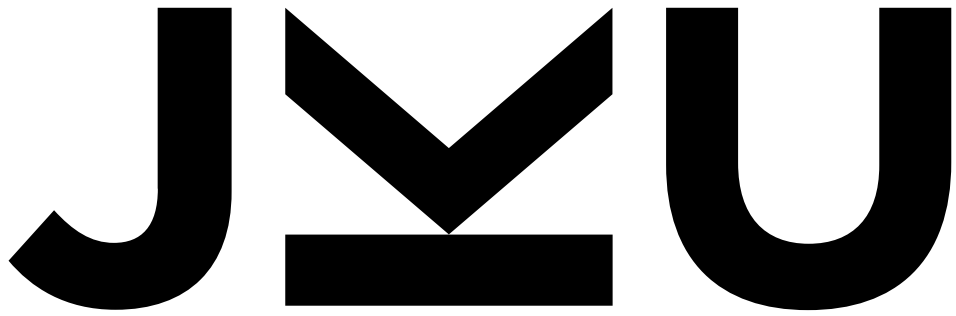
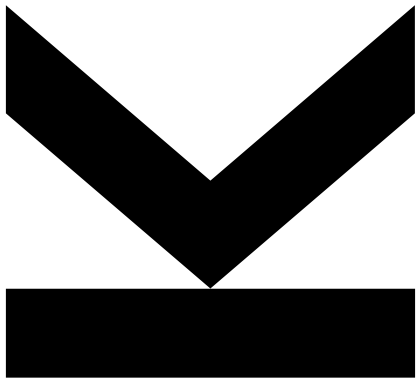


JOHANNES KEPLER
UNIVERSITY LINZ

SPECIAL TOPICS



Audio and Music Processing - Lecture 4: Onsets
344.032
KV, 2h, SS2020

Jan Schlüter
Institute of Computational Perception

OVERVIEW

- goal
  - becoming familiar with onset detection approaches
- topics
  - what are onsets?
  - building blocks
  - preprocessing
  - detection functions
  - peak picking
  - state-of-the-art approaches
  - evaluation and comparison

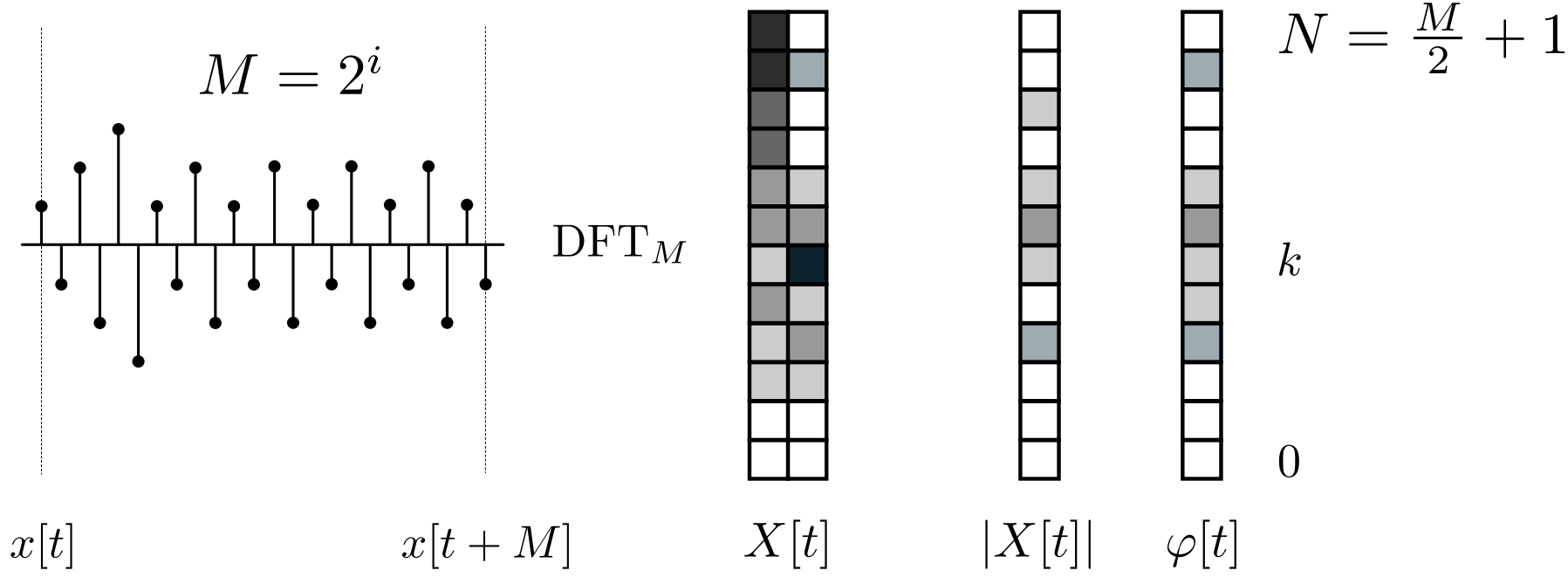
REFRESHER

JKU

NOTATION AND REFRESHER (1)

- $x[t]$ denotes **a discrete audio signal**, at time index t
- **offsets** in the signal are denoted as $x[t \pm k]$
- $X[t]$ is the DFT of length M that **starts** at $x[t]$
- $X[t]$ is a **complex** signal
- $|X[t]|$ is a **real** signal - the **magnitude spectrum**
- $\varphi[t]$ is a **real** signal - the **phases** of $|X[t]|$
- $|X_k[t]|$ denotes the **bin** k in the spectrum
- $\varphi_k[t]$ denotes the **phase** k in the spectrum
- in e^{ix} or e^{jx}, i or j refers to the imaginary number

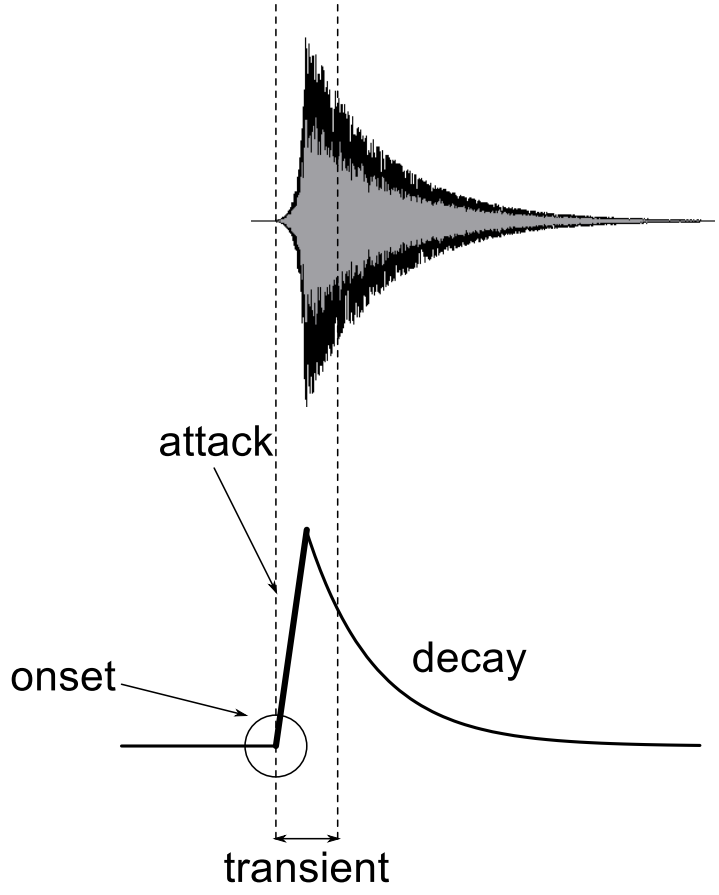
NOTATION AND REFRESHER (2)



BASICS

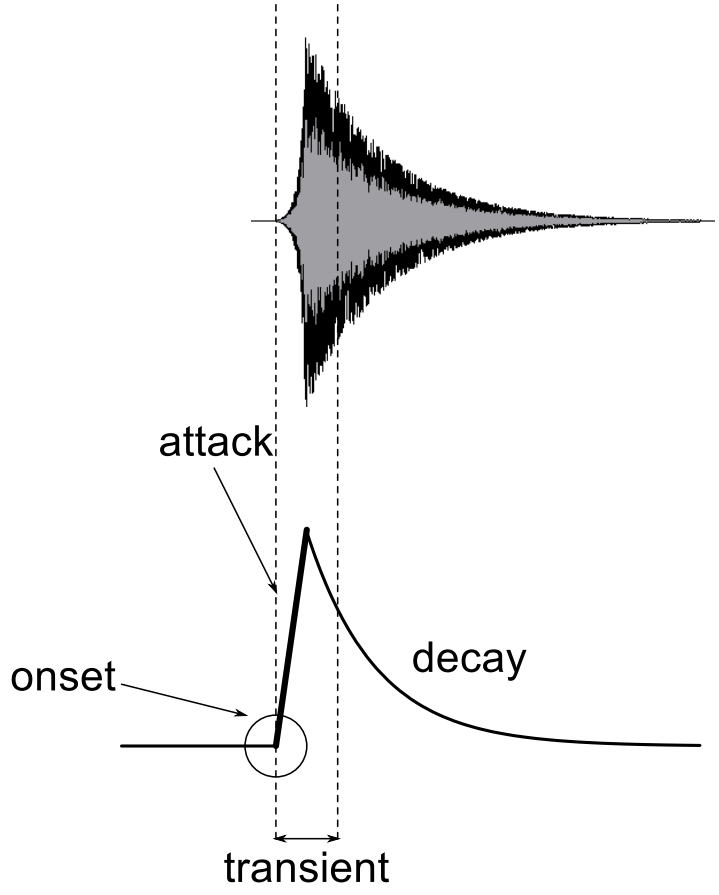
DEFINITIONS (1)

- an **onset** is defined as the **start** of a musical note (or any other sound)
- a **transient** is a **momentary variation** in current, voltage, frequency, and in our case air pressure measurements
- the **starting point** of a transient is chosen as the onset



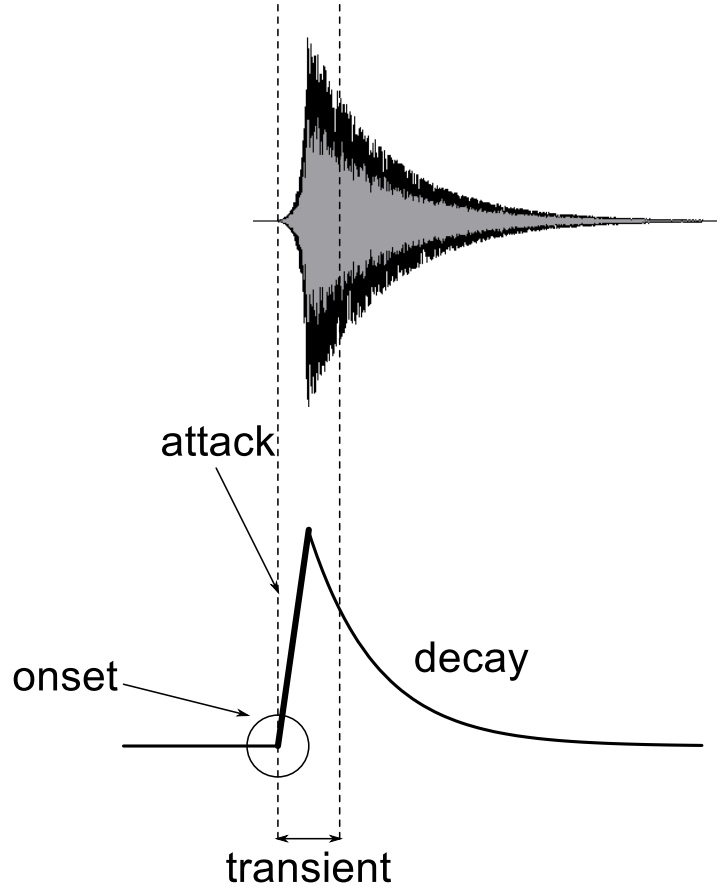
DEFINITIONS (2)

- the amplitude envelope of an instrument sound can be divided into phases
- the **attack** is the part of the sound where the amplitude envelope rises from zero to its maximum
- the **decay** is the part where the amplitude of the sound decays back to zero
- the **onset** is the beginning of the attack phase



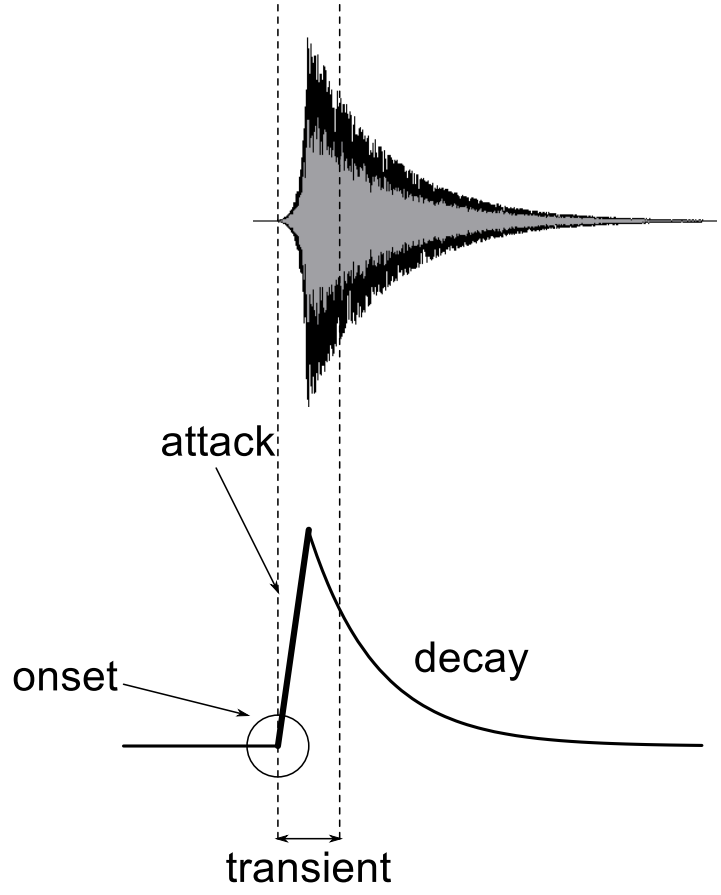
DEFINITIONS (3)

- in audio, the transient refers to a short, non-harmonic phase marking the start of a sound
- mostly non-periodic and unpredictable high-frequency components can be observed

