Music sound and timbre features

Li Su

March 4, 2019

Four essences of audio signals

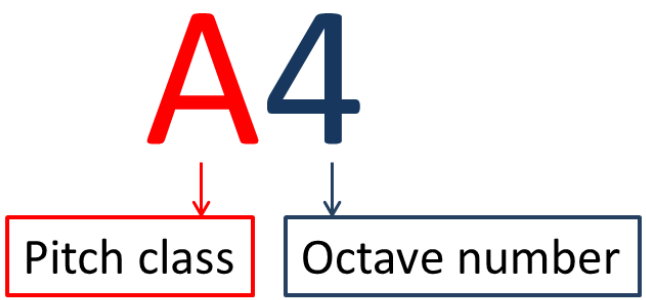
- Pitch
- Loudness
- Timbre
- Direction

Frequency and pitch

- The higher the frequency of a sinusoidal wave, the higher it sounds
- Human's audible frequency: 20 Hz – 20,000 Hz (20 kHz)
- Dog's: $\sim$ 45 kHz; cat's: $\sim$ 64 kHz
- Ultrasound: $>$ 20 kHz; infrasound: $<$ 20 Hz

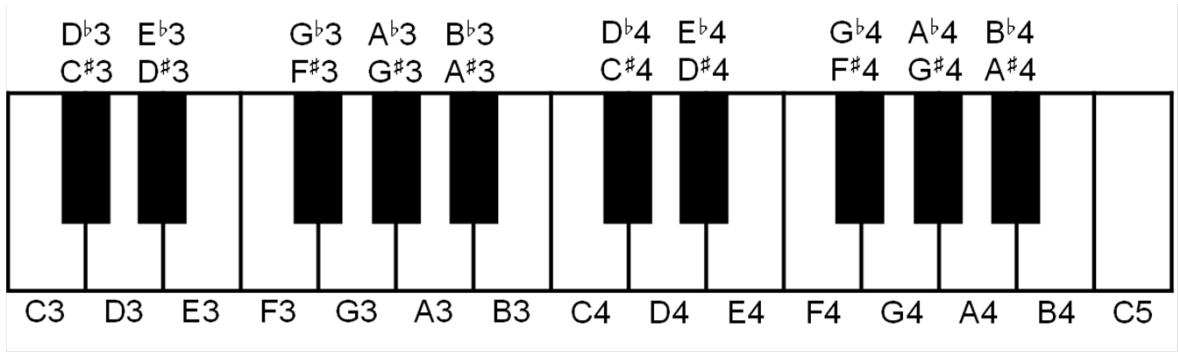
Scientific pitch notation and MIDI number

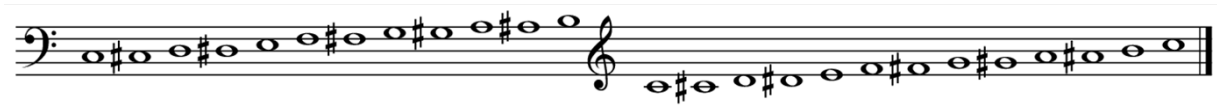
- Musical Instrument Digital Interface (MIDI): 21 – 108 for piano $\in [0, 127]$
- Concert pitch: A4 = 440 Hz
- ▸ Reference



F0 = 440 Hz

MIDI = 69





From: M. Mueller, *Fundamentals of Music Processing*, Chapter 1, Springer 2015

Pitch

頻率2倍高一個度

- Octave equivalence: two frequencies differing by a power of 2 sounds similar
- Semitone: two frequencies (i.e., f_1 and f_2, $f_1 > f_2$) differ by 1 semitone when their ratio is $f_1/f_2 = 2^{\frac{1}{12}} \approx 1.059463$
- One octave contains 12 semitones 半音(12平均率)
- The center frequency $F_{pitch}(p)$ of each pitch with MIDI $= p$ is

$$F_{pitch}(p) = 440 \times 2^{\frac{p-69}{12}} \quad (1)$$

A4 = 69 in MIDI

- Example: we have $F_{pitch}(p+12) = 2F_{pitch}(p)$, $\frac{F_{pitch}(p+1)}{F_{pitch}(p)} = 2^{\frac{1}{12}} \approx 1.059463$

Dynamic, loudness, and intensity

- Dynamic: a term referring to the musical symbols that indicate the volume, like *forte* (f) or *piano* (p)
- Loudness: a perceptual, subjective property, depending on sound intensity, duration and frequency, where the sound can be ordered from quite to loud
- Intensity: a physical property, defined as the sound power per unit area (e.g., W/m^2)
- Threshold of hearing (TOH): the minimal sound intensify of a pure tone (i.e., a sinusoid) a human can hear, $I_{TOH} := 10^{-12} W/m^2$
- Threshold of pain (TOP): $I_{TOH} := 10 W/m^2$
- dB-scaled sound intensity: $\mathrm{dB}(I) = 10 \log_{10} \left(\frac{I}{I_{TOH}} \right)$

