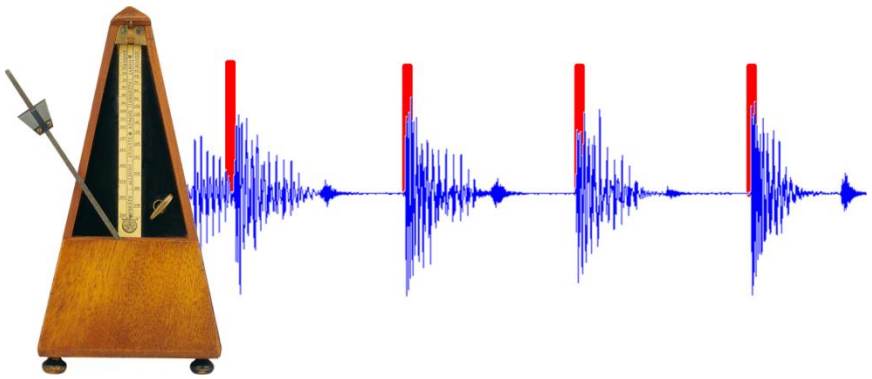
Onset detection

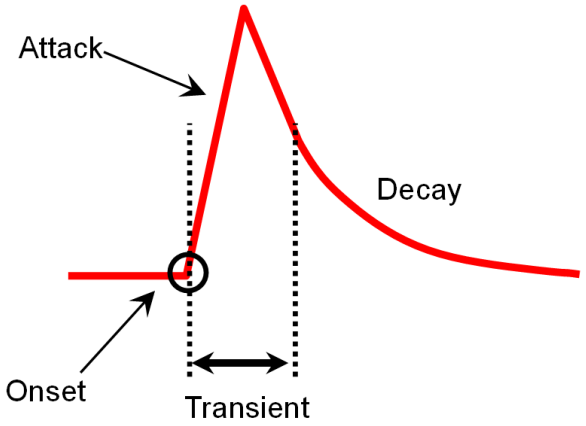
Li Su

2019/04/16

Onset detection 通常看 50 ms 以内算 f-measure

- Recall the attack-decay-sustain-release (ADSR) curve
- Transient: the noise-like sound component of short duration and high amplitude typically occurring at the beginning of a musical tone
- Onset: the instant marking the start of the tranasient





Energy-based novelty (1)

新東西出來

- Playing a note on an instrument often coincides with a sudden increase of the signal's energy
- Local energy: given a window function $h(m)$ supported on $m \in [-M, M]$, we have

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

- Energy-based novelty function:

相當於local energy的一次微分

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

時間

- Half-wave rectification function $|\cdot|_{\geq 0}$:

半波整流(ReLu)

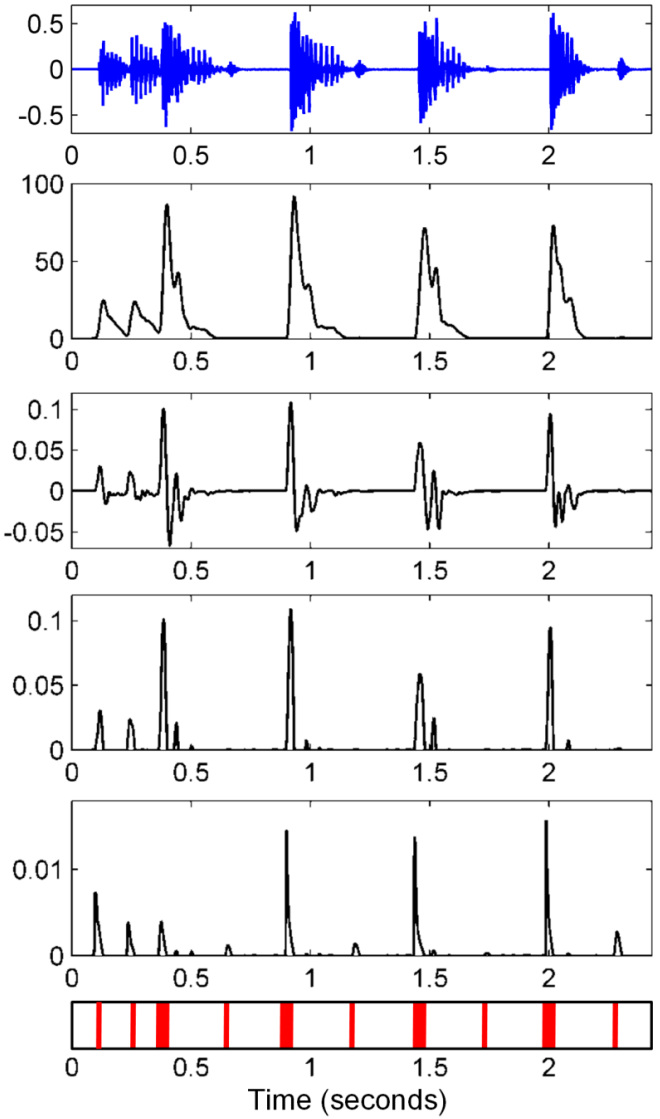
$$|r|_{\geq 0} := \frac{r + |r|}{2} = \begin{cases} r, & if\ r \geq 0 \\ 0, & if\ r < 0 \end{cases}$$

Energy-based novelty

- Human perception of sound intensity is logarithmic in nature
- Log-energy-based novelty function:

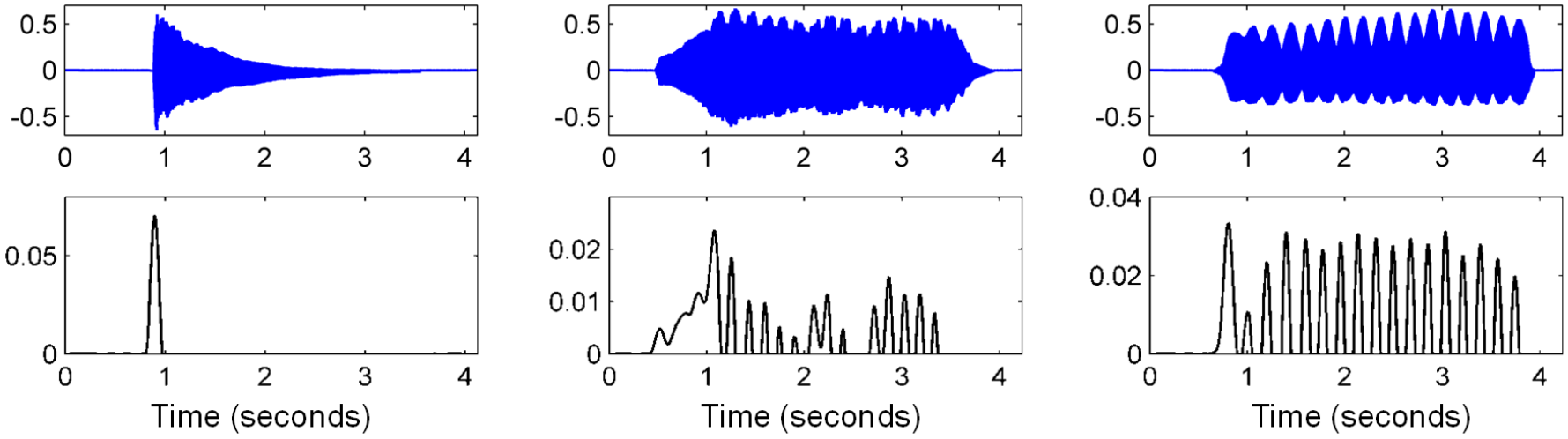
$$\Delta_{\text{Energy}}^{\text{Log}}(n) = |E_h^x(n+1) - E_h^x(n)|_{\geq 0}$$

- Example:



Energy-based novelty (3)

- Waveform and energy-based novelty function of the note C4 (261.6 Hz) played by different instruments – piano (left), violin (middle) and flute (right)



Recap: short-time Fourier transform

- Given a discrete-time signal $x(t)$ sampled at a rate f_s. Let window size N samples, hop size H samples, then the short-time Fourier transform (STFT) $X(n,k)$ is:

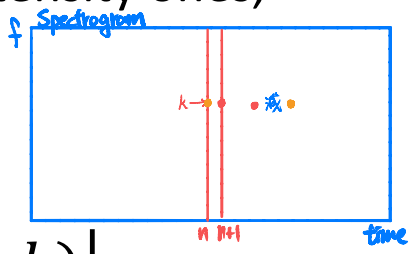
$$X(n,k) = \sum_{m=0}^{N-1} x(m+nH)h(m)e^{-\frac{j2\pi km}{n}}$$

- k: frequency index, $f(k) \coloneqq \frac{kf_s}{N}$
- n: time index , $t(n) \coloneqq \frac{nH}{f_s}$
- Spectrogram: $|X(n,k)|^2$
- Logarithmic compression:

$$Y_\gamma(n,k) \coloneqq \log(1+\gamma|X(n,k)|)$$

Spectral-based novelty (1)

- Energy-based novelty function falls short of:
  - Pitch change, low-intensity note masked by high-intensity ones, frequency-dependent transient, and others …