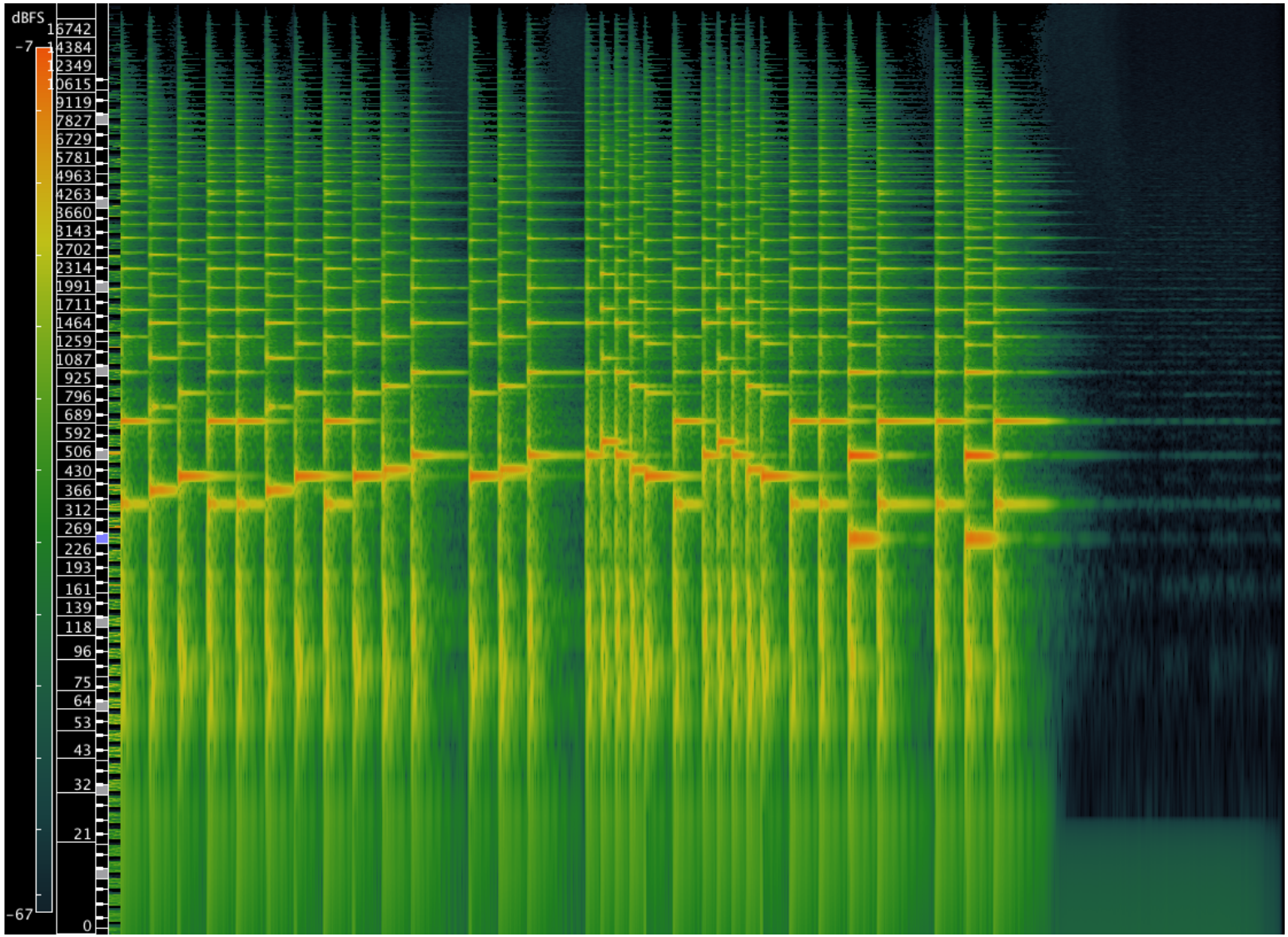


DEFINITIONS (4)

- the **true** position of an onset may very well be a **subjective** matter, and depend on when humans can reliably detect a transient
- soft, harmonic sounds (such as those produced by most string instruments) do not necessarily have clear transients, so the onsets are harder to detect

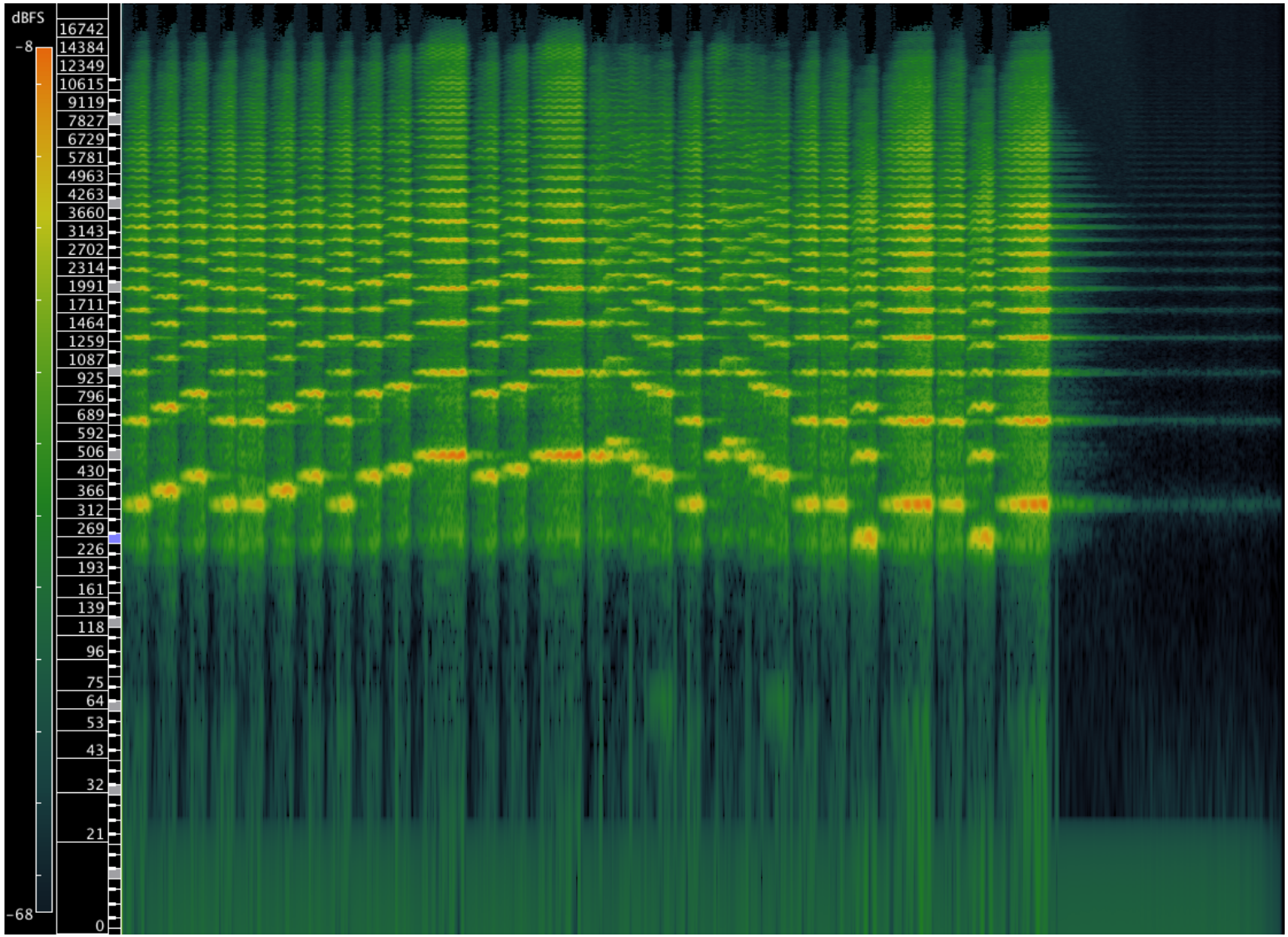


EX: PITCHED PERCUSSIVE



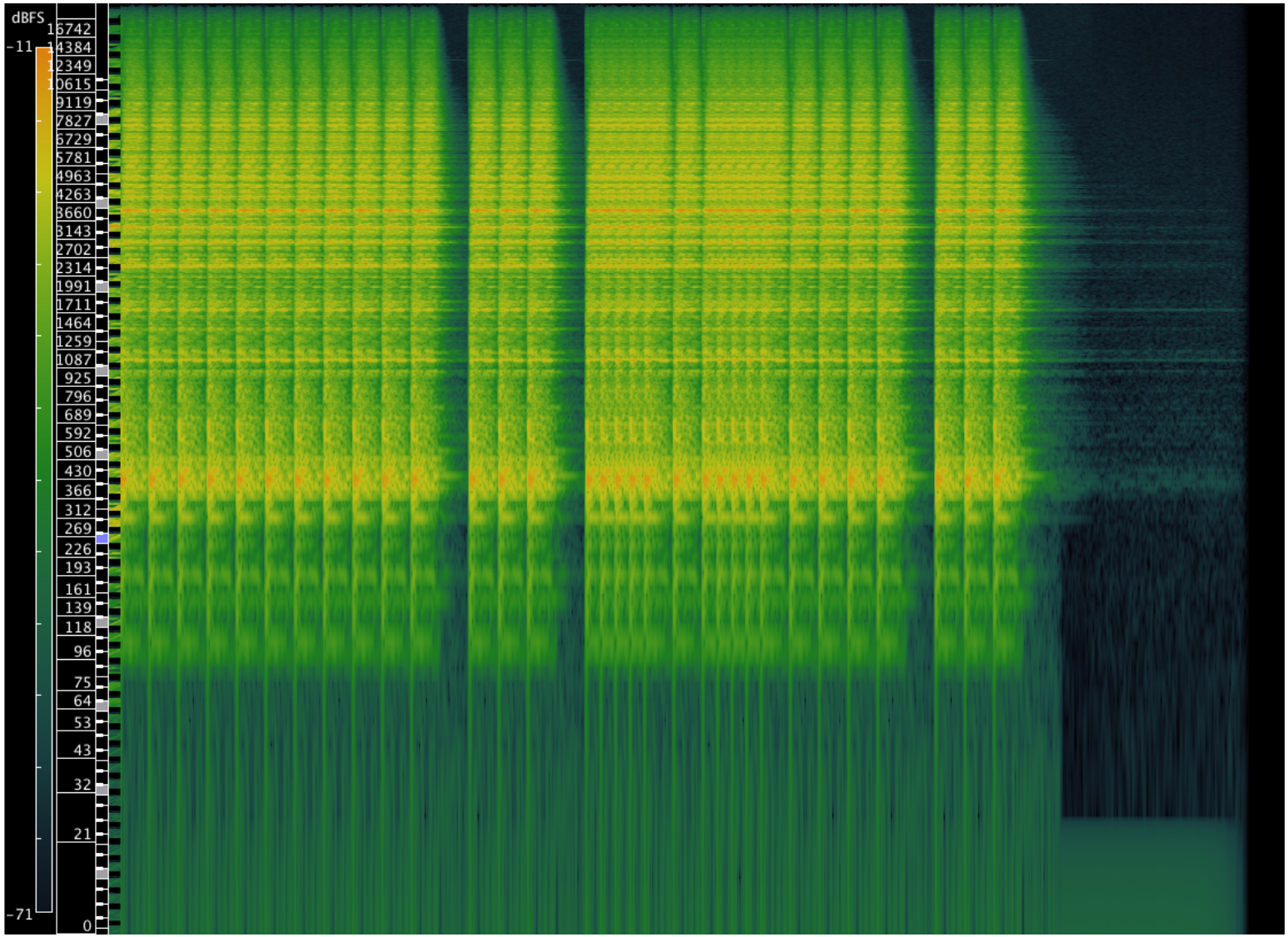
JKU

EX: PITCHED NON-PERCUSSIVE



JKU

EX: NON-PITCHED PERCUSSIVE



JKU

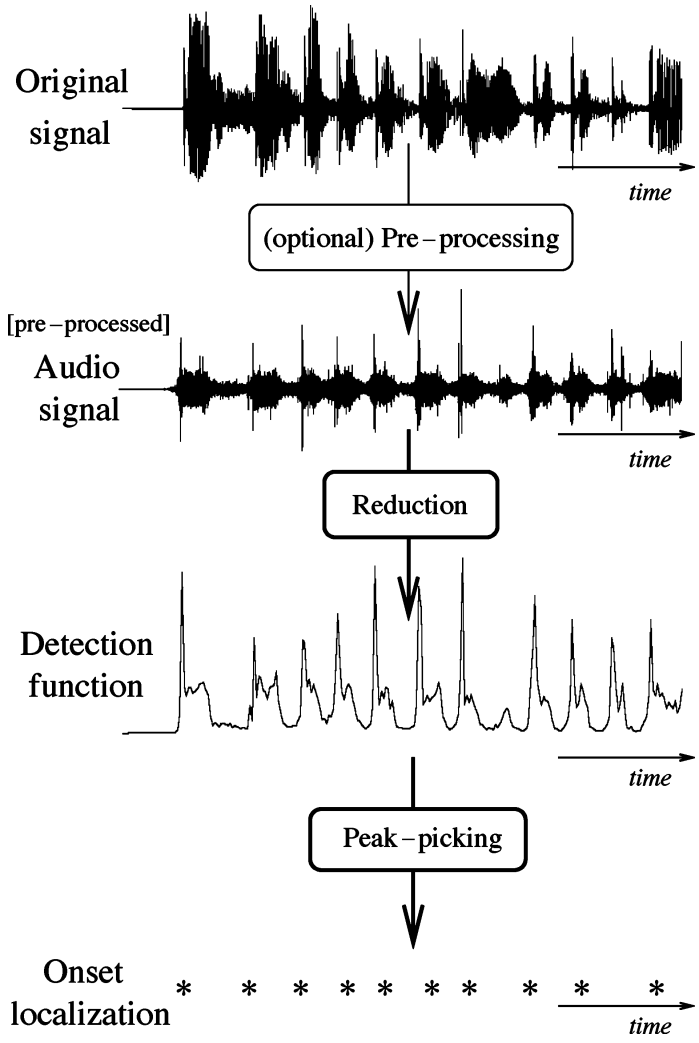
ONSET DETECTION

ONSET DETECTION

- onsets are the **atoms** of **musical timing**
- **a first step** towards high-level feature extraction
- the **basis** for many techniques in music analysis, such as beat tracking, transcription
- **not as trivial** as it may appear at first glance
- real world signals are **noisy**
- most music is **polyphonic**
- different instruments have **different types of onsets**
- onsets may **mask** each other

GENERAL APPROACH

- in general it is not possible to detect onsets directly in noisy polyphonic signals
- approaches generally look as follows:
  - preprocess signal
  - derive a **detection function**
  - postprocess output of detection function
  - pick peaks



PREPROCESSING

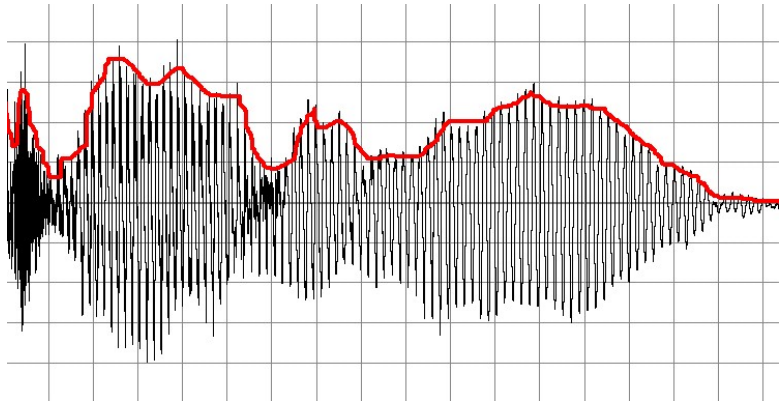
- **accentuate** or **attenuate** various aspects of the signal according to their relevance for onset detection
- compute the STFT or a wavelet transform
- separate the signal into
  - multiple bands
  - transient (percussive) content
  - steady-state (harmonic) content
- we will not look too much into preprocessing in this lecture
- we already know how to compute the STFT, which is the most important preprocessing step

DETECTION FUNCTIONS

- **transform** the audio signal into a (highly sub-sampled) **detection function**
- the detection function should **emphasize transient** regions

SIGNAL ENVELOPE (1)

- very simple approach
- note onsets are usually accompanied by an increase in amplitude
- compute a smoothed version of the **signal envelope** to get the 'outline' of a signal
- most obvious technique
- move a window of size M over the signal x, compute the average of absolute (squared) values, with an optional weighting w



$$E_0[t] = \frac{1}{M} \sum_{m=-\frac{M}{2}}^{\frac{M}{2}-1} |x[t+m]| \; w[m]$$

(or as a variant)

$$E[t] = \frac{1}{M} \sum_{m=-\frac{M}{2}}^{\frac{M}{2}-1} (x[t+m])^2 \; w[m]$$

JKU

SIGNAL ENVELOPE (2)

