Tutorial on “Tempo, Beat, and Downbeat Estimation”: https://tempobeatdownbeat.github.io/tutorial/intro.html


Thank You!