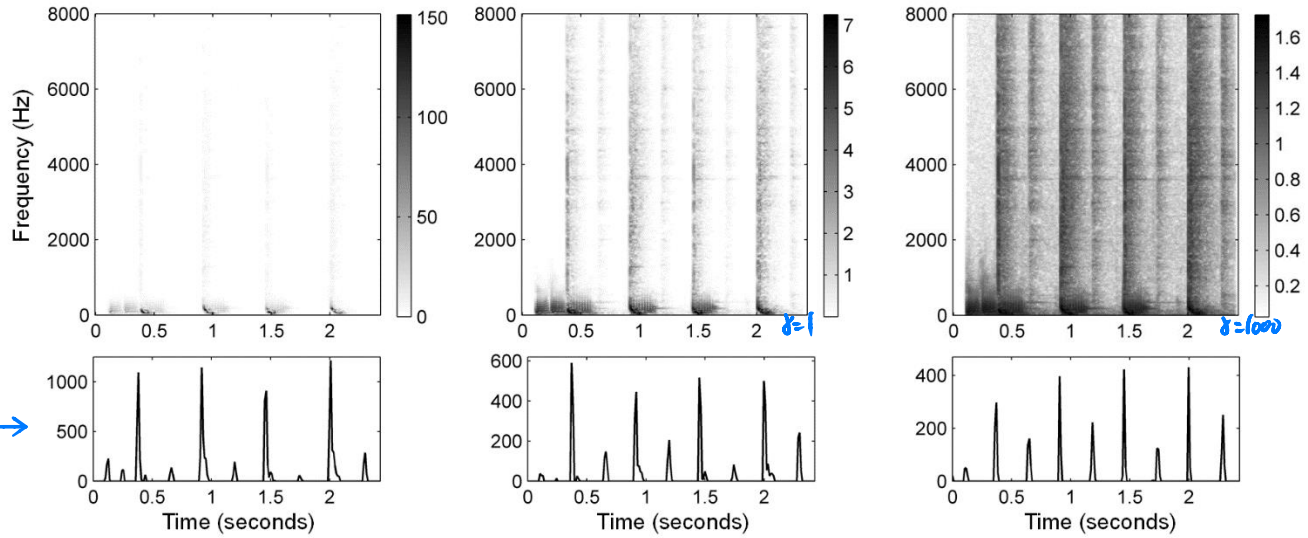


- Spectral flux
  通量

$$\Delta_{\text{Spectral}}(n) := \sum^{K} \left| Y_\gamma(n+1,k) - Y_\gamma(n,k) \right|_{\geq 0}$$

频谱 频率的 index 半波整流

- Example:
  - Spectrogram
  - $\gamma = 1$
  - $\gamma = 1000$



spectral flux curve →

Spectral-based novelty (2): post-processing

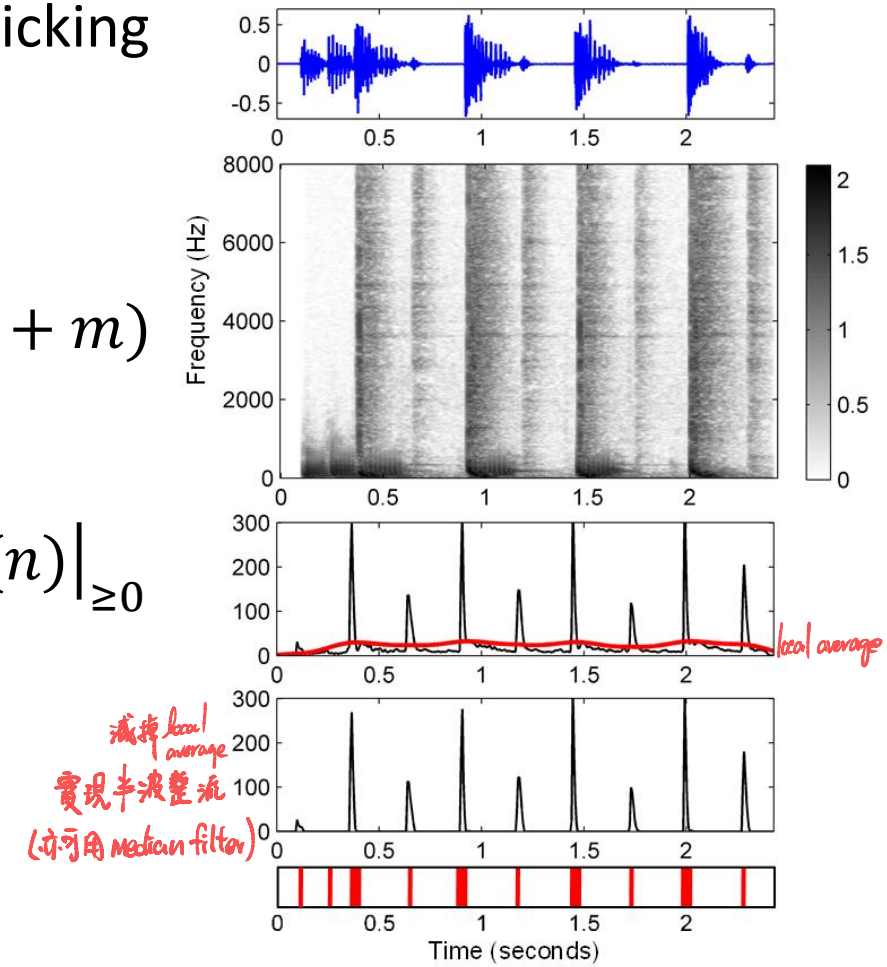
- Adaptive thresholding and peak picking
  - Basic idea: moving average
- Local average function:

$$\mu(n) := \frac{1}{2M+1} \sum_{m=-M}^{M} \Delta_{\text{Spectral}}(n+m)$$

- Enhanced novelty function:

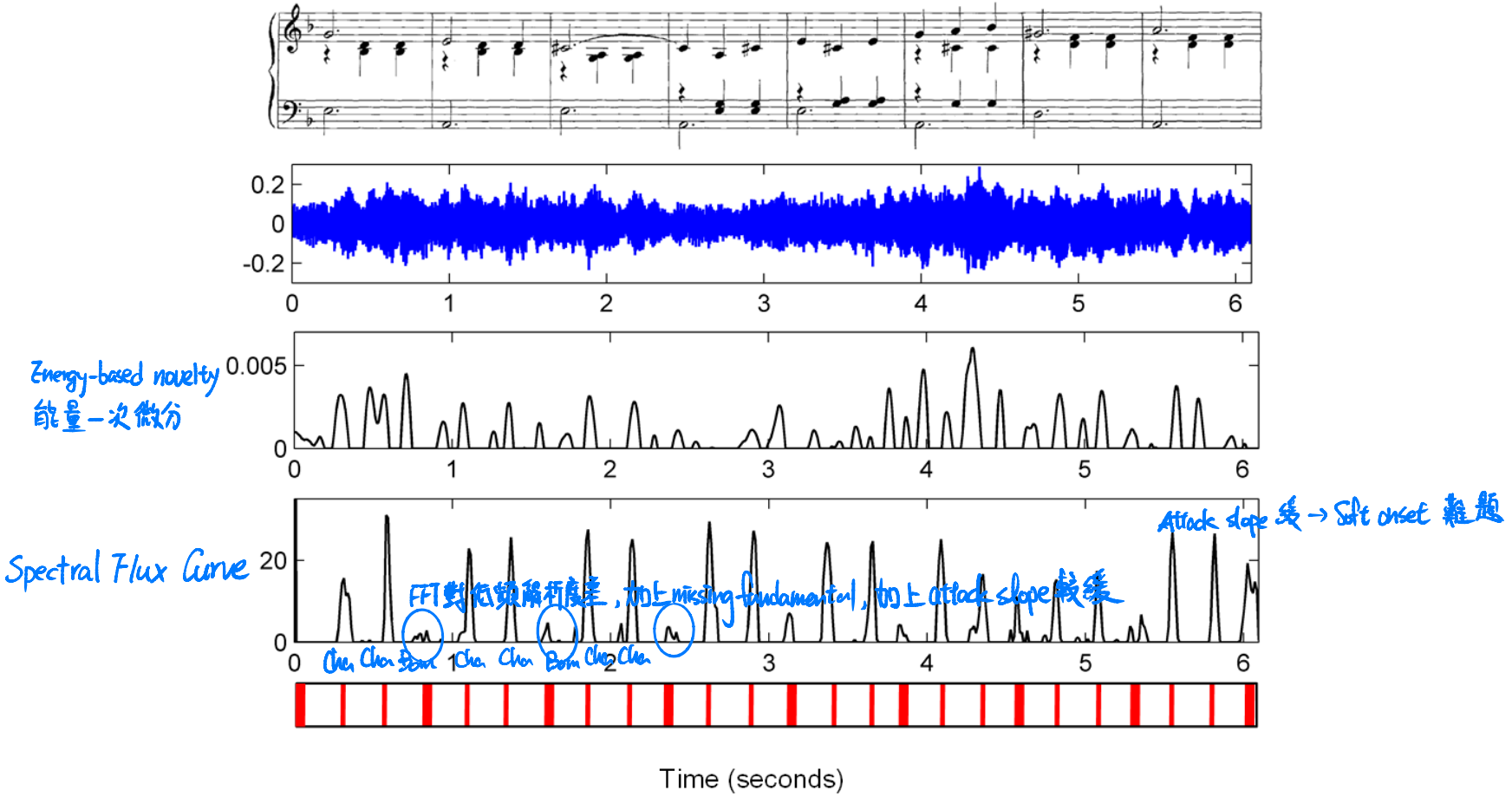
$$\overline{\Delta}_{\text{Spectral}}(n) := \left|\Delta_{Spectral}(n) - \mu(n)\right|_{\geq 0}$$

- Other ideas: median filtering



Spectral-based novelty (3)

- Shostakovich's Waltz No. 2



Spectral-based novelty (4)

較好處理

- Percussive onsets (e.g., percussion instrument, piano) can be considered a solved problem, but soft onsets, vibrato and tremolo are still major challenges
- The variation of frequency in the vibrato are easily considered as onset using the spectral flux, while the variation of amplitude in the tremolo are easily considered as onset using energy-based novelty
- How to improve (Böck et. al, 2013): (1) consider longer time difference $\mu > 1$, (2) maximum filtering

抖音 很不適合 spectral flux

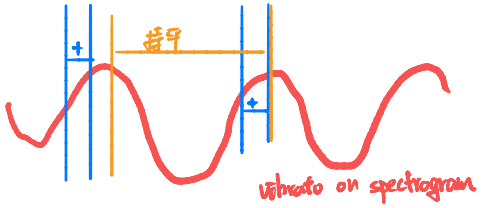
考慮較大的間隔

(看兩一個frame或多個的一個頻帶) 前一個frame的頻帶的最大值

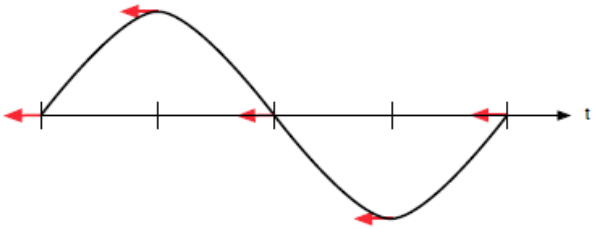
$$SF\left(n+\frac{\mu}{2}\right)=\sum_{k=0}^{K}\left|Y_\gamma(n+\mu,k)-\max_{k-\eta\le k'\le k+\eta}Y_\gamma(n,k)\right|_{\ge 0}$$

Super Flux

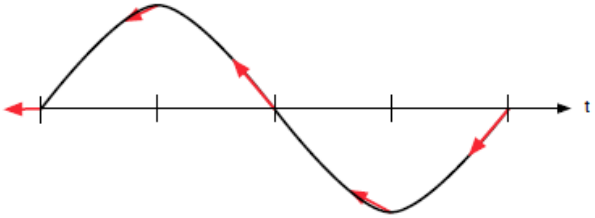
Spectral-based novelty (5)