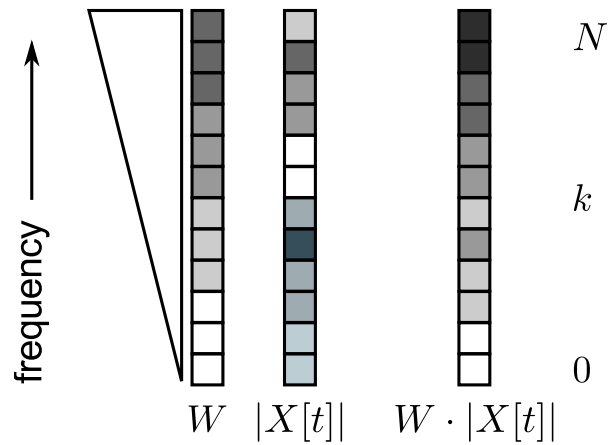
- in general this simple detection function is not very reliable
- possible refinements include:
  - use finite differencing to compute an approximation to the first derivative of the signal (can be forward, backward or central differences)
    - if the approximate derivative is large, this results in large peaks in the detection function
  - use psychoacoustic insights about human perception of loudness
    - perceptually informed equal-loudness curves such as “A-weighting”
    - approximate perceived loudness envelope instead of amplitude envelope

HIGH FREQUENCY CONTENT

- the energy of the signal is usually concentrated in the lower frequencies
- transients appear as **broadband events** in the frequency domain - they stretch across the whole frequency spectrum
- this makes transients **more noticable** in the **higher frequencies**

$$d_{HFC}[t] = \frac{1}{N} \sum_{k=0}^{N} W_k \cdot |X_k[t]|^2$$

where $|X_k[t]|$ is the energy in bin k, at time t
W_k is the weight for bin k



JKU

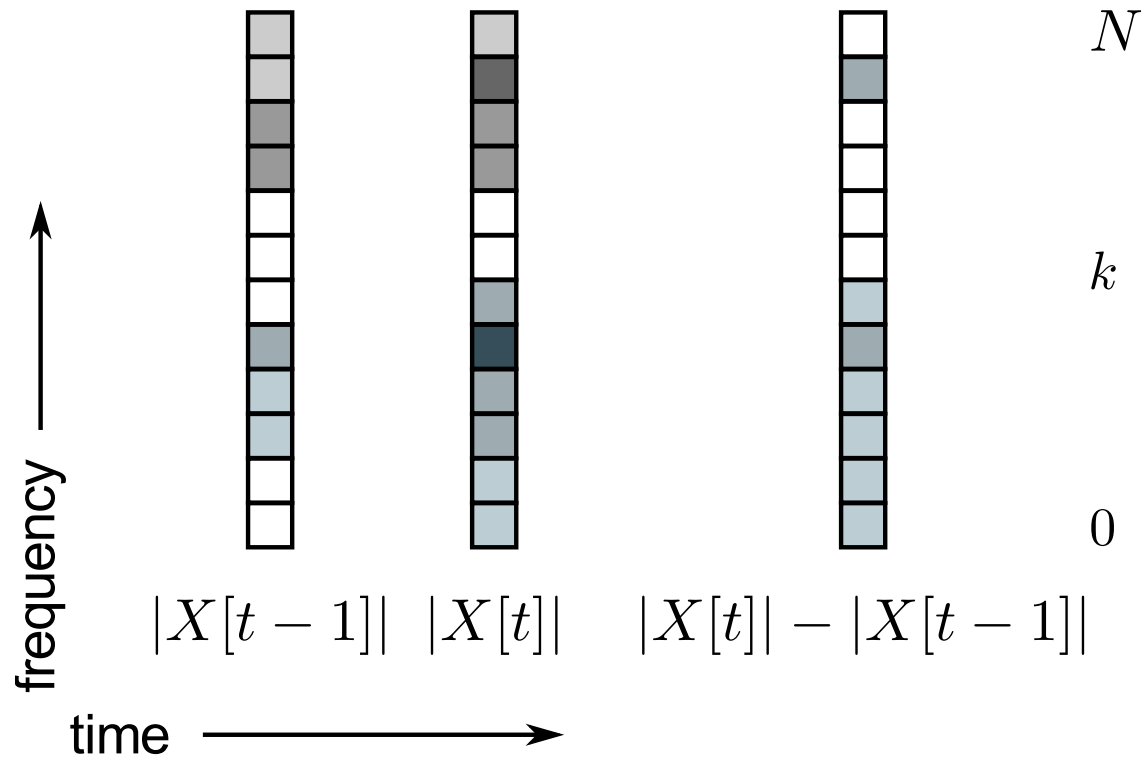
SPECTRAL DIFFERENCE (1)

- take the temporal evolution of the spectrum into account
- treat two successive frames in the STFT as points in an N-dimensional space
- compute the L^p norm of their difference
- $L^p(\mathbf{x}) = \|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$
- common choices for norms: L^1, L^2 ($\|\cdot\|_1, \|\cdot\|_2$)
- optional: only look at positive components of the difference ('half-wave rectification' function H)

$$SD[t] = \sum_{k=0}^{N} [H(|X_k[t]| - |X_k[t-1]|)]^2$$

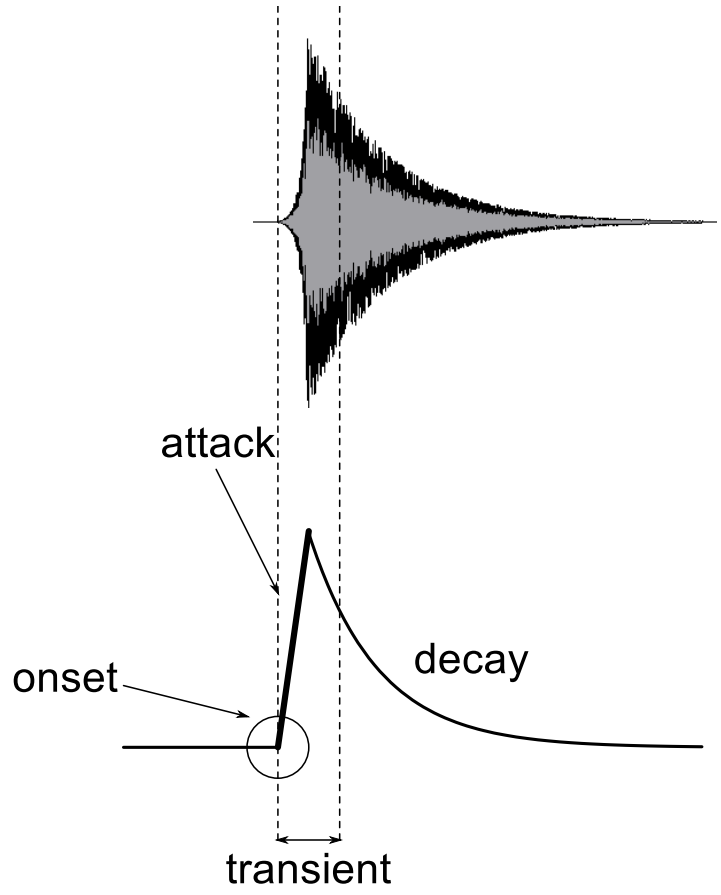
$$H(x) = \max(0, x) = \frac{x + |x|}{2} = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x <= 0 \end{cases}$$

SPECTRAL DIFFERENCE (2)



USING PHASES

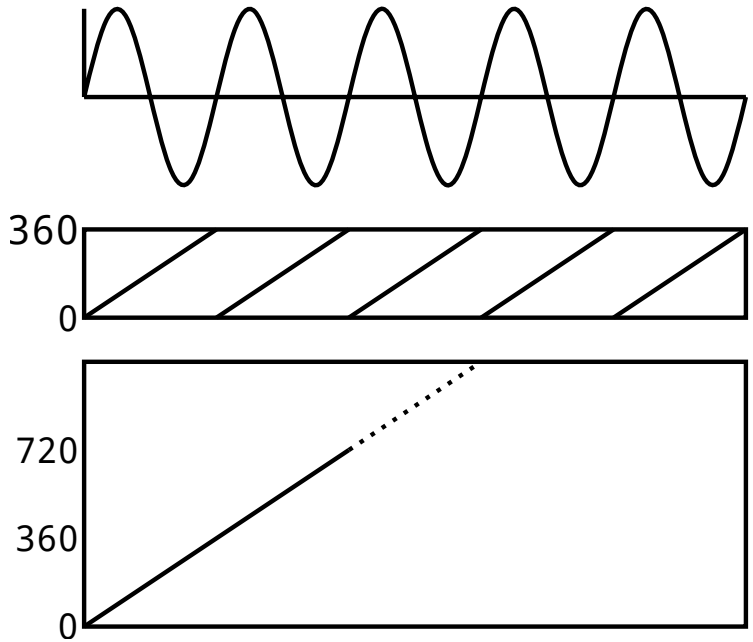
- while a sound sustains / decays, phases “roll” in predictable manner
- a new sound will disturb this, even if it is harmonic
- detecting phase shifts will help detect onsets



JKU

PHASE UNWRAPPING

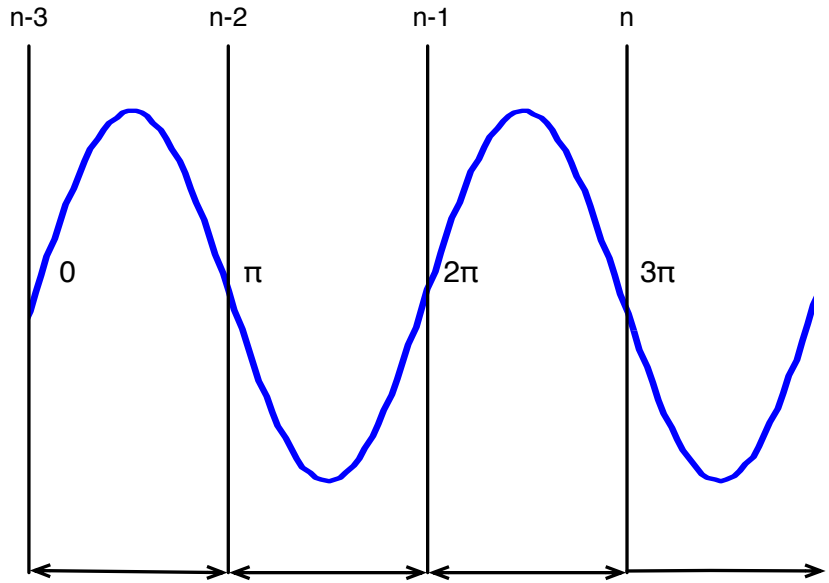
- the phase $\varphi[t]$ is restricted to the interval $[0°, 360°) = [0, 2\pi)$
- we can look at $\varphi[t]$ as a continuous function of t
- for a regular sinusoid, it still jumps from 360° to 0° , which is a jump we want to ignore
- we “unwrap” the phase to do so, and will call it $\tilde{\varphi}[t]$
- the unwrapped phase then **increases linearly** over time, instead of jumping back



JKU

PHASE DEVIATION (1)

- use the phase information for transient detection
- for a **locally stationary** sinusoid, the unwrapped phase increment should be (approximately) **constant** over adjacent windows
- **phase deviation** is given by the second difference of the phase



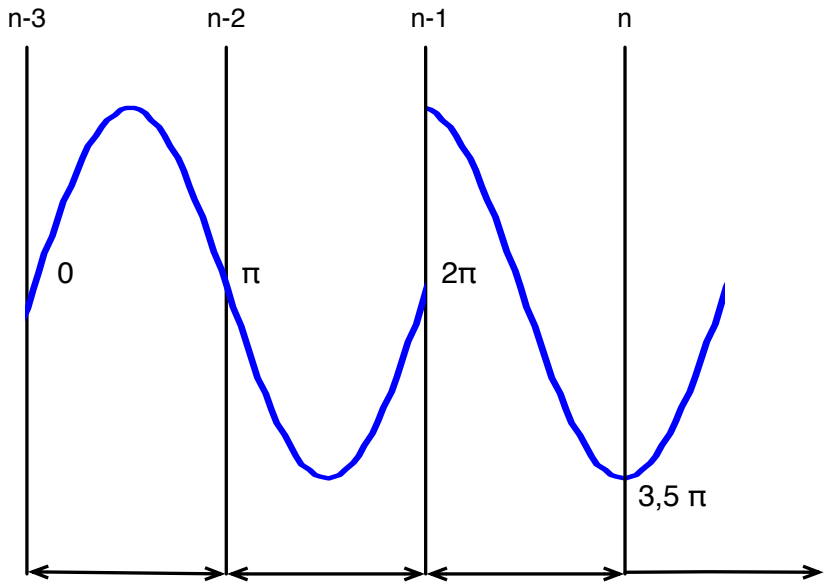
$$\tilde{\varphi}_k[t] - \tilde{\varphi}_k[t-1] \simeq \tilde{\varphi}_k[t-1] - \tilde{\varphi}_k[t-2]$$

$$\Delta\tilde{\varphi}_k[t] = \tilde{\varphi}_k[t] - 2\tilde{\varphi}_k[t-1] + \tilde{\varphi}_k[t-2] \simeq 0$$

PHASE DEVIATION (2)

- during a transient, the quantity $\Delta\tilde{\varphi}_k[t]$ tends to be large
- based on this observation, multiple formulations of detection functions are possible
- an example would be the "mean absolute phase deviation" $\zeta[t]$

$$\zeta[t] = \frac{1}{N}\sum_{k=1}^{N} |\Delta\tilde{\varphi}_k[t]|$$

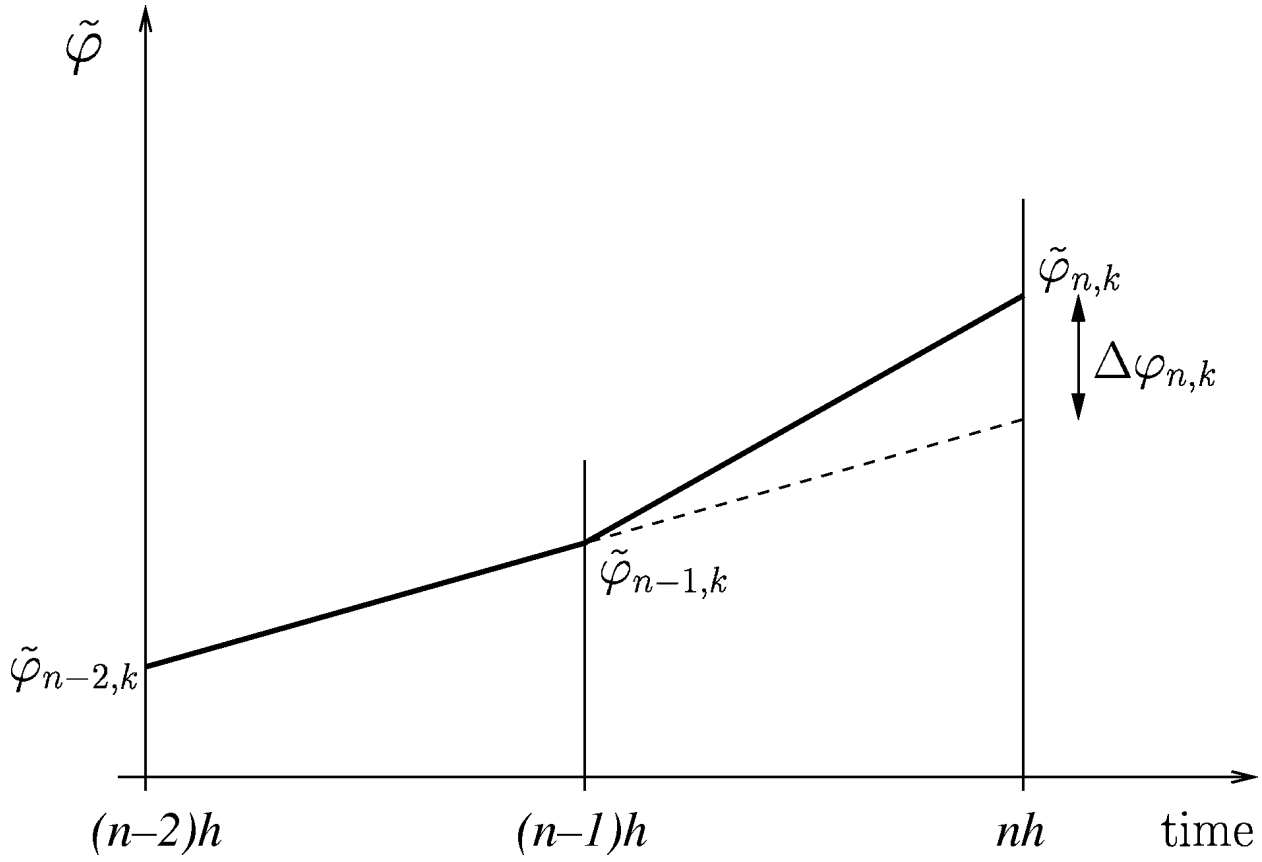


in this graph:

$$\Delta\tilde{\varphi}_k[n] \approx 0.5\pi$$

$$\Delta\tilde{\varphi}_k[n-1] \approx 0$$

PHASE DEVIATION (3)



this phase diagram shows the unwrapped phase over adjacent frames - for a stationary sinusoid the phase deviation should more or less stay constant, as hinted at with the dotted line