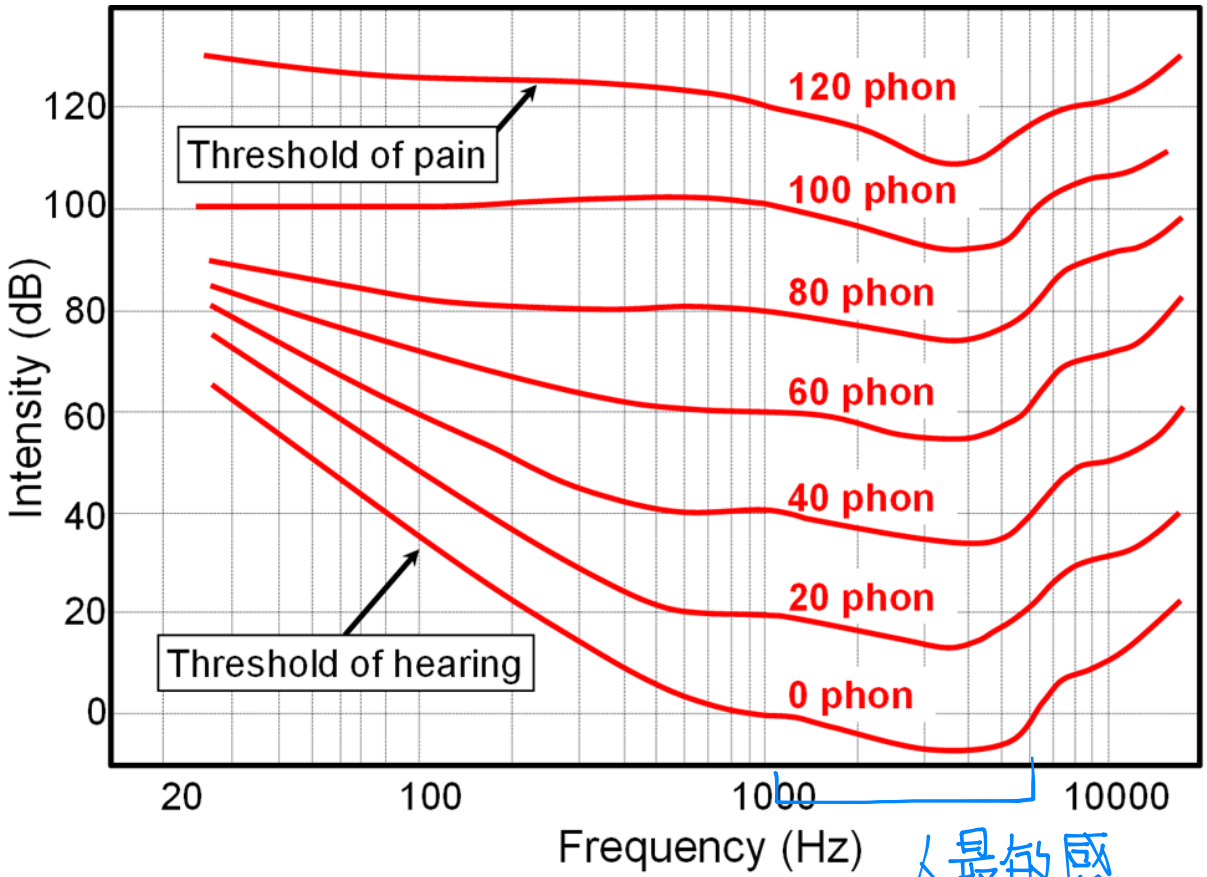
Sound intensity

Source	Intensity	Intensity level	× TOH
Threshold of hearing (TOH)	10^{-12}	0 dB	1
Whisper	10^{-10}	20 dB	10^{2}
Pianissimo	10^{-8}	40 dB	10^{4}
Normal conversation	10^{-6}	60 dB	10^{6}
Fortissimo	10^{-2}	100 dB	10^{10}
Threshold of pain	10	130 dB	10^{13}
Jet take-off	10^{2}	140 dB	10^{14}
Instant perforation of eardrum	10^{4}	160 dB	10^{16}

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 1, Springer 2015

Equal loudness curve

- Loudness is highly correlated with intensity
- Human ears are most sensitive to sounds around 2–4 kHz
- Frequency-dependent unit: *phon*



From: M. Mueller, *Fundamentals of Music Processing*, Chapter 1, Springer 2015

Timbre

- Timbre is the attribute whereby a listener can judge two sounds as dissimilar using any criterion other than pitch and loudness
- Timbre information allows us to tell apart the sounds of a violin, oboe and trumpet, even when the pitch and loudness of them are the same
- Words describing timbre: bright, dark, warm, harsh, cold, ...

Musical instrument families

No unified categories for music instrument families. In common sense:

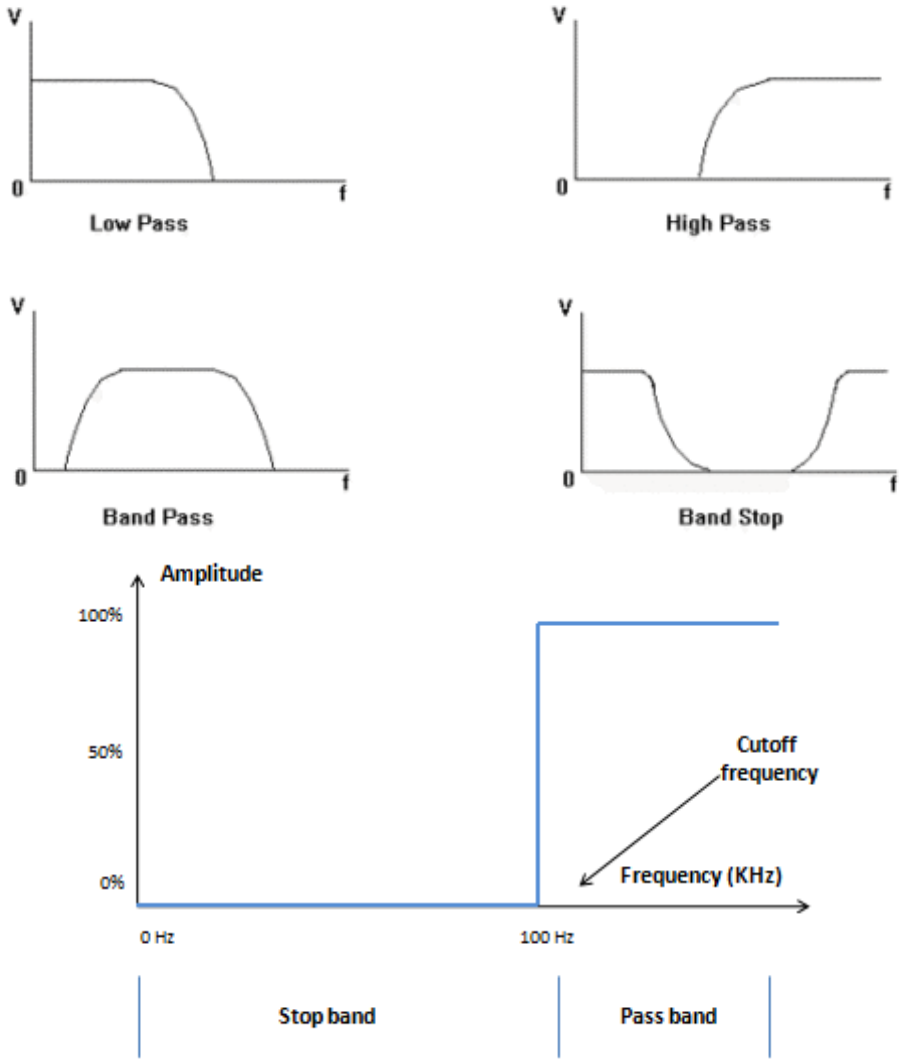
- Strings: violin, cello, guitar, ...
- Brass: trumpet, trombone, horn, ...
- Woodwind: clarinet, oboe, bassoon, ...
- Percussion: drum, cymbal, hi-hat, xylophone, ...

The Hornbostel-Sachs system

- Idiophone: produce sound by vibrating themselves
- Membranophone: produce sound by a vibrating membrane
- Chordophone: produce sound by vibrating strings
- Areophone: produce sound by vibrating air
- (New) electrophone: produce sound by electronic signal

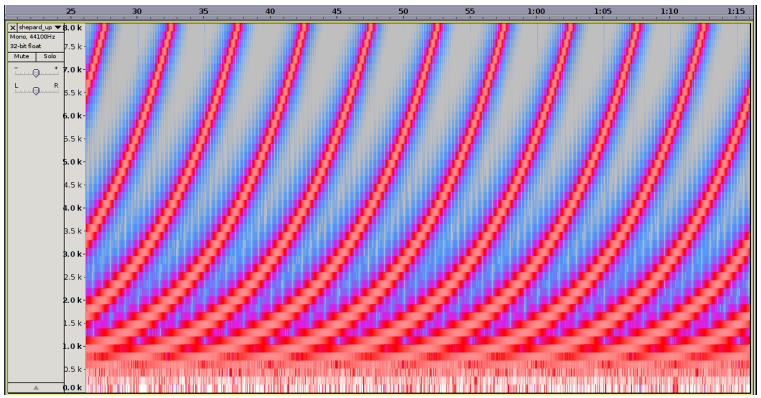
Digital audio effects: filter

- Suppress of remove specific components in a given frequency band
- Example: what will happen if we use a high-pass filter (e.g., suppress low-frequency components) on a signal?
  - Original
  - Cut-off frequency = 100 Hz
  - Cut-off frequency = 200 Hz
  - Cut-off frequency = 500 Hz
  - Cut-off frequency = 1000 Hz