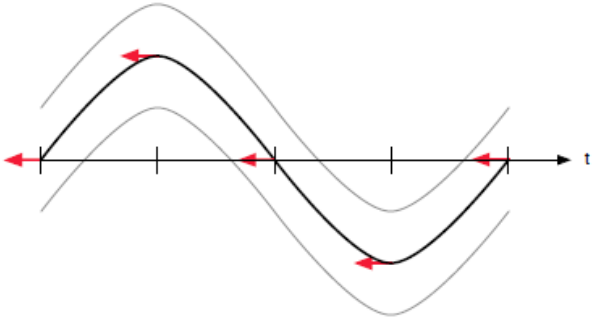
- Vibrato suppression



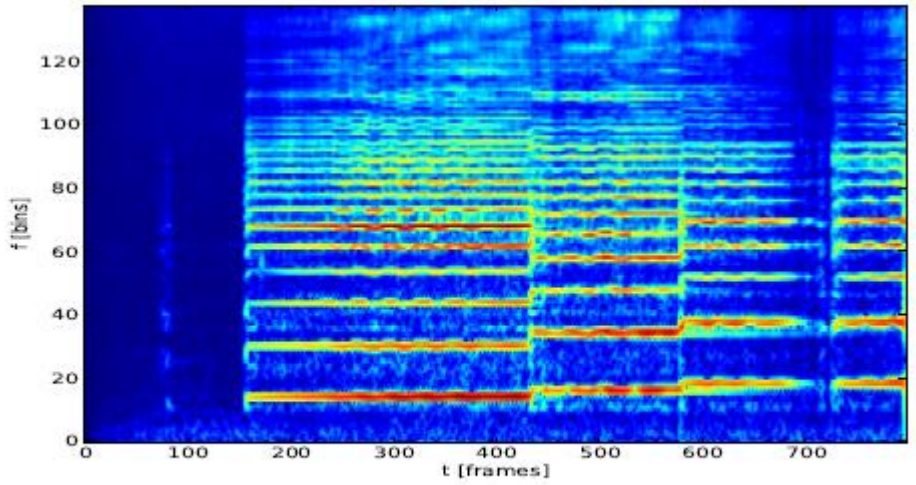
(a) *classical bin-wise difference calculation*

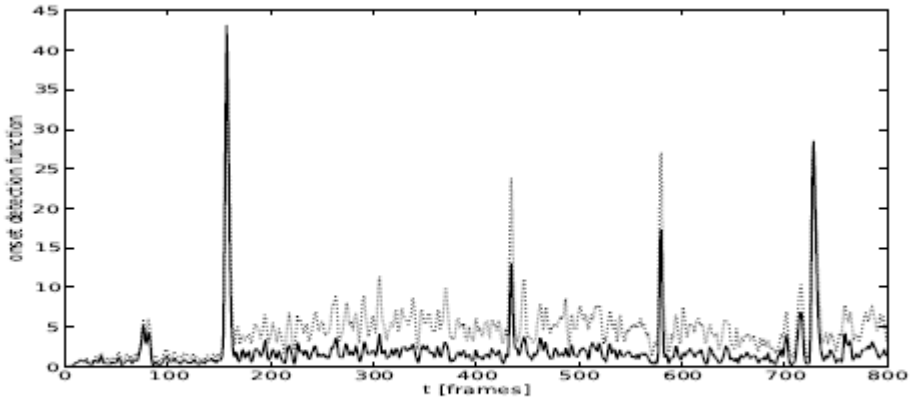


(b) *trajectory tracking-based difference calculation*



(c) *maximum filter-based difference calculation*





(d) *sum of differences*

Phase-based novelty

- Phase is also important
  - Stationary tones have a stable phase (i.e., evolves linearly with time), while transients have a unstable phase
- Polar coordinate representation

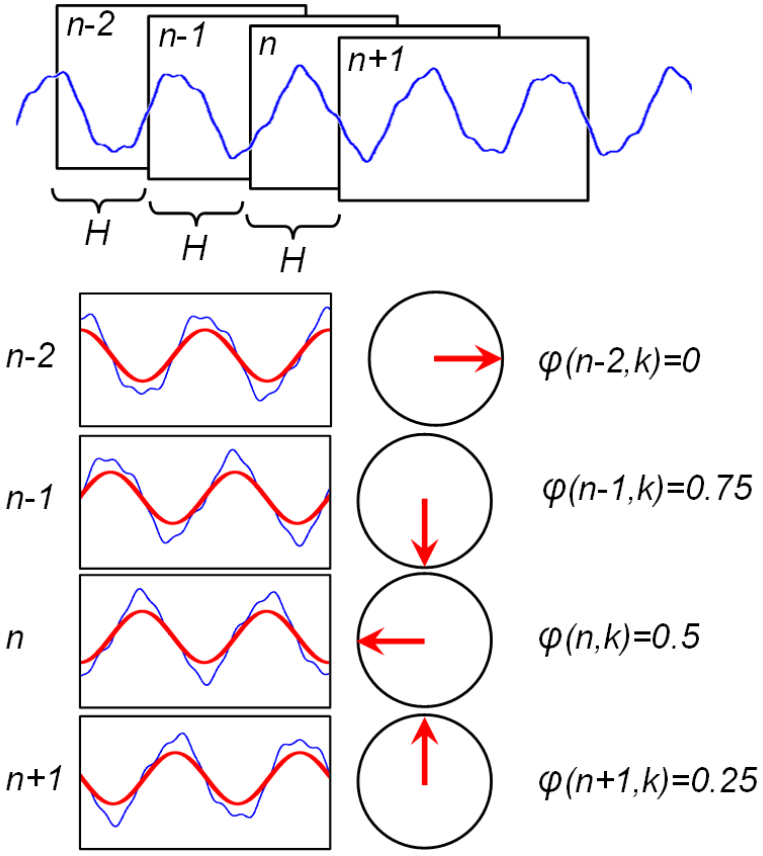
$$X(n,k) = |X(n,k)|e^{2\pi i\phi(n,k)}$$