JKU

COMPLEX DOMAIN

- amplitude and phase can be considered jointly to detect departures from steady-state behaviour
- the stationarity of the k th spectral bin is quantified by the L_1 **distance** between the **observed** and the **predicted** value, where $\hat{X}_k[t]$ is the prediction:

$$\hat{X}_k[t] = |X_k[t-1]|e^{j(\varphi_k[t-1]-\Delta\varphi_k[t-1])}$$

- these distances are then summed to obtain an onset detection function, which is the sum of absolute deviations from target values for each bin:

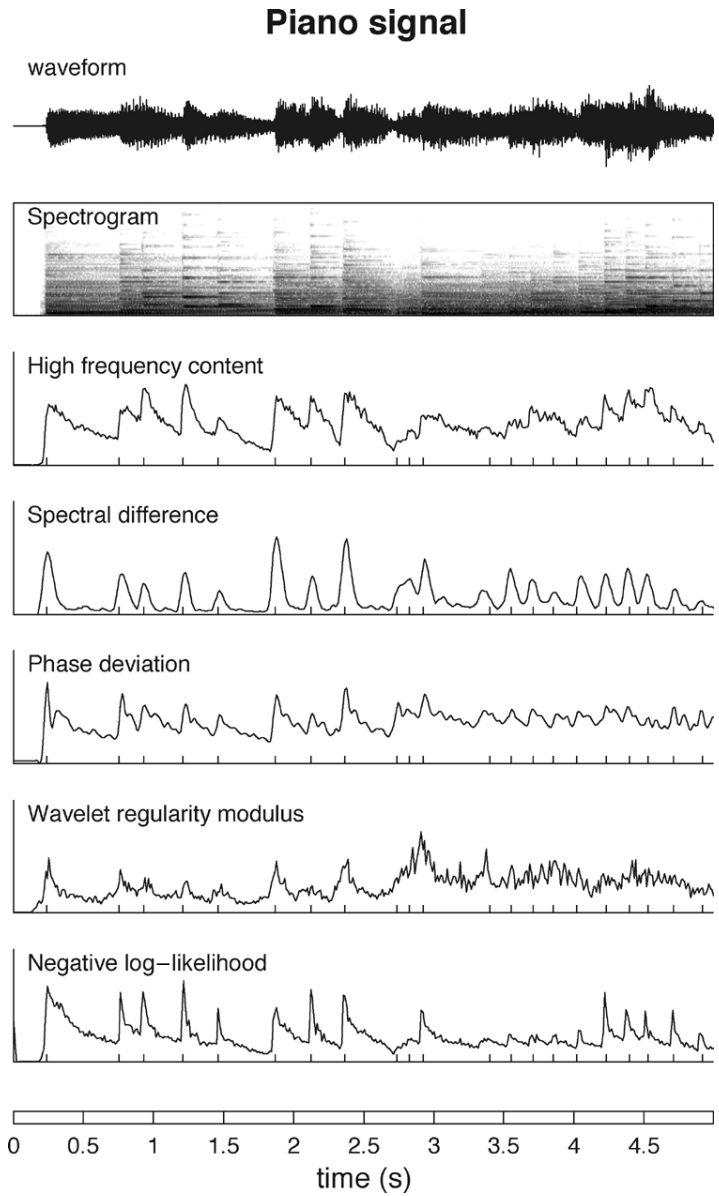
$$d_{CD}[t] = \sum_{k=1}^{N} |X_k[t] - \hat{X}_k[t]|$$

OTHER METHODS

- there is a variety of other approaches for detection functions
- mostly different versions of the approaches presented so far
- time-frequency and time-scale analysis via wavelet regularity modulus
  - large wavelet coefficients are related to transients
  - they are not evenly spread, but form 'structures'
- statistical methods
  - sequential probability ratio test
  - negative log-probability ("surprising moments" in the signal)
- machine learning
  - learn a detection function from data

COMPARISON - PIANO

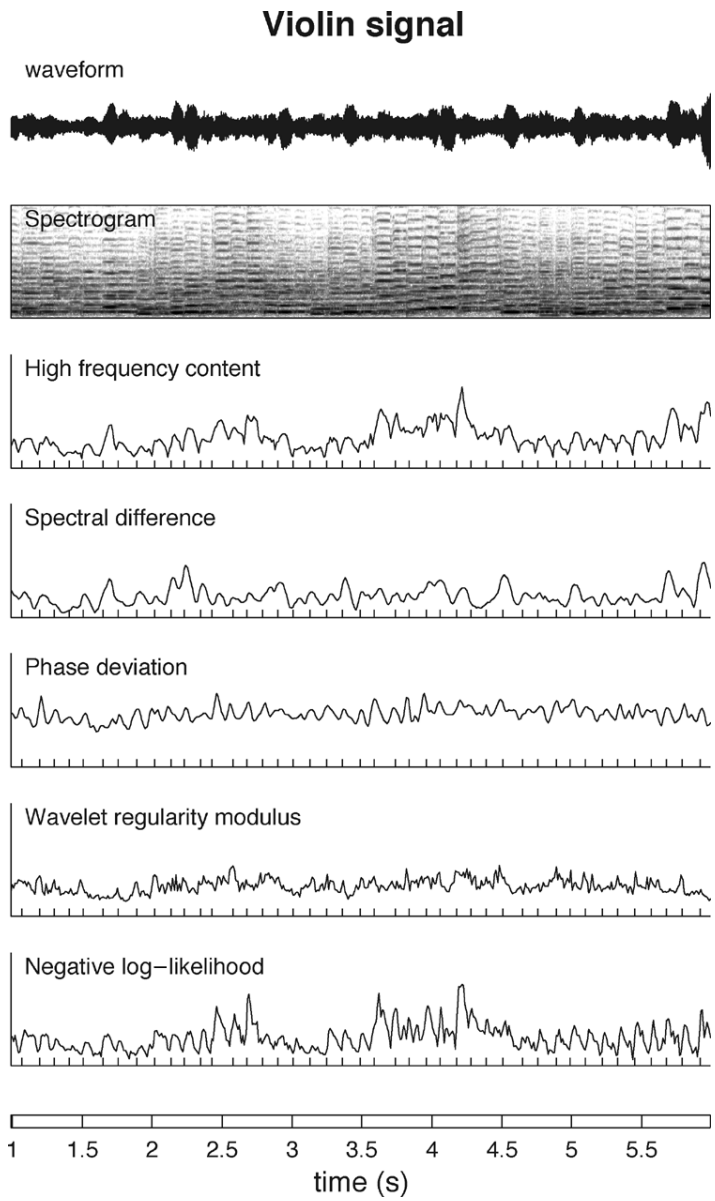
- pitched, percussive onsets
- spectral difference methods are well suited for this kind of signal
- the detection functions we know do well on this kind of signal



JKU

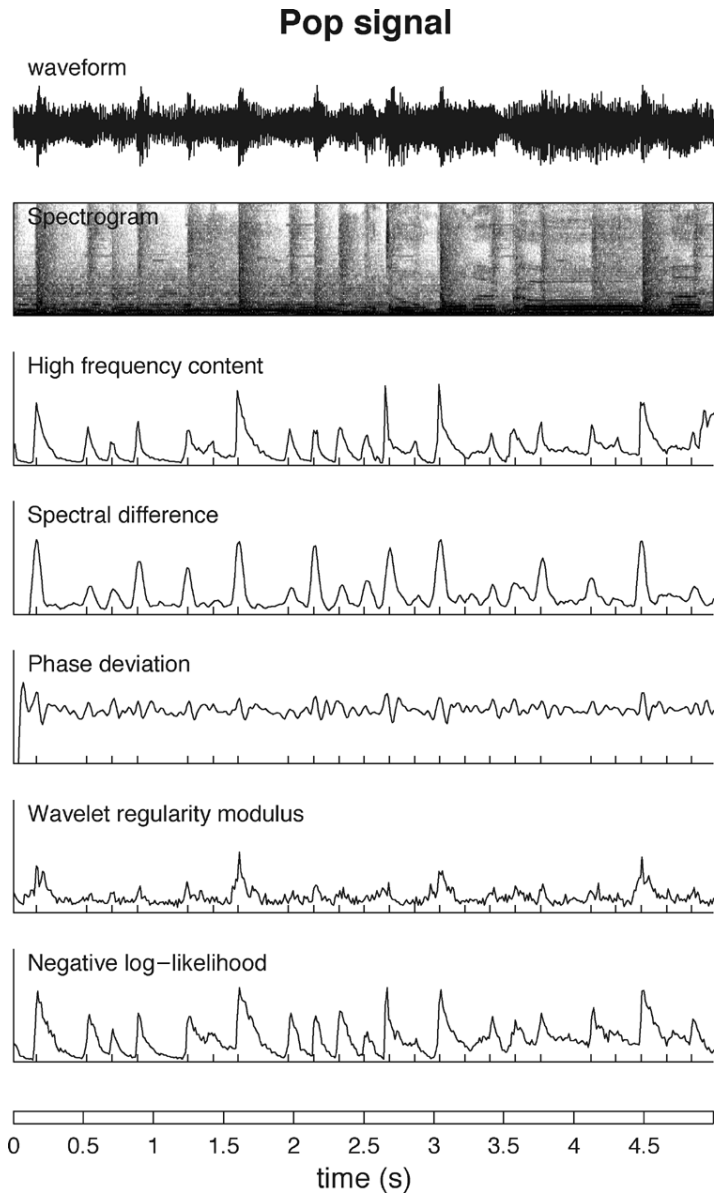
COMPARISON - VIOLIN

- pitched, non-percussive onsets
- phase deviation methods work reasonably well for this kind of signal
- onset detection in these kinds of signals appears to be much harder then in the piano signal we looked at before



COMPARISON - POP MUSIC

- a complex mixture of onsets of different variety
- high-frequency content analysis works well for these kinds of onsets (with typical “pop-percussions”)
- phase deviation suffers from the noise introduced by the percussive instruments



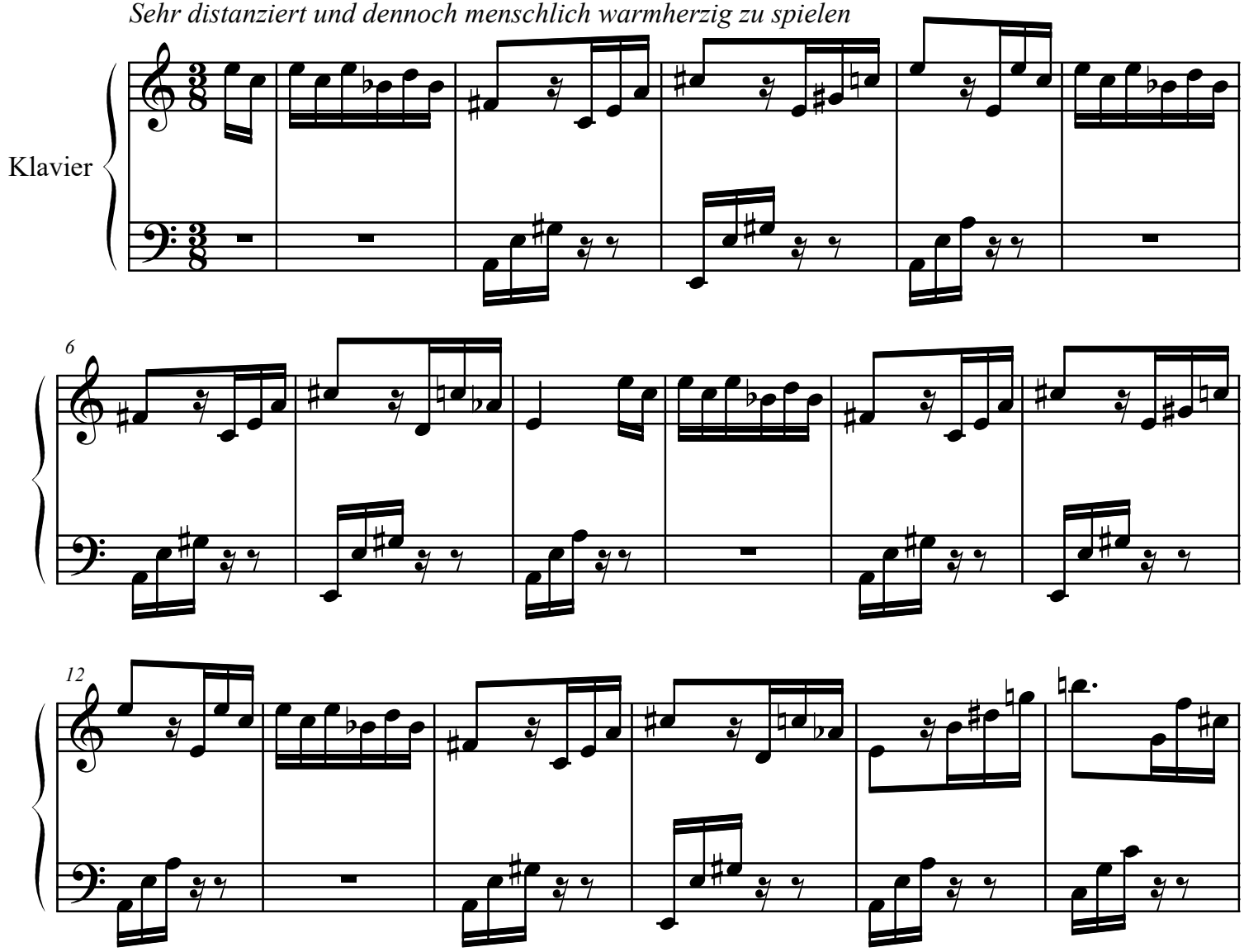
RELAXATION

"Für Elise" (Corona-Version)

Ludwig van Beethoven / Arno Lücker (März 2020)

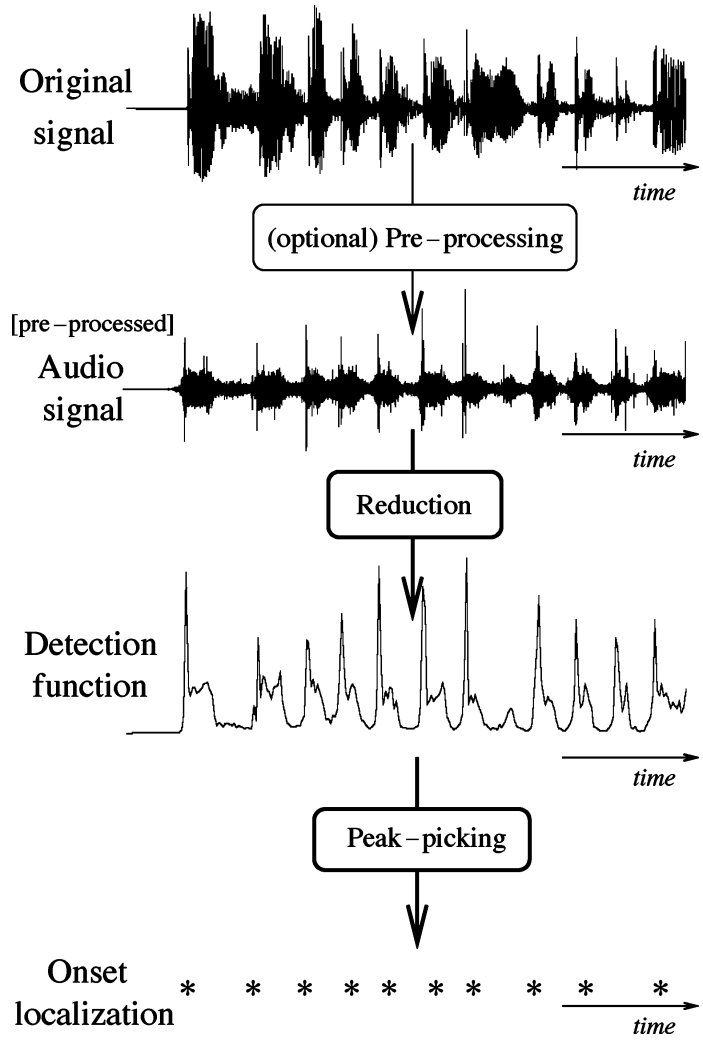
In Zeiten der Krise: Keine Chromatik mehr! Abstand halten auch zwischen Tönen! Es gilt mindestens eine große Terz Abstand!