Digital audio effects: flanging

- Flanging: combining two identical signals together, with a small time difference (around 20 ms)
- Behaves like a comb filter: ▸ About comb filter
- The history of flanging: ▸ Link
- Other audio effects (e.g., phasing, chorus effect, etc.): visit Wikipedia for resources
- "Infinite flanging": the Shepard tone effect (the sonic barber pole) ▸ Audio



Energy features

The instantaneous RMS energy 能量

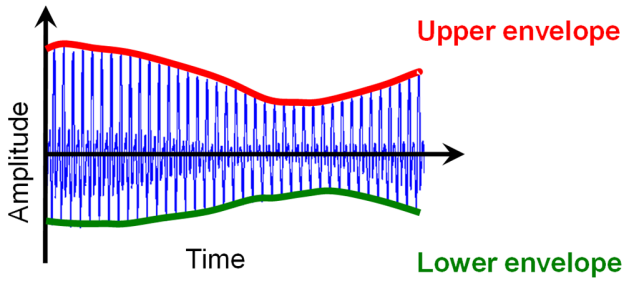
$$E(n) = \sqrt{\frac{1}{2N+1} \sum_{i=-N}^{N} x(n+i)^2} \tag{2}$$

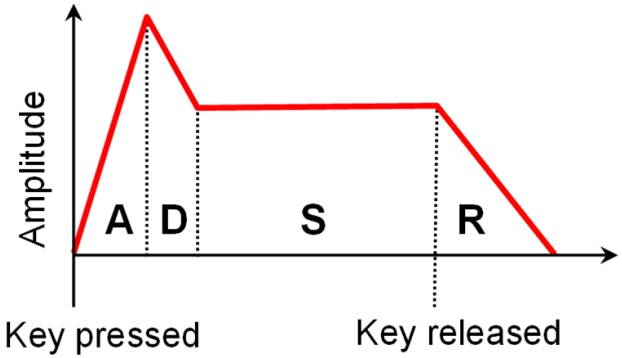
取 2N+1 個訊號

The ADSR curve

onset 起音點 50ms 內誤差從計算上可接受
offset 結束點

- Temporal dynamics of sounds are very critical to the perception of timbre
- A general model of the temporal amplitude envelope 包波
- Attach-Decay-Sustain-Release (ADSR) 很重要
- RMS amplitude envelope: low-pass filtering $E(n)$ with cut-off frequency around 30 Hz
- Other methods?





From: M. Mueller, *Fundamentals of Music Processing*, Chapter 1, Springer 2015

Attack time

LAT：木琴，鐵琴 < 鋼琴 < 小提琴 < 大提琴

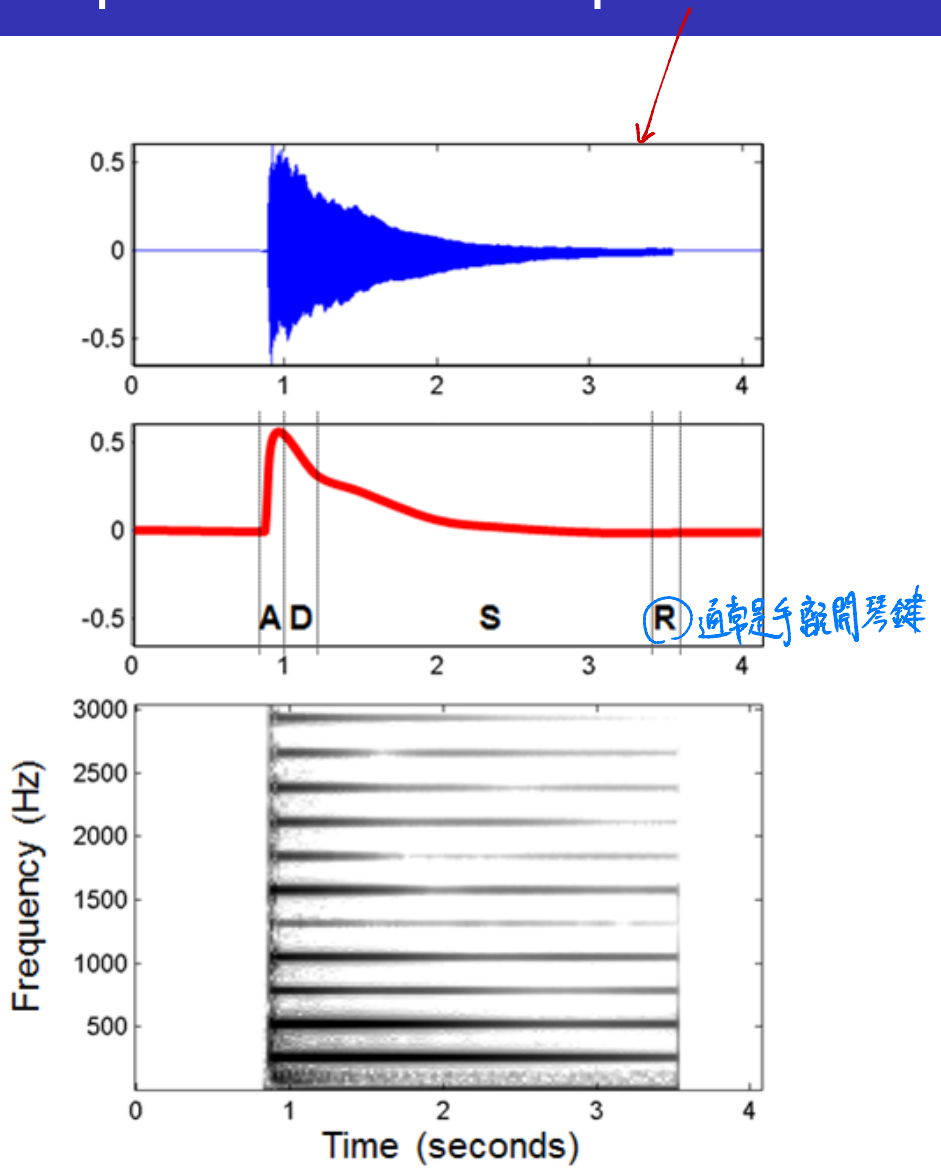
- "Rise time": no strick definition
- One definition: the time interval between the point the audio signal reaches 20% and 80% of its maximum value
- Log attack time (LAT): 20% 能量 → 80% 能量的時間，適合用來分辨樂器