- Phase derivative
  - $\phi'' = 0$ when steady state

$$\phi'(n,k) := \phi(n,k) - \phi(n-1,k)$$
$$\phi''(n,k) := \phi'(n,k) - \phi'(n-1,k)$$

- Phase-based novelty function

$$\Delta_{\text{Phase}} = \sum_{k=0}^{K} \phi''(n,k)$$



Complex-domain novelty

- Problems in phase-based novelty
  - Phase jumps from $-\pi$ to π, needs a procedure called phase unwrapping (unstable) 能量小 phase 亂，但能量小不應作為 onset 參考
  - When $X(n,k)$ is very small, $\phi(n,k)$ could be very chaotic (large $\phi''(n,k)$)
- Considering both magnitude and phase

$$\hat{X}(n+1,k) = |x(n,k)|e^{2\pi i\left(\phi(n,k)+\phi'(n,k)\right)}$$

??

$$\hat{X}^{+}(n,k) = \begin{cases} \left|\hat{X}(n,k) - X(n+1,k)\right| & \text{for } |X(n,k)| > |X(n-1,k)| \\ 0 & \text{otherwise} \end{cases}$$

- Complex-domain novelty function:

$$\Delta_{\text{Complex}}(n,k) = \sum_{k=0}^{K} X^{+}(n,k)$$

General strategies in modeling *causal* time-domain behaviors (1)

現代作法

現在只和過去有關：可能可以被預測

- Causality: the Now is determined by the Past rather than the Future; time never goes back

$$x(n,k) = f(x(n,1), x(n,2), \cdots x(n,k-1))$$

- Methods that model causal time sequences:
  - Ordinary differential equation (ODE): for continuous domain

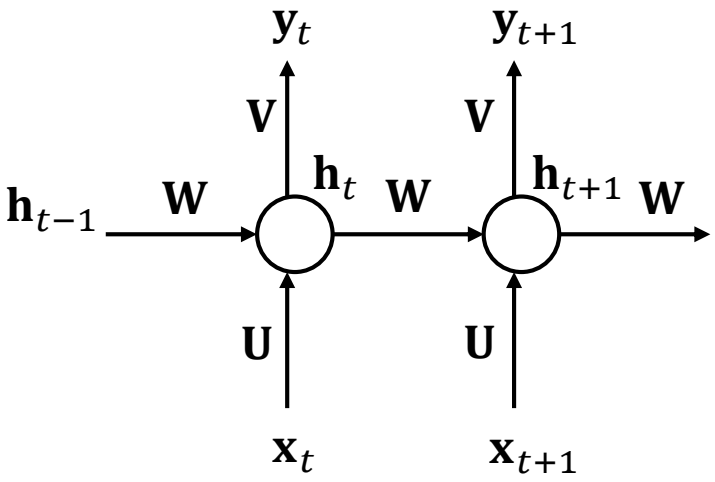
$$\frac{df(t)}{dt} + \frac{d^2 f(t)}{d^2 t} + \cdots = g(t)$$

  - Linear prediction (or, ODE in the discrete case / linear auto-regression): for discrete domain

$$x(n,k) = \sum_{i=1}^{K} a_i x(n,k-i)$$

General strategies in modeling *causal* time-domain behaviors (2)

- Methods that model causal time sequences:
- Kalman filters
- Recurrent neural networks (RNN) 隨著時間變化都用同一組參數來描述訊號的關係
- $\mathbf{h}_t = \sigma(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{b})$
- $\mathbf{y}_t = g(\mathbf{V}\mathbf{h}_t)$
- Variants of RNN: long-short term memory (LSTM), GRU, …

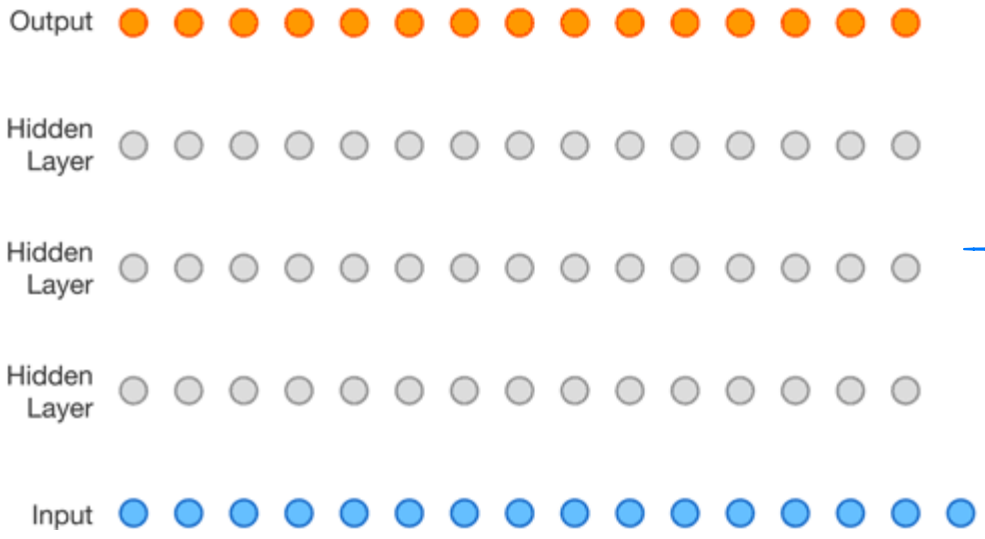


General strategies in modeling *causal* time-domain behaviors (3)