Sehr distanziert und dennoch menschlich warmherzig zu spielen



PEAK PICKING

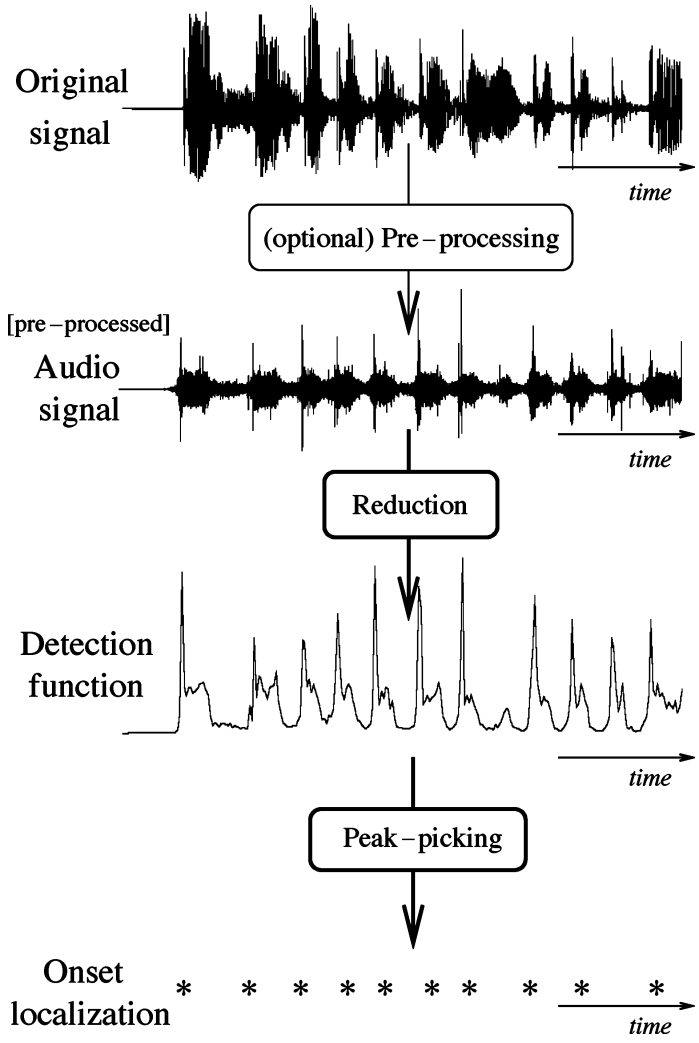
- extract the onset timing from the detection function
- generally detection functions are designed such that onsets in the signal result in peaks (local maxima) in the detection function



JKU

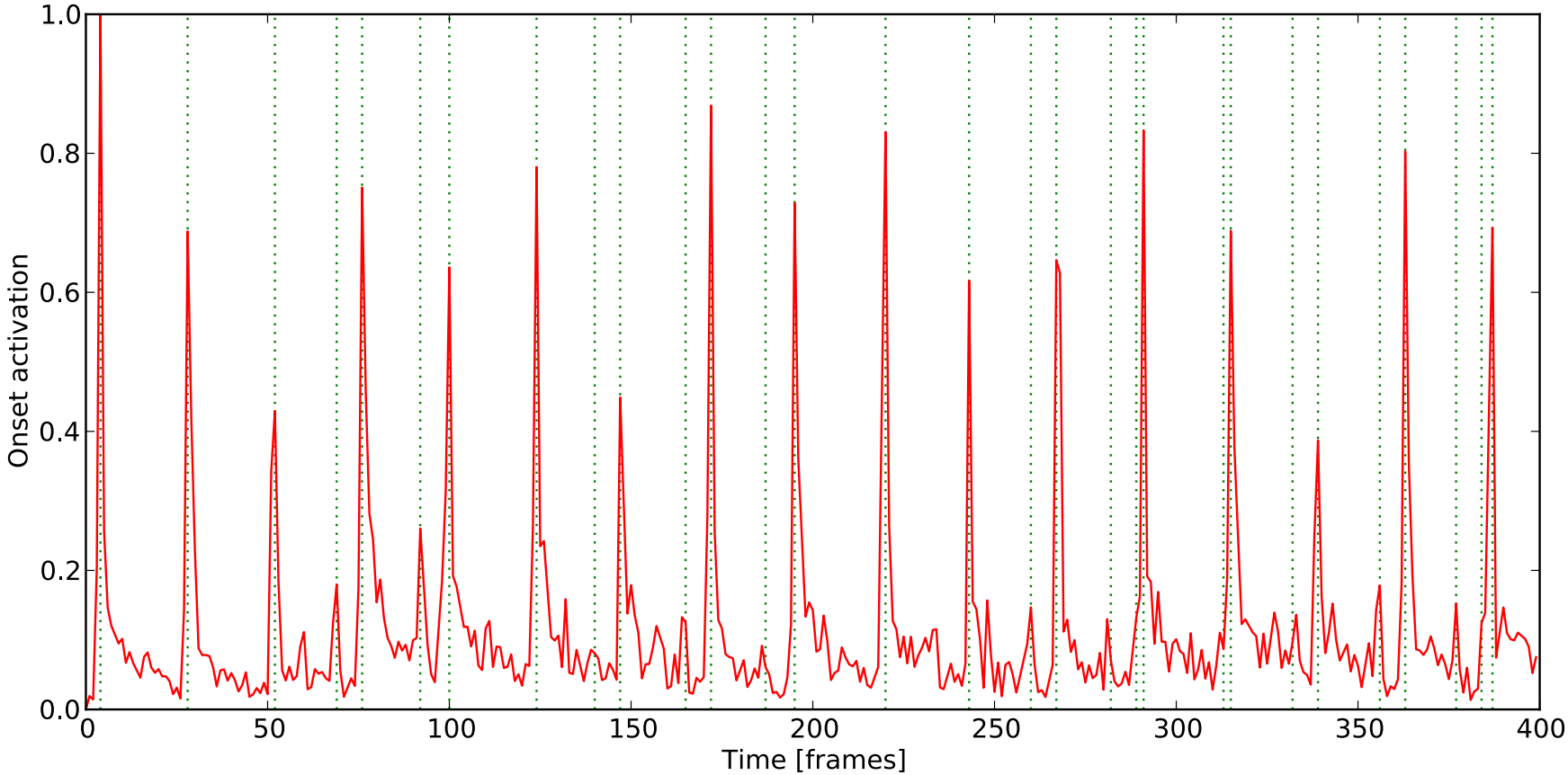
PEAK PICKING

- these peaks come in a variety of shapes and sizes
- peaks may be masked by noise in the signal or musical aspects, like vibrato
  - post-processing of the detection function
  - thresholding
  - more complex decision process



JKU

RAW DETECTION FUNCTION

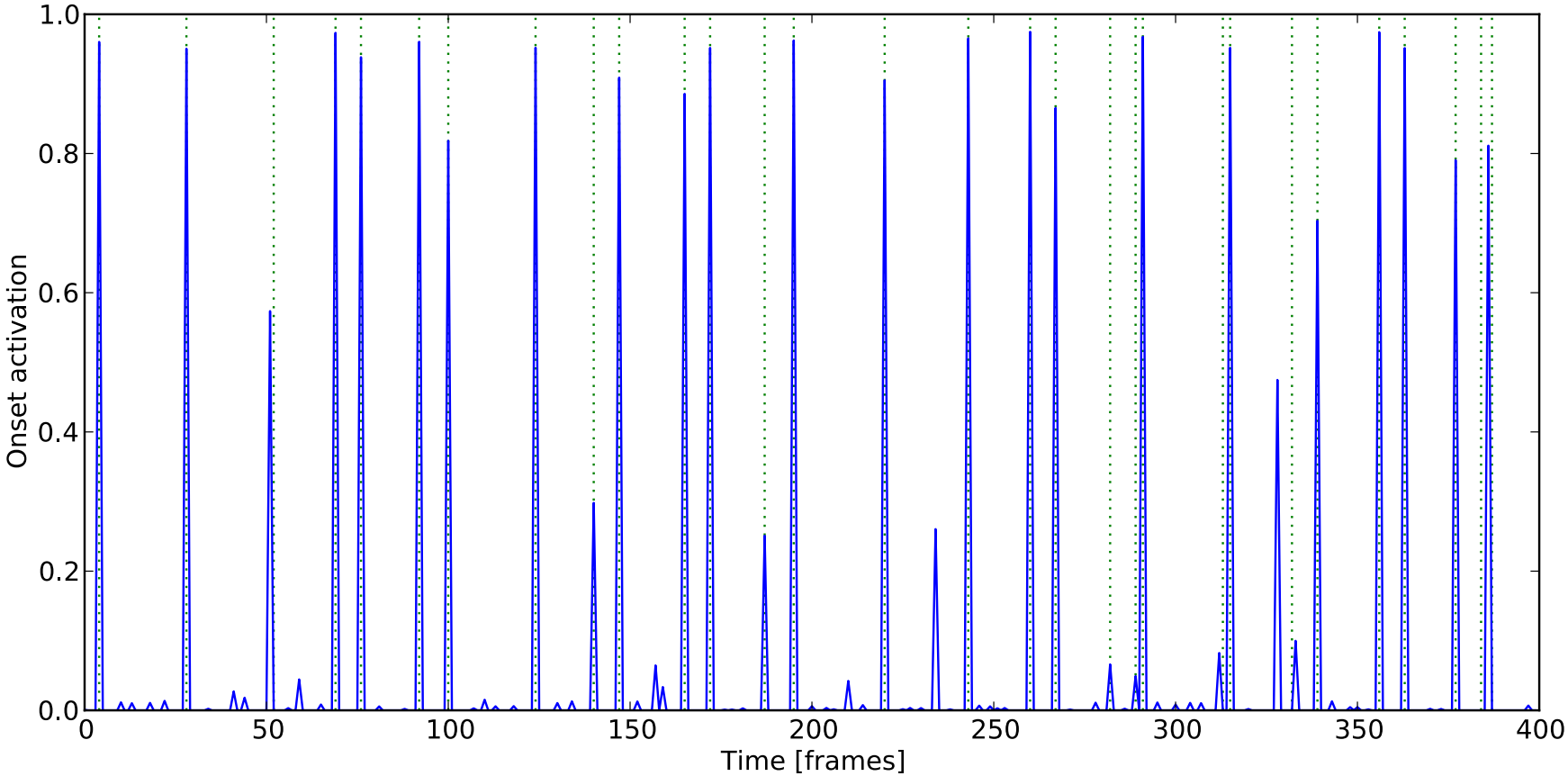


JKU

POST PROCESSING

- increase the uniformity and consistency of onset-related peaks in the detection function
- in the ideal case, this would transform the peaks into isolated and easily detectable local maxima
- possible steps include
  - smoothing (reduce effects of noise)
  - normalization
  - DC removal

AFTER POST PROCESSING

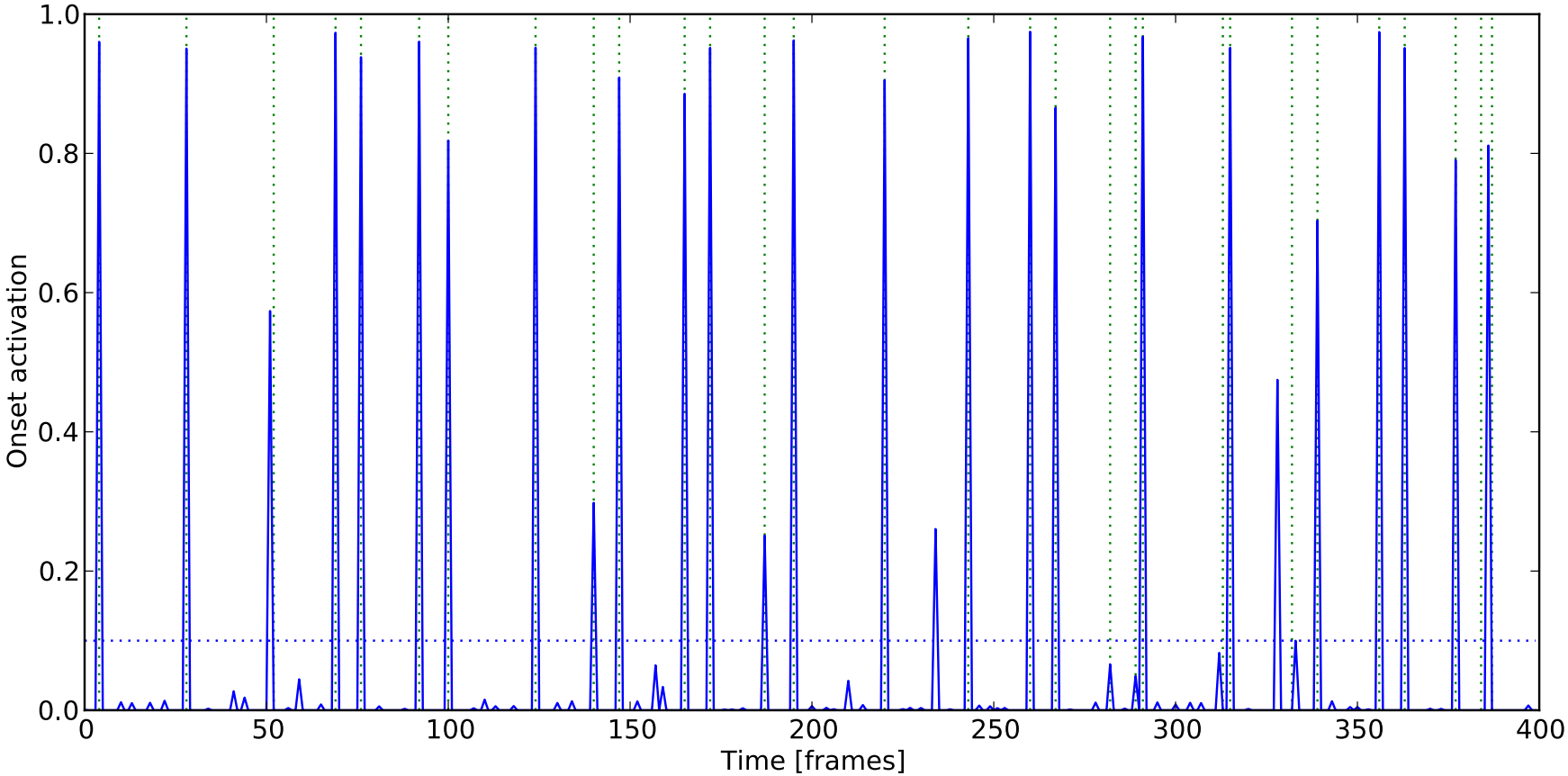


JKU

FIXED THRESHOLD

- even after postprocessing there will be a number of peaks which are not related to onsets
- it is necessary to define a threshold which separates onset-related and non-related peaks
- simplest idea is to use a **fixed threshold** δ
- but: many detection functions scale with sound intensity
- for signals with little dynamic variation, this is unproblematic
- in general, music tends to change in intensity/loudness
- fixed thresholds will **miss** onsets in quiet passages
- fixed thresholds will **over-detect** onsets during loud passages