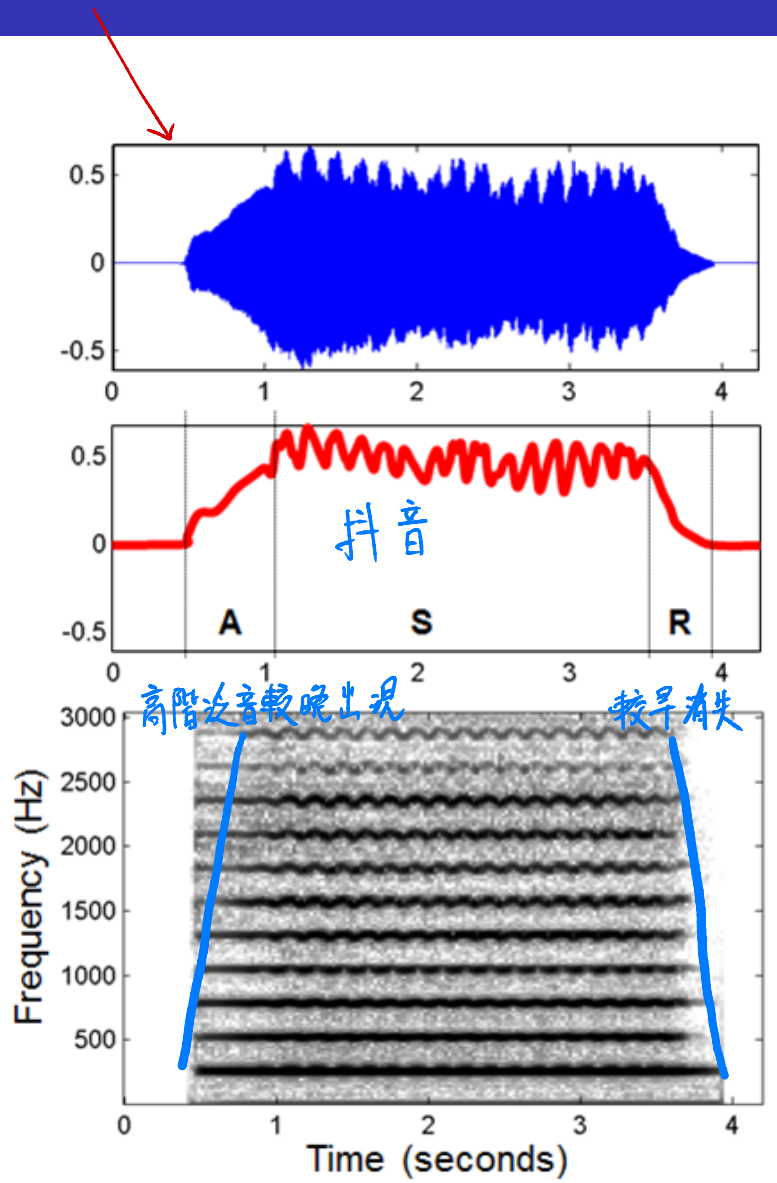
$$\mathrm{LAT} = \log_{10}(t_{80} - t_{20}) \tag{3}$$

- Temporal centroid of a note:

$$C_t = \frac{\sum_\Omega nE(n)}{\sum_\Omega E(n)}, \quad \Omega := \{n : \text{onset time} < n < \text{offset time}\} \tag{4}$$

Temporal features: piano and violin





From: M. Mueller, *Fundamentals of Music Processing*, Chapter 1, Springer 2015

Vibrato and tremolo

- *Tremolo*: periodic variations in amplitude (i.e., amplitude modulation), in some cases called *shimmer*
- *Vibrato*: periodic variations in frequency (i.e., frequency modulation), in some cases called *jitter*

通常有 Vibrato 都會跟著有 Tremolo

Log-scale spectrum

- Sampling rate f_s, window size N, hop size H

$$X(n,k) = \sum_{m=0}^{N-1} x(m+nH)h(m)e^{-\frac{j2\pi km}{N}} \tag{5}$$

$$\mathcal{X}(n,k) = |X(n,k)|^2 \tag{6}$$

- The index k corresponds to the frequency $f(k) := \frac{kf_s}{N}$ 在 spectrogram 上的刻度
- The index n corresponds to the time $t(n) := \frac{nH}{f_s}$
- Human perception of loudness is of log-scale: $\log \mathcal{X}(n,k)$
- Human perception of pitch is also of log-scale: define for each pitch p

參考12平均律

$$P(p) := \{k : F_{pitch}(p-0.5) \leq k < F_{pitch}(p+0.5)\} \tag{7}$$

音的上下半個半音的範圍當作那個 pitch 　Ex: A4 440Hz → $440\times 2^{\frac{1}{24}} \sim 421.47$, → $440\times 2^{\frac{1}{24}} \sim 452.89$

- The log-frequency spectrogram: $\mathcal{Y}(n,p) := \sum_{k\in P(p)} \mathcal{X}(n,k)$