- Autoregressive models (e.g., WaveNet)
- $p(x_1, x_2, \ldots, x_T) = \prod_{t-1}^{T} p(x_t | x_1, x_2, \ldots, x_{t-1})$
- Sequence-to-sequence models
- Attention models

for generation

for onset detection or beat tracking



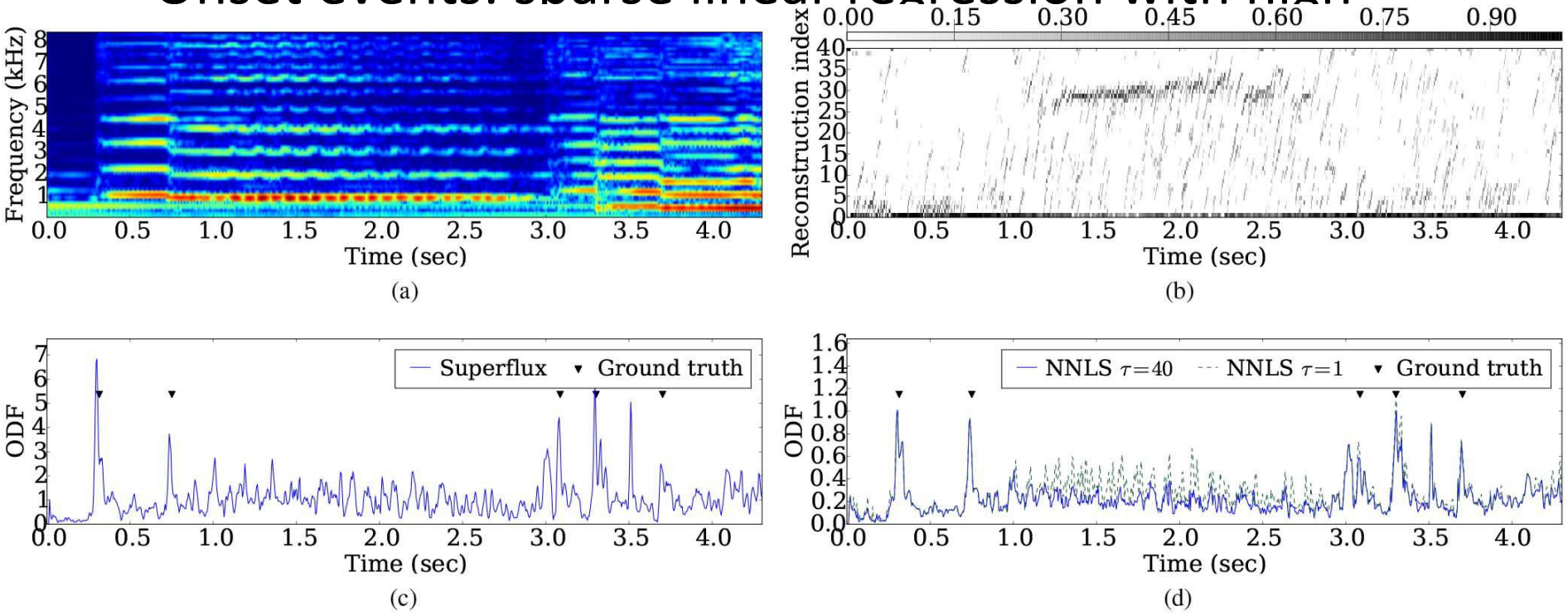
→ WaveNet Illustration

Sparse linear prediction for onset detection

$x(t) = \sum_{i=1}^{M} \alpha_i x(t-i)$ 很适合正弦波等稳定的讯号

Transient 很新被 linear prediction reconstruct，所以可以看 reconstruction loss 分辨 onset

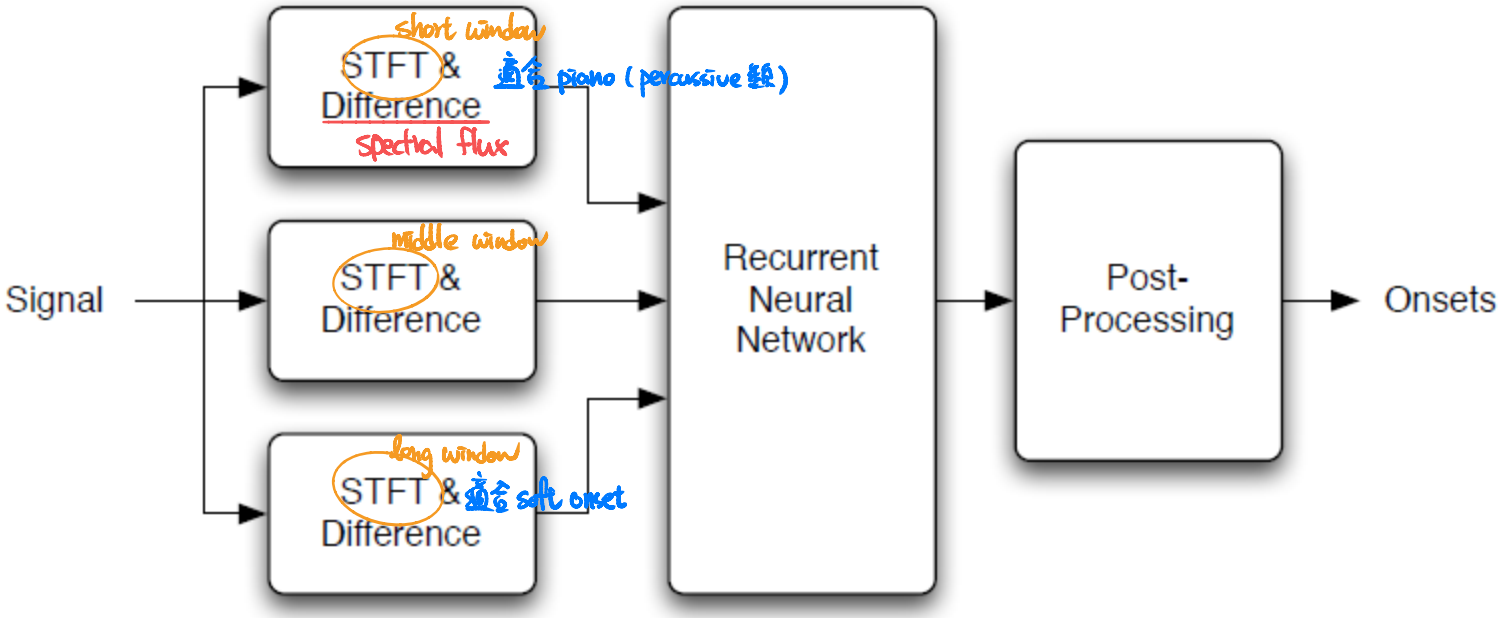
- Onsets are those which are “hard to be predicted”
- Onset events: sparse linear regression with high



Che-Yuan Liang, Li Su, and Yi-Hsuan Yang. "Musical onset detection using constrained linear reconstruction." *IEEE Signal Processing Letters* 22.11 (2015): 2142-2146.

Neural networks for onset detection

- LSTM-based recurrent neural networks (LSTM-RNN)
- State-of-the-art



Böck, Sebastian, et al. "Online real-time onset detection with recurrent neural networks." *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK*. 2012. state-of-the-art till now

Offset detection

很難介定

通常會改討論 semantic 問題（音符種類4分,8分,附點?）