SIMPLE THRESHOLDING



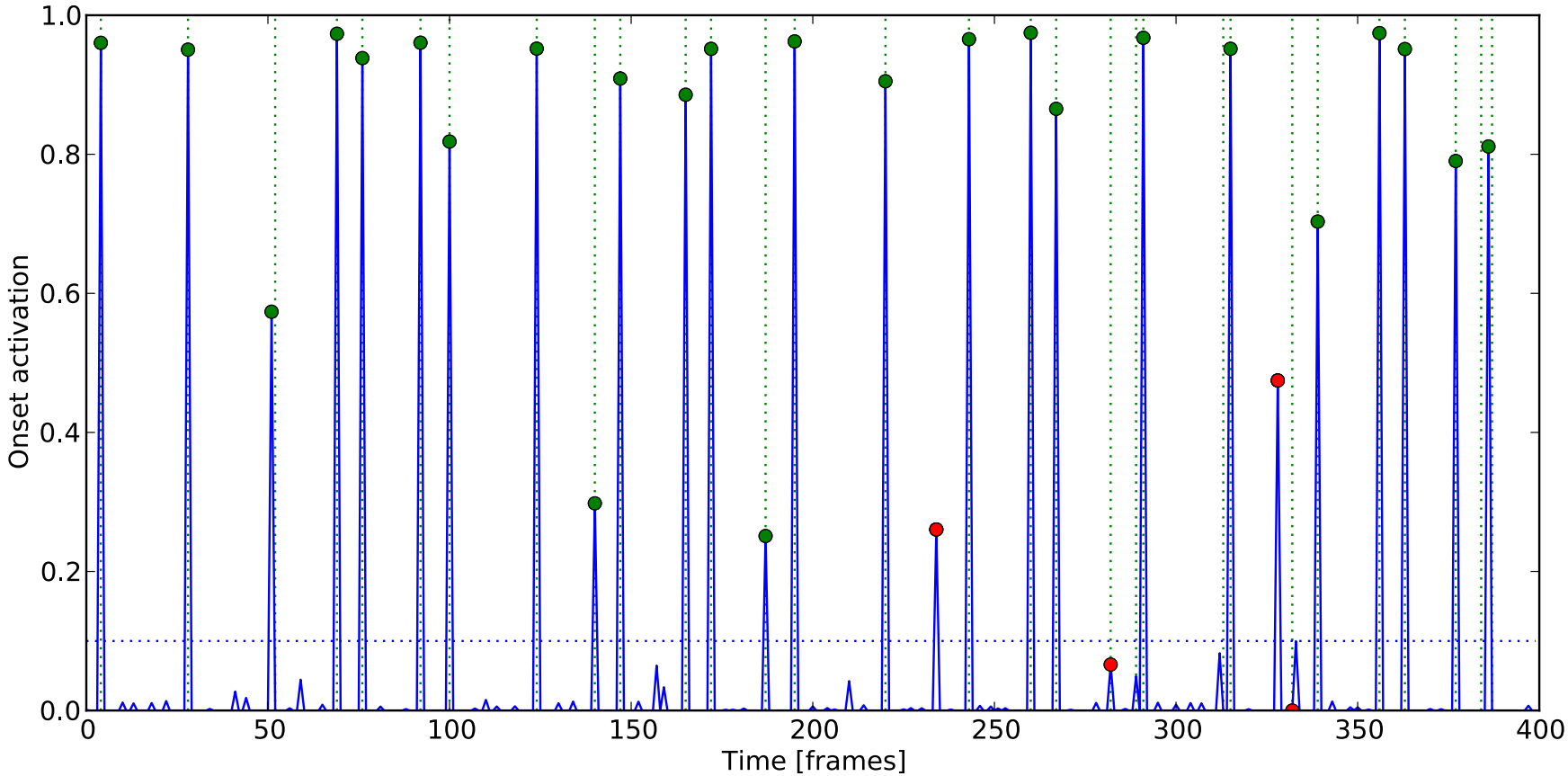
JKU

ADAPTIVE THRESHOLDING

- an adaptive threshold is computed as a smoothed version of the detection function
- smoothing is done via a low-pass filter
- large peaks in the detection function tend to **mask** smaller adjacent peaks
- smoothing may be based on local statistics, such as the median of detection function values in a window centered around t

$$\delta_t[t] = \delta + \lambda \ \mathrm{median}\{d[t-k]\}_{k=-M/2}^{M/2}$$

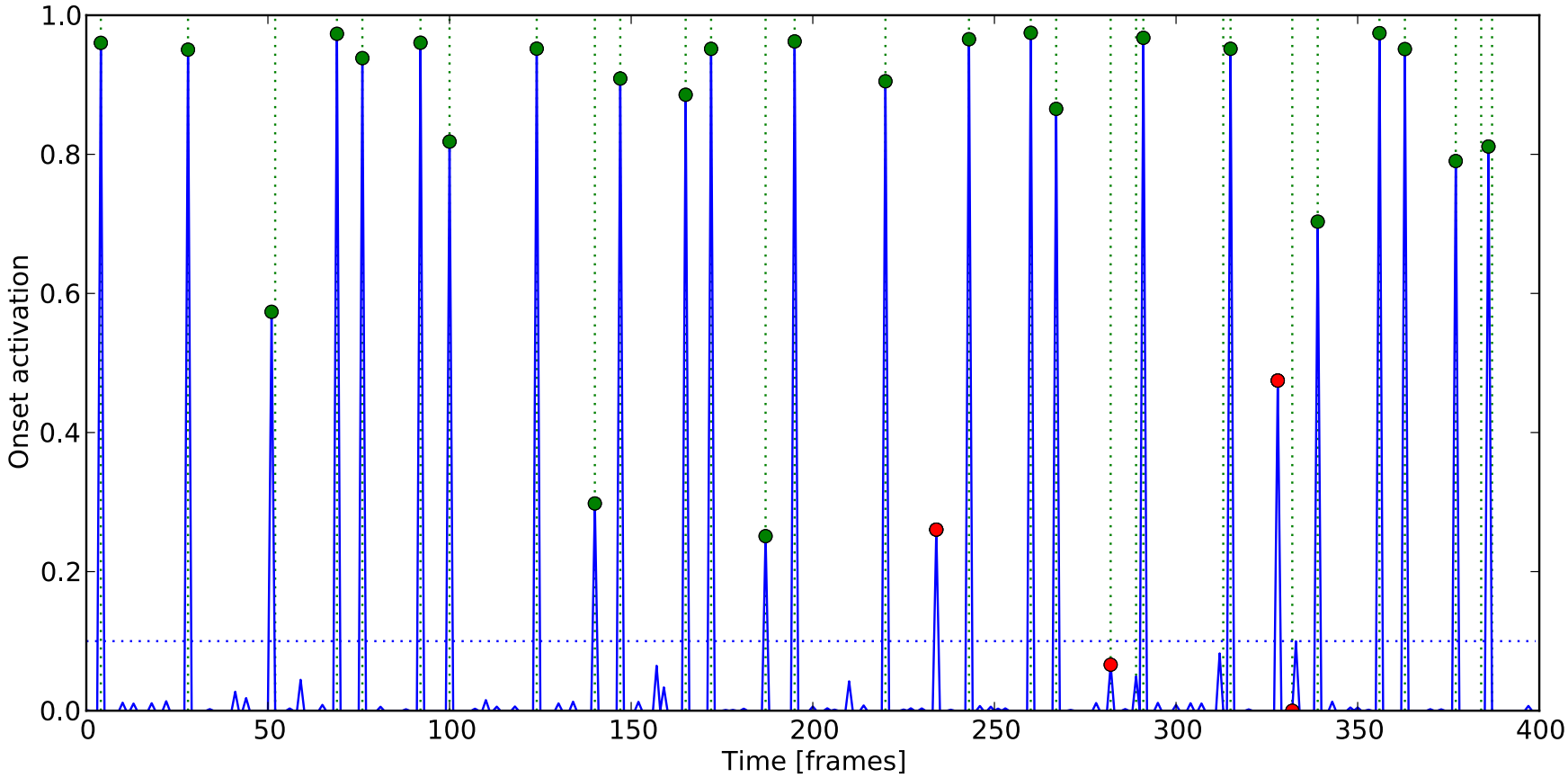
PEAK PICKING



after post processing and thresholding, “peak picking” means identifying local maxima above the defined threshold (in this example, two peaks are wrongly detected, one peak is missed)

JKU

PEAK PICKING



possible implementation: apply sliding maximum, compare with original $\rightarrow$ where both are equal, it's a local maximum

JKU

STATE OF THE ART

you do not have to understand every detail in this chapter,
try to remember the basic ideas

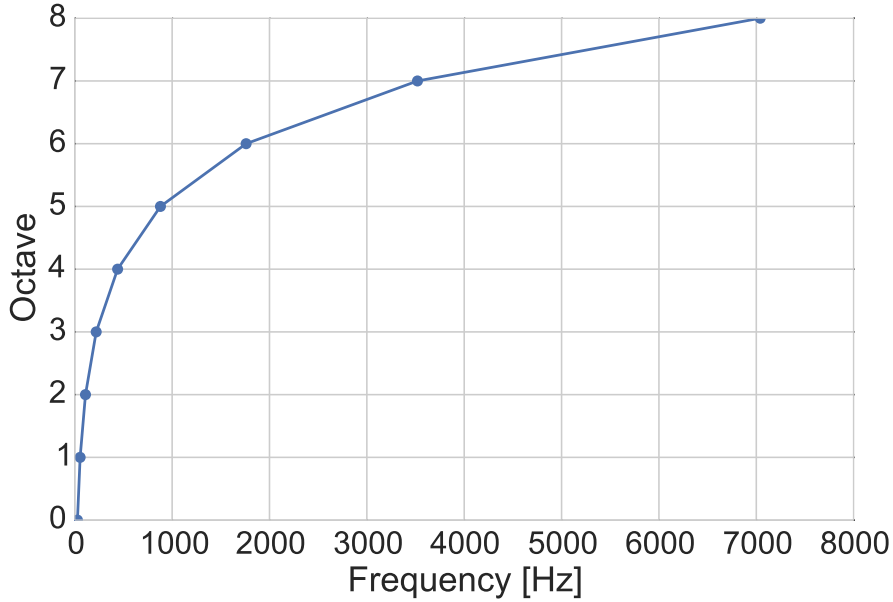
PSYCHOACOUSTICS

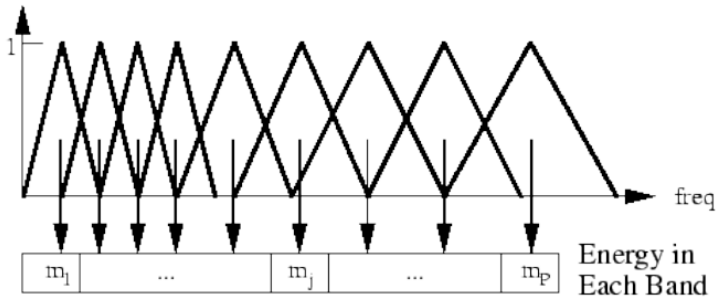
- map the results of the spectral analysis stage (the STFT) to a psychoacoustically meaningful scale
- apply an onset detection function such as spectral flux
- different scales are possible
  - bark scale
  - mel scale
  - semitone scale
- aforementioned scales are widely used in music processing

SEMITONE SCALE

- map the linearly spaced frequency bins from the STFT to a semitone scale
- this map takes the form of a filterbank with triangular filters
- filters are centered on pitches
- filterwidth is given by neighboring pitches
- normalized by area under filter

frequencies of note A:





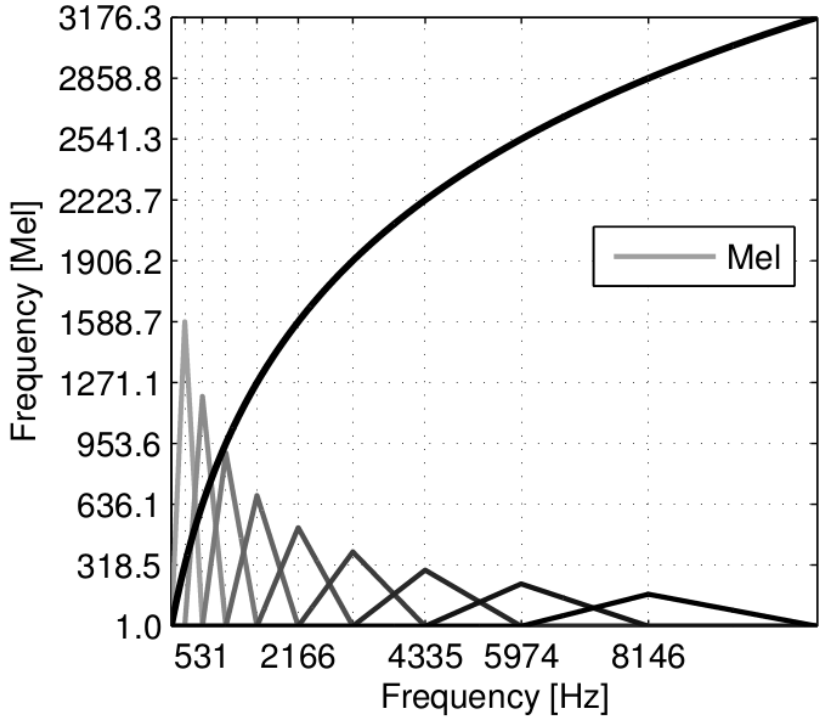
JKU

THE MEL SCALE

- perceptual scale of pitches judged by listeners to be equal in distance from one another
- given frequency f in Hertz, the corresponding pitch in mel can be computed by