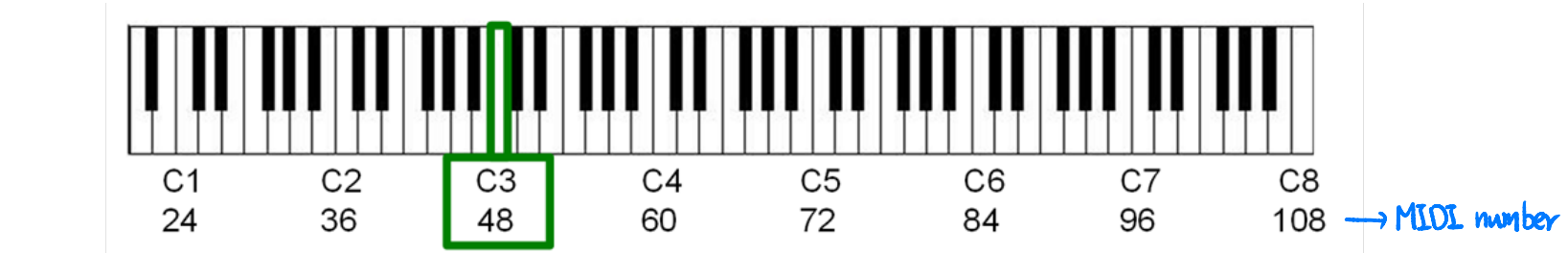
Pitch name, MIDI, and frequency

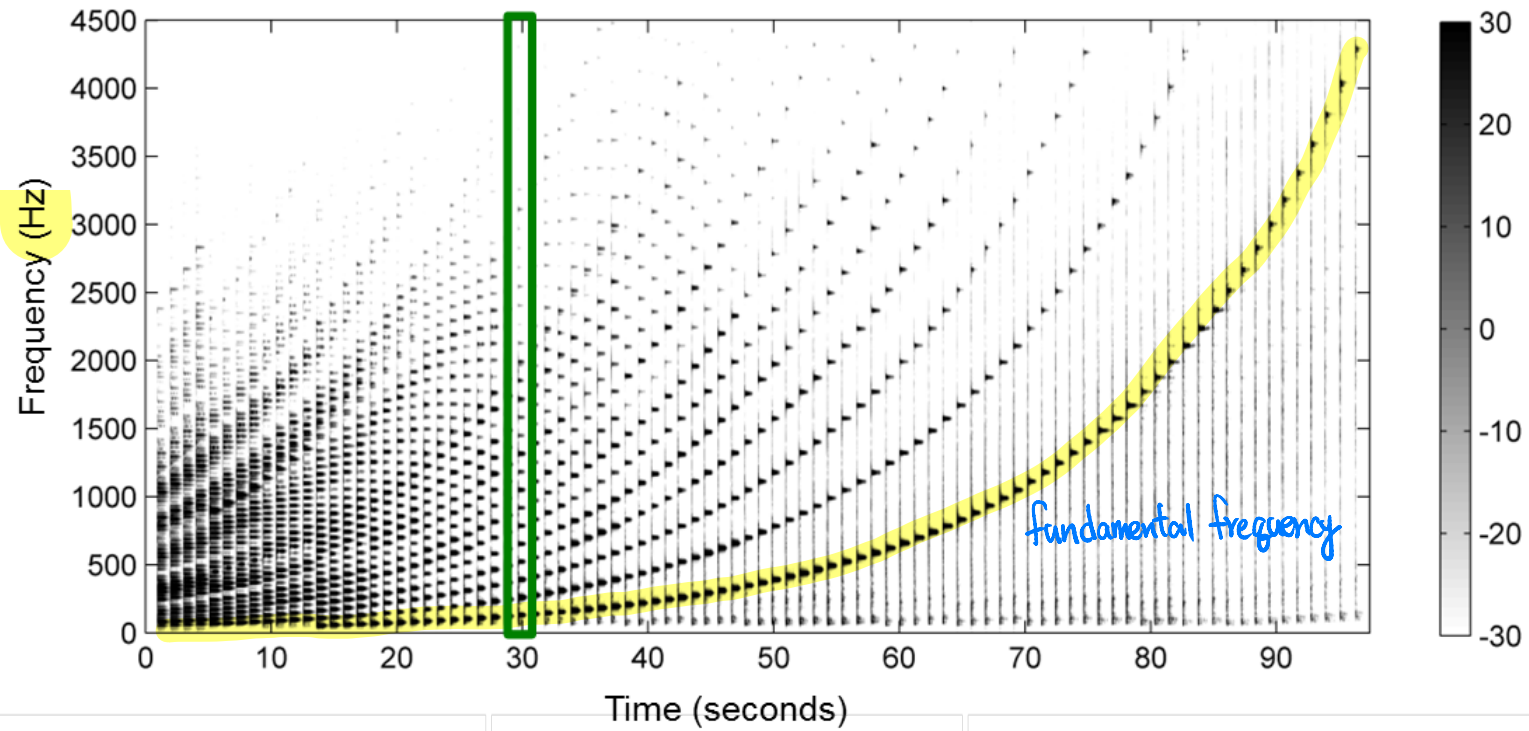
Band Width 頻寬

Note	p	$F_{\text{pitch}}(p)$	$F_{\text{pitch}}(p-0.5)$	$F_{\text{pitch}}(p+0.5)$	$\text{BW}(p)$
C4	60	261.63	254.18	269.29	15.11
C♯4	61	277.18	269.29	285.30	16.01
D4	62	293.66	285.30	302.27	16.97
D♯4	63	311.13	302.27	320.24	17.97
E4	64	329.63	320.24	339.29	19.04
F4	65	349.23	339.29	359.46	20.18
F♯4	66	369.99	359.46	380.84	21.37
G4	67	392.00	380.84	403.48	22.65
G♯4	68	415.30	403.48	427.47	23.99
A4	69	440.00	427.47	452.89	25.41
A♯4	70	466.16	452.89	479.82	26.93
B4	71	493.88	479.82	508.36	28.53
C5	72	523.25	508.36	538.58	30.23