- More difficult; less investigated
- Some possible directions:
  - Energy threshold
  - Pitch confidence
  - State prediction
- Depend on the instrument and playing technique!

- Reference: Che-Yuan Liang, Li Su, Yi-Hsuan Yang, Musical offset detection of pitched instruments: the case of violin, ISMIR 2015

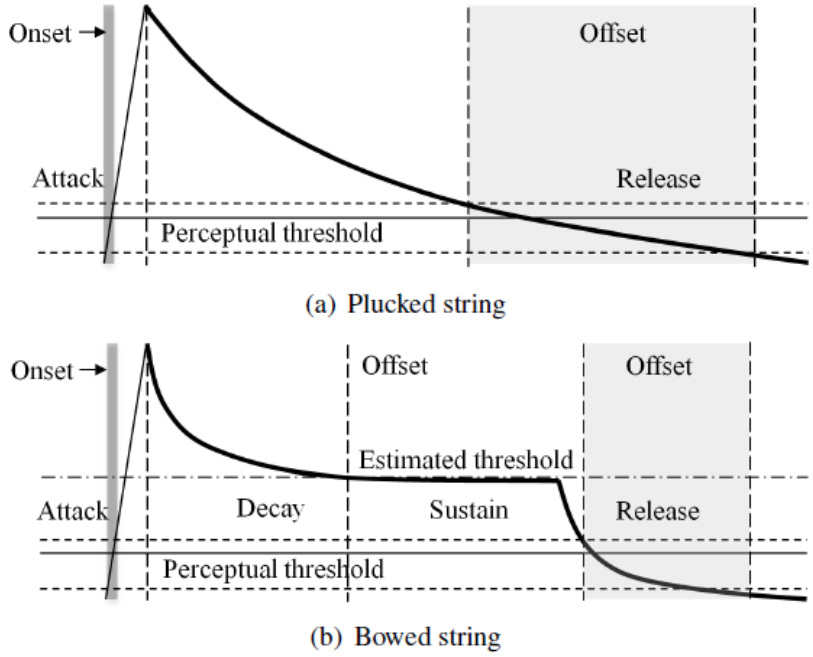


Figure 1. The ADSR envelops of a plucked string (upper) and a bowed string signal (lower). The gray blocks show the ambiguity of onset (dark) and offset (light) due to the variation of hearing threshold. The bold-line segments of the envelopes are the possible regions to detect an offset.

Further readings

- Sebastian Böck and Gerhard Widmer, “Maximum filter vibrato suppression for onset detection”, Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13), 2013
- S. Dixon, “Onset detection revisited,” in Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06), Montreal, Quebec, Canada, September 2006, pp. 133–137.
- A. Holzapfel, Y. Stylianou, A.C. Gedik, and B. Bozkurt, “Three dimensions of pitched instrument onset detection,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1517–1527, 2010.