$$m = 2959 \log_{10} \left(1 + \frac{f}{700} \right)$$

- normally around 40 bins equally spaced on the mel scale are used



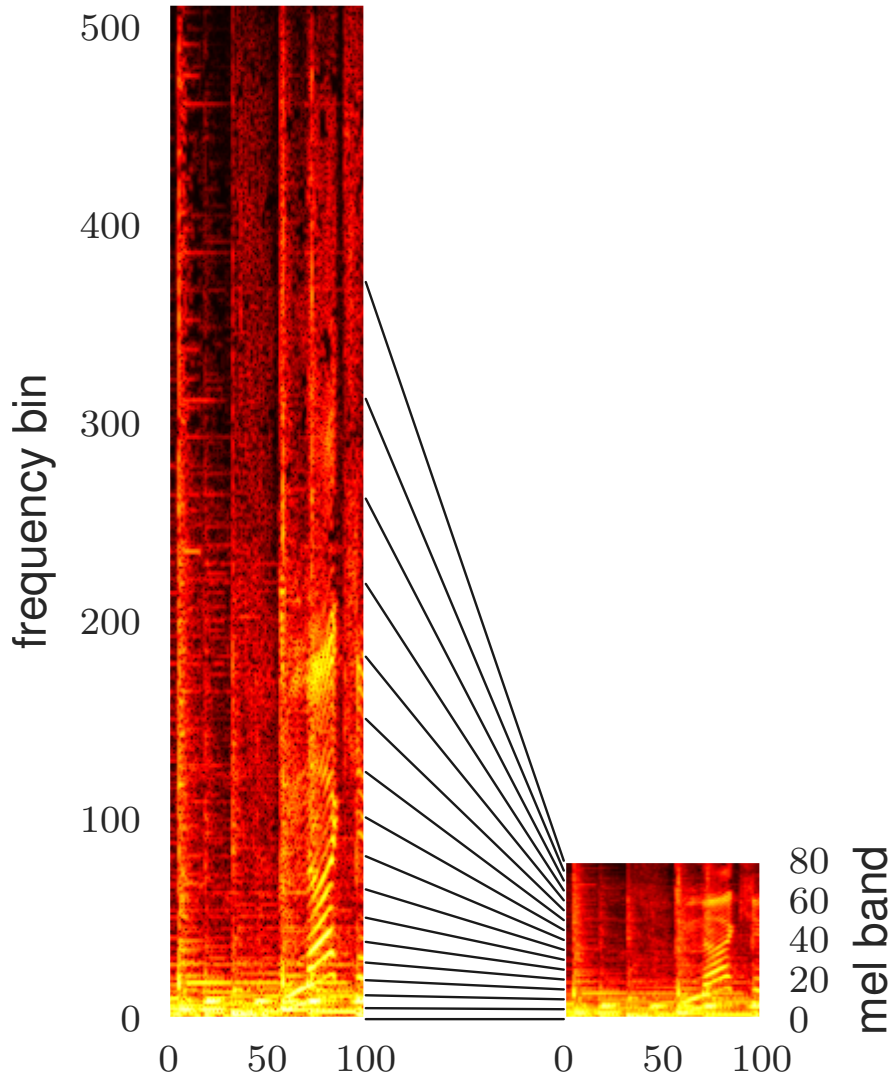
JKU

THE MEL SCALE

- perceptual scale of pitches judged by listeners to be equal in distance from one another
- given frequency f in Hertz, the corresponding pitch in mel can be computed by

$$m = 2959 \log_{10} \left(1 + \frac{f}{700} \right)$$

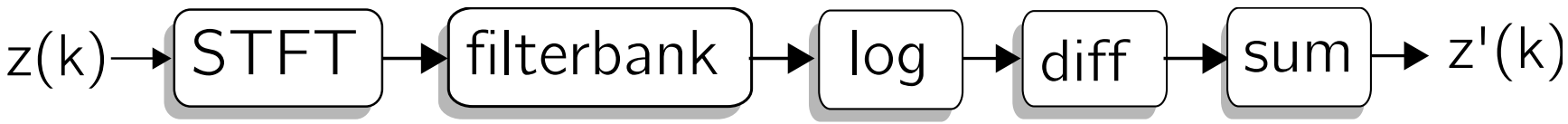
- normally around 40 bins equally spaced on the mel scale are used



JKU

“LogFiltSpecFlux” [3]

- relatively simple algorithm, capable of producing state-of-the-art results
- based on spectral differences
- psychoacoustically informed
- optimized peak picking



LFSF - PREPROCESSING

- compute the STFT (window size 23 [ms], hop size 10 [ms])

$$X_k[t]$$

- apply filterbank - map the STFT to the semitone scale

$$|X_f[t]| = \sum_{k=1}^{N} |X_k[t]| \cdot F_k[f]$$

- convert filtered spectrogram to logarithmic magnitude

$$\log(1 + \lambda \cdot |X_f[t]|)$$

- shifting by 1 and multiplying by a scalar λ compresses the dynamic range of the signal

LFSF - DETECTION FUNCTION

- the detection function is defined as
  (N_f is the number of triangular filters)

$$d_{LFSF}[t] = \sum_{f=0}^{N_f-1} H\big(\log\left(1+\lambda|X_f[t]|\right) - \log\left(1+\lambda|X_f[t-1]|\right)\big)$$

- $H(\cdot)$ above is again the half-wave rectifier

$$H(x) = \max(x,0) = \frac{x+|x|}{2} = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x <= 0 \end{cases}$$

LFSF - PEAK PICKING

- a point in the reduction function has to fulfill three properties to be considered an onset
  - it has to be a local maximum in a window (w_1, w_2)
  - it has to be greater than the mean in a window (w_3, w_4)
  - it has to have a minimum distance in time from other detected onsets (w_5)
- these parameters will be set to maximize the performance on a so called “validation dataset” (more on that later)

$$d(t) = \max(d(t - w_1 : t + w_2))$$
$$d(t) \geq \text{mean}(d(t - w_3 : t + w_4)) + \delta$$
$$t - t_{last\ onset} > w_5$$

“SuperFlux” [4]

- an extension to “LogFiltSpecFlux”
- applies a small maximum-filter on the STFT **before** computing the spectral difference
- the maximum-filter acts on a **region** in time **and** frequency
- this helps to **reduce false detections** caused by notes with **vibrato**

ENERGY BASED METHODS

- “An Energy- and Pitch-based Approach to Audio Onset Detection” [10]
- combine pitch and energy information
- **classify** the audio recording according to **which types of onsets** it contains (pitched percussive, pitched non-percussive, non-pitched percussive)
- use **specialized** onset detection function for **each category**

EBM - EXCERPT CLASSES

- estimate **percussiveness** - abrupt energy changes over the whole spectrum as well as a limited frequency band (1 kHz - 5 kHz)
- estimate **pitch presence and tuning** - determine whether pitches are present, and how far they deviate from a reference tuning
- these **features** allow the **classification** of musical excerpts according to their prevalent type of onset, where the class $c \in \{\mathrm{PP}, \mathrm{PNP}, \mathrm{NPP}\}$
- decisions are made using various **manually derived heuristics**

EBM - ONSET DETECTION

- **depending** on the **class** of musical excerpt, **different** onset detection algorithms are used
- percussiveness - compare two successive audio frames in the frequency domain via finite differencing
- pitch changes - computed by comparing changes in pitch pairs (F_0 and first partial)
- for each of these steps, categorization and detection, different parametrizations are used; think “coarse” to “fine”-grained

ML INTERLUDE (1)

- given a set of pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$
- a decision function $\hat{f} : \mathcal{X} \to \mathcal{Y}, \hat{y} = \hat{f}(\mathbf{x})$
- a misclassification cost $l(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{if } \hat{y} \neq y \end{cases}$
- the empirical risk $\mathrm{R}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} l(\hat{f}(\mathbf{x}_i), y_i)$
- we **do not know** the **true function** f that produced the original set of pairs, so we need to **approximate** it
- we **learn** a function f_* that **minimizes empirical risk**

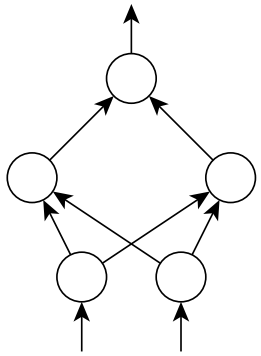
$$f_* = \underset{\hat{f} \in \mathcal{F}}{\arg\min}\ \mathrm{R}(\hat{f})$$

ML INTERLUDE (2)

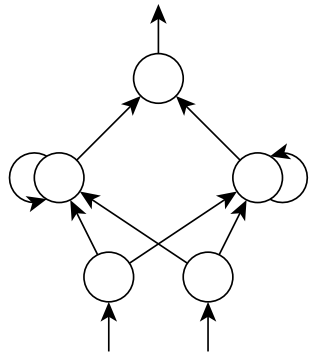
- how does $\mathcal{F}$ look like ?
- it may be a **line** in $\mathbb{R}^2$ or generally a **hyperplane** in $\mathbb{R}^D$
- it may be a **tree**, or a **forest** of **random** trees
- it may be a function $f_L(\mathbf{x}) = \sigma(\mathbf{W}_L f_{L-1}(\mathbf{x}) + \mathbf{b}_L), f_0(\mathbf{x}) = \mathbf{x}$
- we can choose whatever **model class** we think is best suited
- all of these model classes have **parameters**
- **learning** is the process of **searching** for these parameters

RECURRENT NEURAL NETWORKS

- a feed-forward artificial neural network is a **function approximator**, very loosely inspired by biological neural networks
- **recurrent neural nets** [11] are a type of neural network that have feedback connections
- they are used to approximate functions on **sequences**



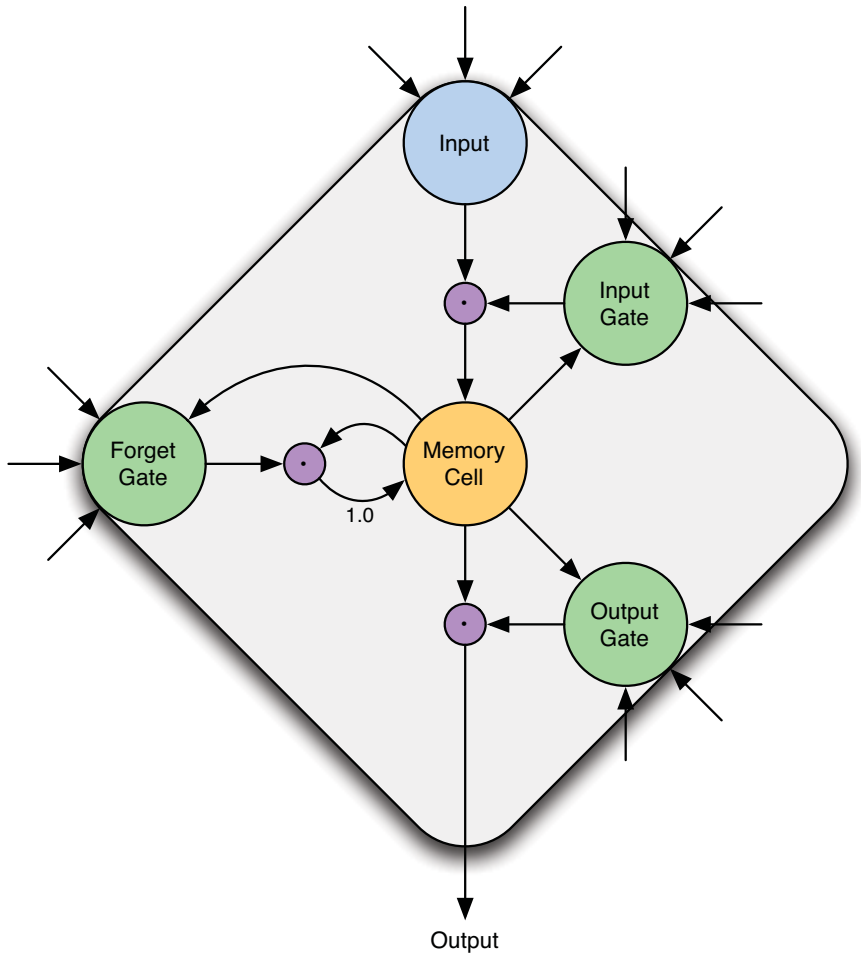
a) simple feedforward net



b) simple recurrent net

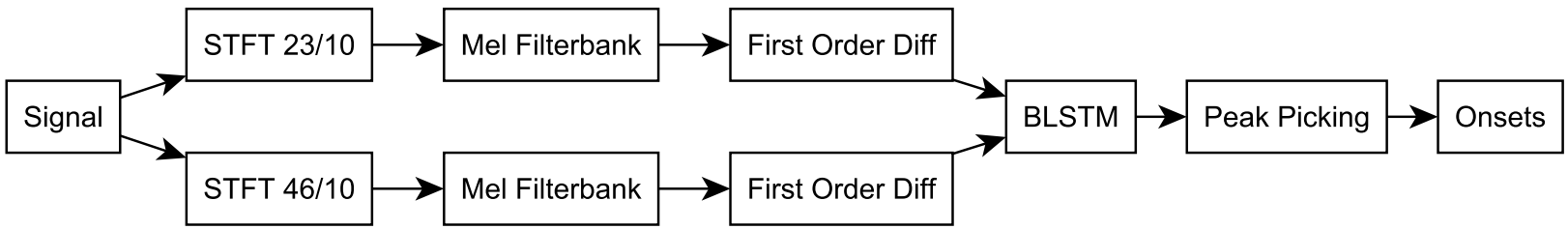
(B)LSTM NETWORKS [8, 7]

- short for “(Bi-Directional) Long-Short Term Memory”
- a special kind of recurrent neural network
- it can “remember” sequences for a longer period of time than simple networks



BLSTM - ONSETS [6]

- do **not** try to come up with a detection function **by hand**
- instead, **learn** the function from annotated **data**
  $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$
- choose $\mathcal{F}$ to be BLSTM networks
- apply standard peak picking and thresholding afterwards



BLSTM - INPUT DATA

- two STFTs with $23|46$ [ms] window size, 10 [ms] hop size
- both STFTs are reduced in dimensionality with a Mel-Scale filterbank
- loudness is converted to logarithmic scale
- **first order differences** of two consecutive Mel spectra are given as additional inputs to the neural network

BLSTM - PEAK PICKING

- the amplitude of the output of the neural network depends only weakly on the amplitude of the signal
- no adaptive thresholding is needed
- simply select the local maxima as onset times

BLSTM - RESULTS

- the method was developed by our group and won the 2010 MIREX Competition in Onset Detection (and again in 2011, and again in 2012 ...)
- see `http://nema.lis.illinois.edu/nema_out/mirex2010/results/aod/summary.html`

BLSTM - RESULTS

- the method was developed by our group and won the 2010 MIREX Competition in Onset Detection (and again in 2011, and again in 2012 ...)
- see `http://nema.lis.illinois.edu/nema_out/mirex2010/results/aod/summary.html`
- it was only superseded by a CNN (Convolutional Neural Network) in 2014 (and again in 2015, 2016 ...)