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 3, Springer 2015

The chromatic scale of piano: $\log \mathcal{X}$

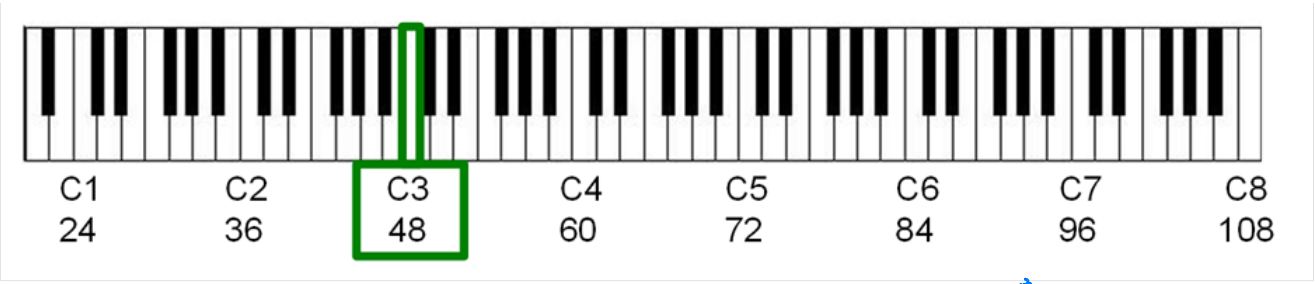


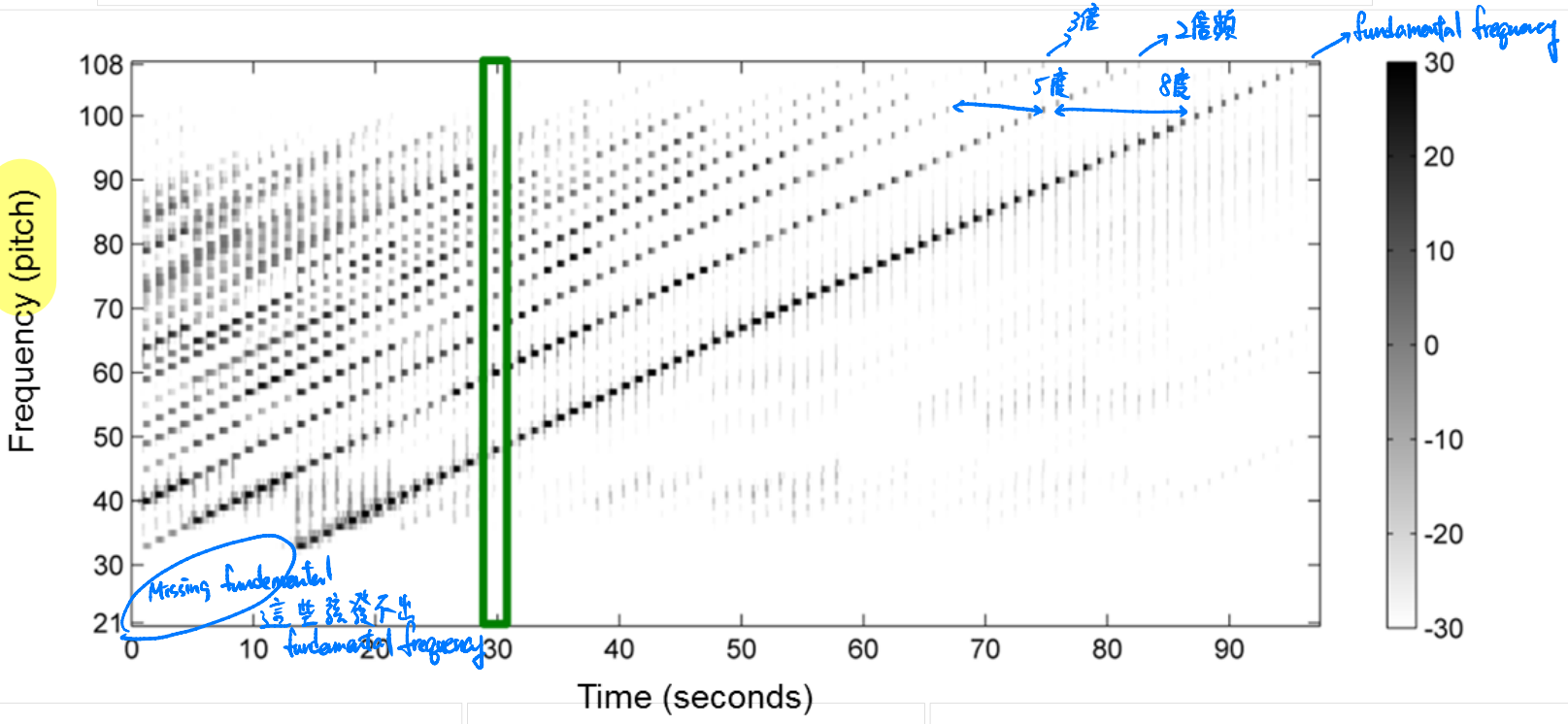


From: M. Mueller, *Fundamentals of Music Processing*, Chapter 3, Springer 2015

The chromatic scale of piano: $\log \mathcal{Y}$

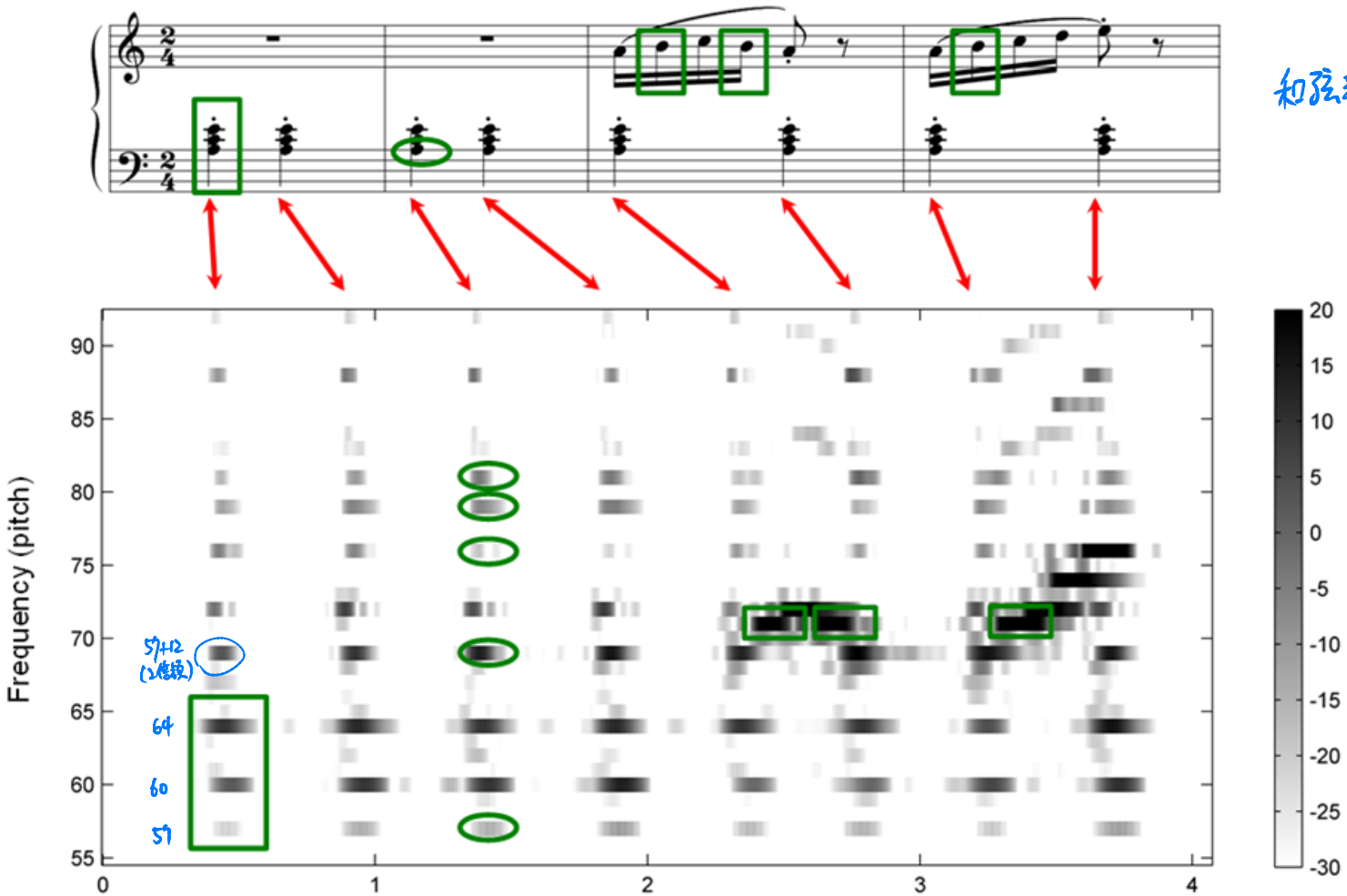
半音階





From: M. Mueller, *Fundamentals of Music Processing*, Chapter 3, Springer 2015

More examples



From: M. Mueller, *Fundamentals of Music Processing*, Chapter 3, Springer 2015

Mel-scale spectrogram

- Mel scale simulates human's perception of pitch

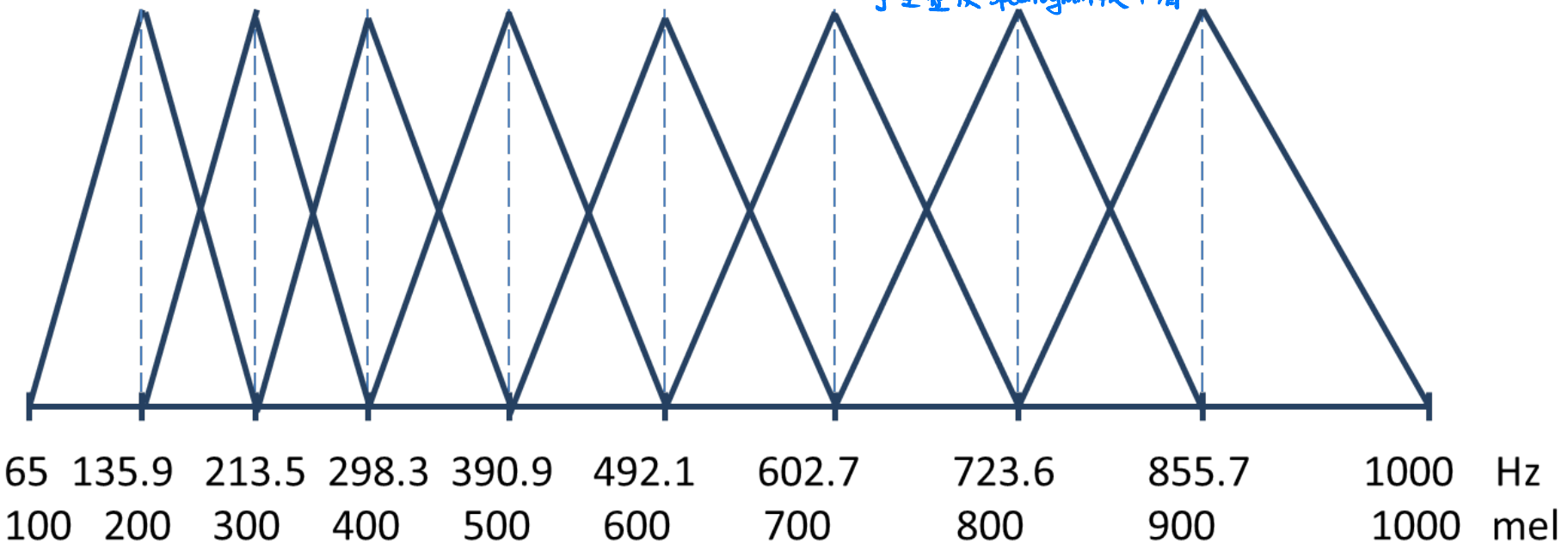
針對語音優化(非音樂)