BLSTM - RESULTS

- the method was developed by our group and won the 2010 MIREX Competition in Onset Detection (and again in 2011, and again in 2012 ...)
- see `http://nema.lis.illinois.edu/nema_out/mirex2010/results/aod/summary.html`
- it was only superseded by a CNN (Convolutional Neural Network) in 2014 (and again in 2015, 2016 ...) also developed by our group [9]

BLSTM - RESULTS

- the method was developed by our group and won the 2010 MIREX Competition in Onset Detection (and again in 2011, and again in 2012 ...)
- see `http://nema.lis.illinois.edu/nema_out/mirex2010/results/aod/summary.html`
- it was only superseded by a CNN (Convolutional Neural Network) in 2014 (and again in 2015, 2016 ...) also developed by our group [9]
- RNN/BLSTM: learnable nonlinear IIR filter
- CNN: learnable nonlinear FIR filter

EVALUATION

EVALUATION (1)

- to evaluate, we need a so called **groundtruth**
- obtaining such a ground truth is usually very laborious (and sometimes impossible)
- for onset detection, the actual time of the onset is needed
- can be obtained via manual annotation (laborious, error prone, disagreement over annotations)
- can be **obtained** by using **special instruments**
  - computer piano
  - electronic wind instruments
  - drum & percussion controllers
  - MIDI pickups for stringed instruments
- “good” evaluation measures (which is not that easy, sometimes)

EVALUATION (2)

- **detected** onset times are **compared** with the ones from the **groundtruth**
- for each **true onset** it is checked whether there was a **predicted onset** within a **small margin around** the true onset
- there are **four** possibilities:

		Predicted	
		Onset	No Onset
True	Onset	TP	FN
	No Onset	FP	TN

EVALUATION (3)

TP	**True Positive**	“hit the onset with the prediction”
FP	**False Positive**	“hit a place where no actual onset is”
TN	**True Negative**	“didn’t hit an onset where there is none”
FN	**False Negative**	“missed an onset where there was one”
P	**Precision**	“proportion of predictions we had correct”
R	**Recall**	“proportion of true onsets we actually hit”
F	**F-measure**	harmonic mean of precision and recall

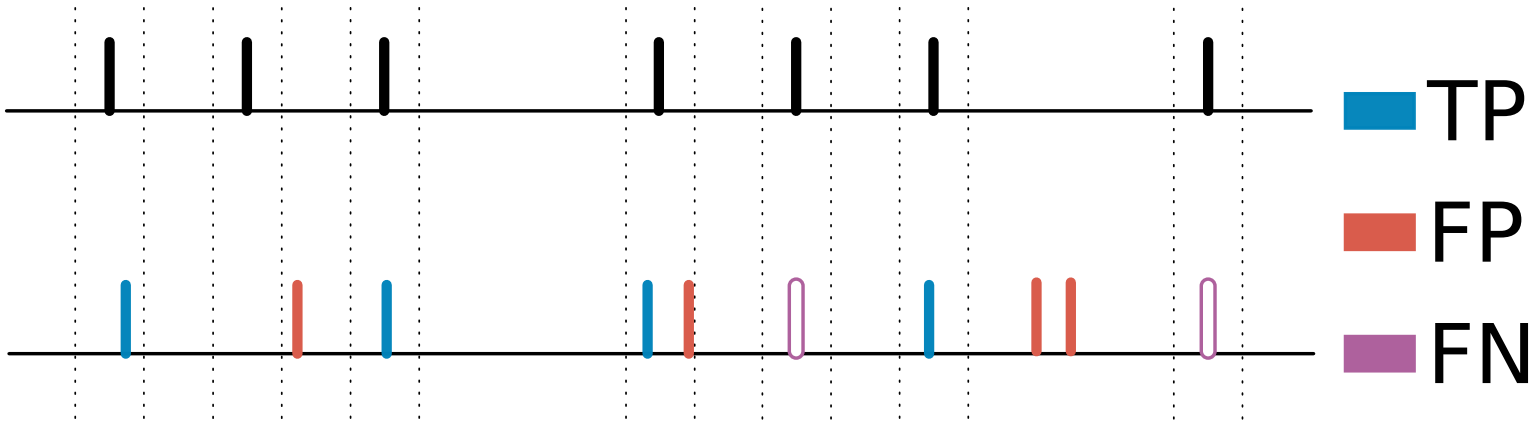
$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

JKU

EVALUATION (4)



NOTE

in the following slides there will be two references
in the tables and graphs, namely “[5]”, and “[21]”

because the tables and graphs were copied over,
these actually refer to [1] and [5].

tables / graphs		this slide deck
“[5]”	$\rightarrow$	[1]
“[21]”	$\rightarrow$	[5]

ABBREVIATIONS

HFC	High Frequency Content
SD	Spectral Difference
PD	Phase Deviation
WPD	Weighted Phase Deviation
NWPD	Normalized Weighted Phase Deviation
WRM	Wavelet Regular Modulus
NLL	Negative Log Likelihood
SF	Spectral Flux
CD	Complex Domain
RCD	Rectified Complex Domain
BLSTM	Bidirectional Long Short-Term Memory

NON-PITCHED PERCUSSIVE

NPP	Precision	Recall	F-measure	TP[%]	FP[%]
HFC [5]	1.000	0.967	**0.983**	96.7	0.0
SD [5]	0.935	0.816	0.871	81.6	5.5
PD [5]	0.934	0.807	0.866	80.7	5.5
WRM [5]	0.974	0.887	0.928	88.7	2.2
NLL[5]	0.980	0.929	0.954	92.9	1.7
SF [21]	0.959	0.975	0.967	97.5	4.2
PD [21]	0.750	0.933	0.831	93.3	31.1
WPD [21]	0.974	0.958	0.966	95.8	2.4
NWPD [21]	0.950	0.966	0.958	96.6	5.2
CD [21]	0.948	0.924	0.936	92.4	5.2
RCD [21]	0.944	0.983	0.963	98.3	5.7
BLSTM $(orig, \omega_{100})$	0.991	0.995	**0.993**	99.5	0.9
BLSTM (mod, ω_{100})	0.991	0.995	**0.993**	99.5	0.9
BLSTM (mod, ω_{50})	0.986	0.986	**0.986**	98.6	1.4
BLSTM $(comb, \omega_{100})$	0.991	0.995	**0.993**	99.5	0.9
BLSTM $(comb, \omega_{50})$	0.986	0.986	**0.986**	98.6	1.4

PITCHED PERCUSSIVE

PP	Precision	Recall	F-measure	TP[%]	FP[%]
HFC [5]	0.947	0.941	0.944	94.1	5.4
SD [5]	0.983	0.949	0.966	94.9	1.6
PD [5]	0.996	0.955	0.975	95.5	0.3
WRM [5]	0.948	0.927	0.937	92.7	5.1
NLL[5]	0.968	0.924	0.945	92.4	3.1
SF [21]	0.981	0.988	**0.984**	98.8	1.8
PD [21]	0.482	0.865	0.619	86.5	93.0
WPD [21]	0.899	0.925	0.912	92.5	5.4
NWPD [21]	0.961	0.981	0.971	98.1	10.4
CD [21]	0.971	0.984	0.978	98.4	2.9
RCD [21]	0.983	0.979	0.981	97.9	1.6
BLSTM $(orig, \omega_{100})$	0.987	0.987	**0.987**	98.7	1.3
BLSTM (mod, ω_{100})	0.992	0.992	**0.992**	99.2	0.8
BLSTM (mod, ω_{50})	0.983	0.979	0.981	97.9	1.7
BLSTM $(comb, \omega_{100})$	0.986	0.993	**0.989**	99.3	1.4
BLSTM $(comb, \omega_{50})$	0.974	0.986	0.980	98.6	2.6

JKU

PITCHED NON-PERCUSSIVE

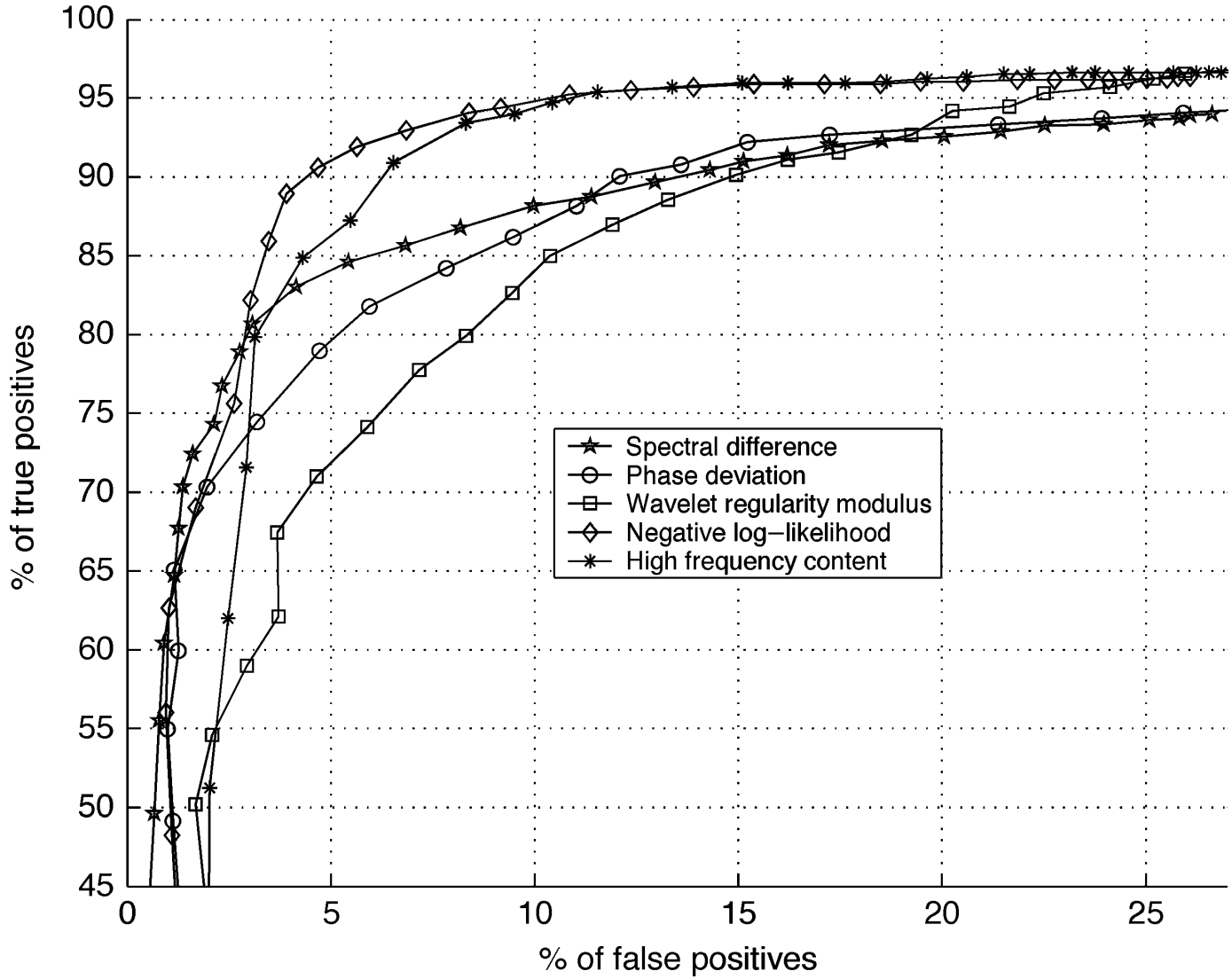
PNP	Precision	Recall	F-measure	TP[%]	FP[%]
HFC [5]	0.844	0.817	0.830	81.7	14.7
SD [5]	0.910	0.871	0.890	87.1	8.6
PD [5]	0.957	0.957	0.957	95.7	4.3
WRM [5]	0.905	0.925	0.915	92.5	10.1
NLL[5]	0.968	0.968	**0.968**	96.8	3.2
SF [21]	0.938	0.968	0.952	96.8	6.5
PD [21]	0.654	0.935	0.770	93.5	49.5
WPD [21]	0.937	0.957	0.947	95.7	6.5
NWPD [21]	0.909	0.968	0.938	96.8	9.7
CD [21]	0.946	0.946	0.946	94.6	5.4
RCD [21]	0.948	0.978	0.963	97.8	5.4
BLSTM $(orig, \omega_{100})$	0.968	0.968	**0.968**	96.8	3.2
BLSTM $(comb, \omega_{100})$	0.968	0.968	**0.968**	96.8	3.2
BLSTM $(comb, \omega_{50})$	0.939	0.939	0.939	93.9	6.1

MIXED ONSET TYPES

MIX	Precision	Recall	F-measure	TP[%]	FP[%]
HFC [5]	0.888	0.845	0.866	84.5	10.8
SD [5]	0.886	0.804	0.843	80.4	10.4
PD [5]	0.764	0.801	0.782	80.1	24.7
WRM [5]	0.768	0.819	0.793	81.9	24.7
NLL[5]	0.889	0.860	0.874	86.0	10.8
SF [21]	0.882	0.882	**0.882**	88.2	11.8
PD [21]	0.663	0.749	0.704	74.9	38.0
WPD [21]	0.843	0.830	0.836	83.0	15.5
NWPD [21]	0.916	0.845	0.879	84.5	7.7
CD [21]	0.941	0.819	0.876	81.9	5.2
RCD [21]	0.945	0.819	0.877	81.9	4.8
BLSTM $(orig, \omega_{100})$	0.941	0.897	**0.918**	89.7	5.6
BLSTM (mod, ω_{100})	0.947	0.930	**0.938**	93.0	5.3
BLSTM (mod, ω_{50})	0.921	0.896	**0.909**	89.6	7.9
BLSTM $(comb, \omega_{100})$	0.938	0.918	**0.928**	91.8	6.2
BLSTM $(comb, \omega_{50})$	0.907	0.878	**0.893**	87.8	9.3

JKU

RESULTS (“[5]” [1] ONLY)



JKU

LFSF ONLINE SHOOTOUT

"LogFiltSpecFlux" vs. "The Rest" in an **online** setting
dataset contains ~ 26K onsets

Online algorithm	% F-meas.	% Prec.	% Rec.
SF	74.5	76.3	72.8
SF aw	75.7	78.0	73.4
SF log	76.1	78.3	74.0
SF log filtered	**80.3**	**88.3**	73.5
CD	71.1	72.4	69.8
CD aw	75.8	76.4	**75.1**
CD log	74.1	77.4	71.1
WPD	69.7	68.8	70.6
WPD aw	71.4	70.8	72.0
WPD log	70.9	74.6	67.5
OnsetDetector.LL [4]	81.7	85.0	78.7

JKU

CONCLUSIONS

ONLINE VS. OFFLINE

- **online** onset detection seems much **harder than offline**
- no peeking at the future
- this **restricts** smoothing and peak detection
- we risk more false positives
- we have to accept some delay in the detection of onsets

IS ONSET DETECTION SOLVED?

- basic onset detection ideas are **simple**
- current **state of the art** may be very **complicated**
- **basic** algorithms already provide **solid performance**
- **certain** types of music has **instrumentation** which makes onset detection **difficult**
- there is always room for improvement!
- onset detection provides a basis for many music analysis tasks, like **beat tracking**, **music transcription**, **score following** ...

MAIN SOURCES

- some of the main sources for this lecture were not explicitely cited, because they would have to be cited everywhere
- there are **lots** of references in the cited paper's own reference sections
- don't forget about the papers with state-of-the-art approaches!

REFERENCES I

[1] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler.
A tutorial on onset detection in music signals.
IEEE Transactions on Speech and Audio Processing, 13(5):1035–1047, 2005.

[2] Sebastian Böck.
Onset, beat, and tempo detection with artificial neural nets.
Master's thesis, TU München, 2010.

REFERENCES II

[3] Sebastian Böck, Florian Krebs, and Markus Schedl.
Evaluating the online capabilities of onset detection methods.
In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012*, pages 49–54, 2012.

[4] Sebastian Böck and Gerhard Widmer.
Maximum filter vibrato suppression for onset detection.
In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland*, pages 55–61, 2013.

REFERENCES III

[5] Simon Dixon.
Onset detection revisited.
In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06), Montréal, Québec, Canada*, pages 133–137, 2006.

[6] Florian Eyben, Sebastian Böck, Björn W. Schuller, and Alex Graves.
Universal onset detection with bidirectional long short-term memory neural networks.
In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, pages 589–594, 2010.

REFERENCES IV

[7] Alex Graves and Jürgen Schmidhuber.
Framewise phoneme classification with bidirectional lstm and other neural network architectures.
Neural Networks, 18(5):602–610, 2005.

[8] Sepp Hochreiter and Jürgen Schmidhuber.
Long short-term memory.
Neural Computation, 9(8):1735–1780, 1997.

[9] Jan Schlüter and Sebastian Böck.
Improved musical onset detection with convolutional neural networks.
In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6979–6983, Florence, Italy, May 2014.

REFERENCES V

[10] Hui Li Tan, Yongwei Zhu, and Lekha Chaisorn.
An energy-based and pitch-based approach to audio onset detection.
MIREX (2009), Onset Detection Contest. (cited on pages 52, 82, 94, 95, 96, 97, and 99), 2009.

[11] Ronald J Williams and David Zipser.
A learning algorithm for continually running fully recurrent neural networks.
Neural Computation, 1(2):270–280, 1989.