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right) \tag{8}$$

$f < 700$ 時為線性

- Example: 8 mel-scale, triangular filter banks for 65 – 1000 Hz

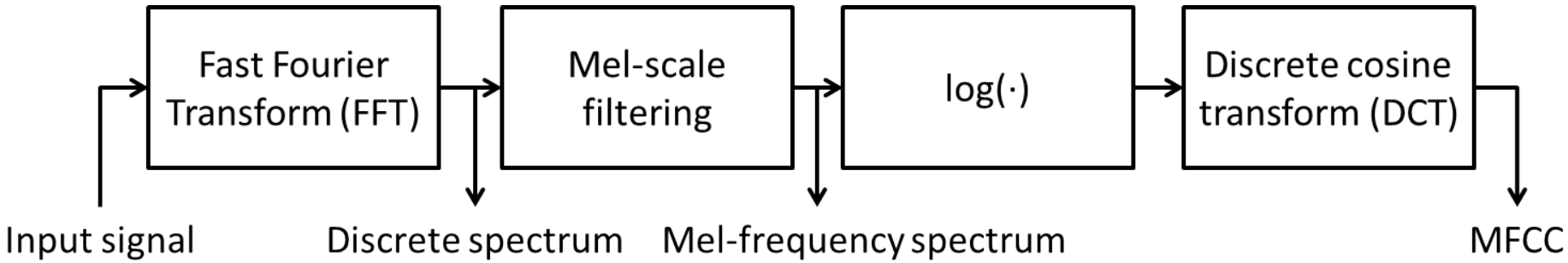
可重疊使 spectrogram 較平滑



MFCC

↳ DNN之前語音辨識和MIR的 state-of-the-art

- Cepstrum: the inverse FFT of the log-magnitude spectrum
- Mel-frequency cepstral coefficients (MFCC): a cepstral feature derived from mel-frequency spectrum
- Common usage: 13-, 20-, or 40-term MFCC
- 1st and 2nd temporal differences of MFCC are also important feature
- Building blocks:



Window size, pitch and bandwidth

時間的解析度 v.s. 頻率的解析度

- Recall *spectral leakage*: every spectral peak (of a sinusoidal component) has finite width
- Recall the *chromatic scale*: low pitches distribute denser than high pitches 低頻醜：因為低頻 Bandwidth 較小
- Recall *Heisenberg uncertainty principle*: longer window gives sharper peaks, and vice versa in terms of 波長 not #samples
- Q: What do we mean by "long"? A: In terms of the *wavelength* (frequency) of the signal!
- Main idea: using long window for low-frequency parts, while using short window for high-frequency part

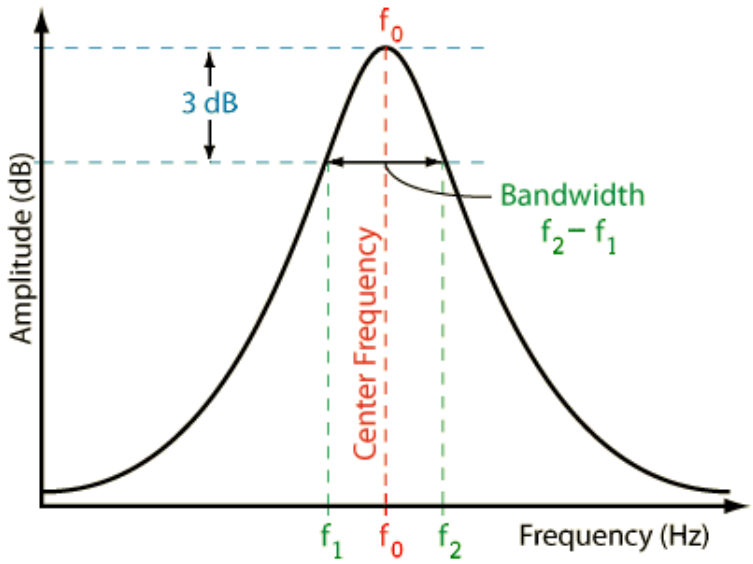
Q factor

- 3-dB bandwidth: $f_1 - f_2$
- Q-factor: 頻譜散開的幅度，Q愈大愈好，Q愈小愈糟

#signal in FFT

$$Q = \frac{f_0}{f_1 - f_2} = \frac{Nf_0}{\Delta f_s} \quad \text{最高點 / 3dB頻寬} \tag{9}$$

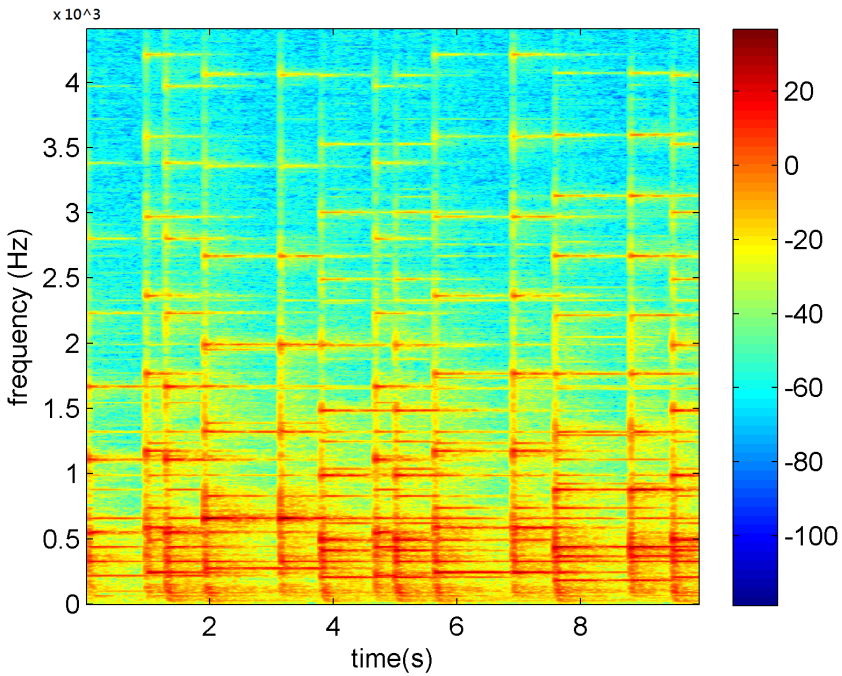
- Δ: the 3-dB bandwidth of the window function
- For Hann window, $\Delta \approx 1.50$ DFT bins
- To achieve constant Q factor, we have $N \propto \frac{1}{f_0}$

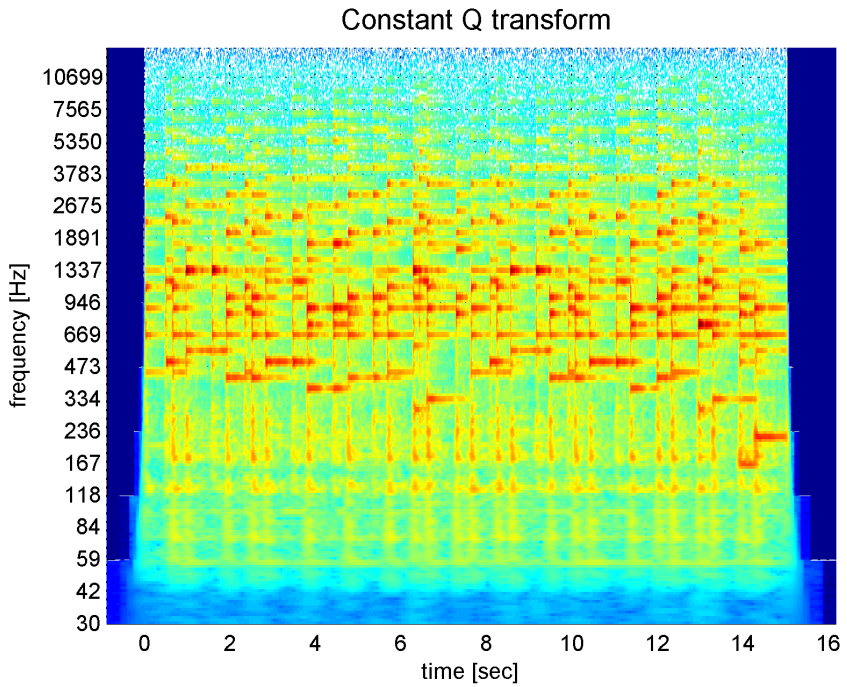


Constant-Q transform (CQT)

- 24 bins per octave ($Q = 22.75$)
- Reference: C. Schorkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing", in Proceedings of the 7th Sound and Music Computing Conference, Barcelona, Spain, 2010.

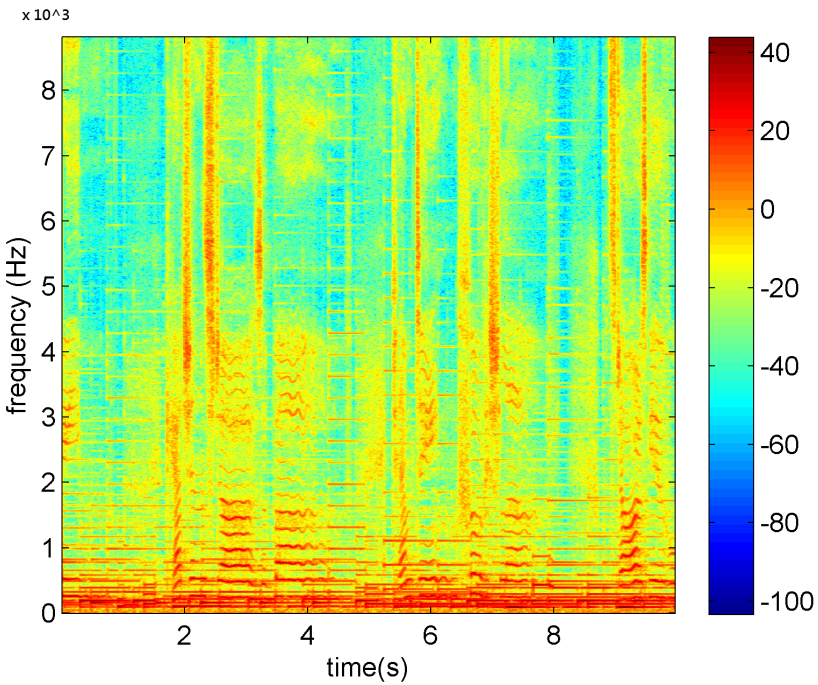
▸ Play sound

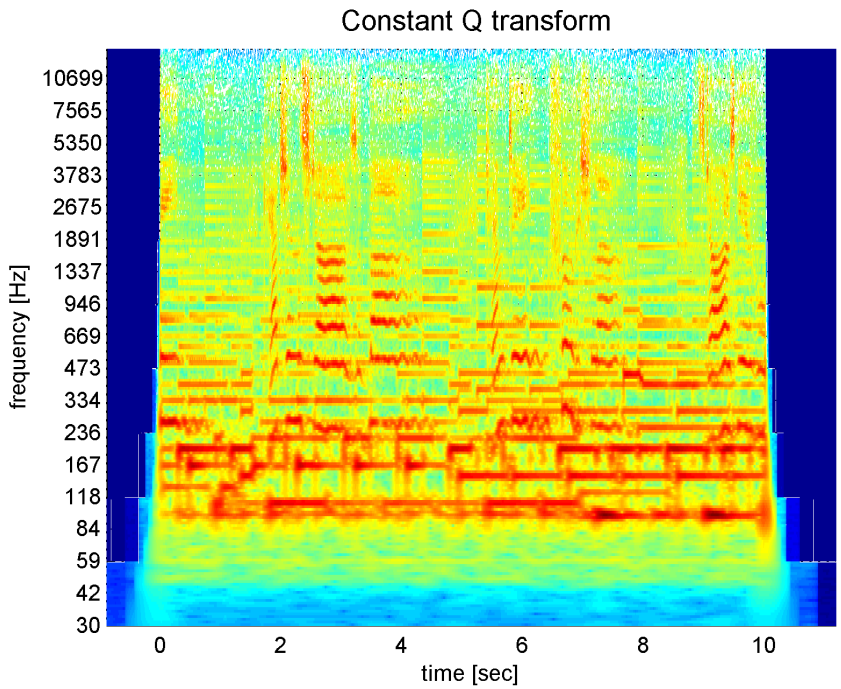




Constant-Q transform

▸ Play sound





Timbre features: high-frequency energy

Spectral *roll-off*

- The frequency such that a certain fraction of the total energy is contained below that frequency
- The minimal R such that $\sum_{k=1}^{R} X[n,k] \geq \gamma \sum_{k=1}^{N} X[n,k]$, where $\gamma = 0.95$ or 0.85 的訊號能量

Spectral *brightness*

- The amount of energy above a *cut-off* frequency f_c (e.g., 1500 Hz) compared to all spectral energy 某個頻率以上的能量

Timbre features: derived from $X[n, k]$

Spectral *roughness*

- Sensory dissonance related to the beating phenomenon whenever pair of sinusoids are closed in frequency (Plomp and Levelt, 1965)
- A model: for two pure tones with frequencies f_1 and f_2, with amplitudes a_1 and a_2, the roughness d is

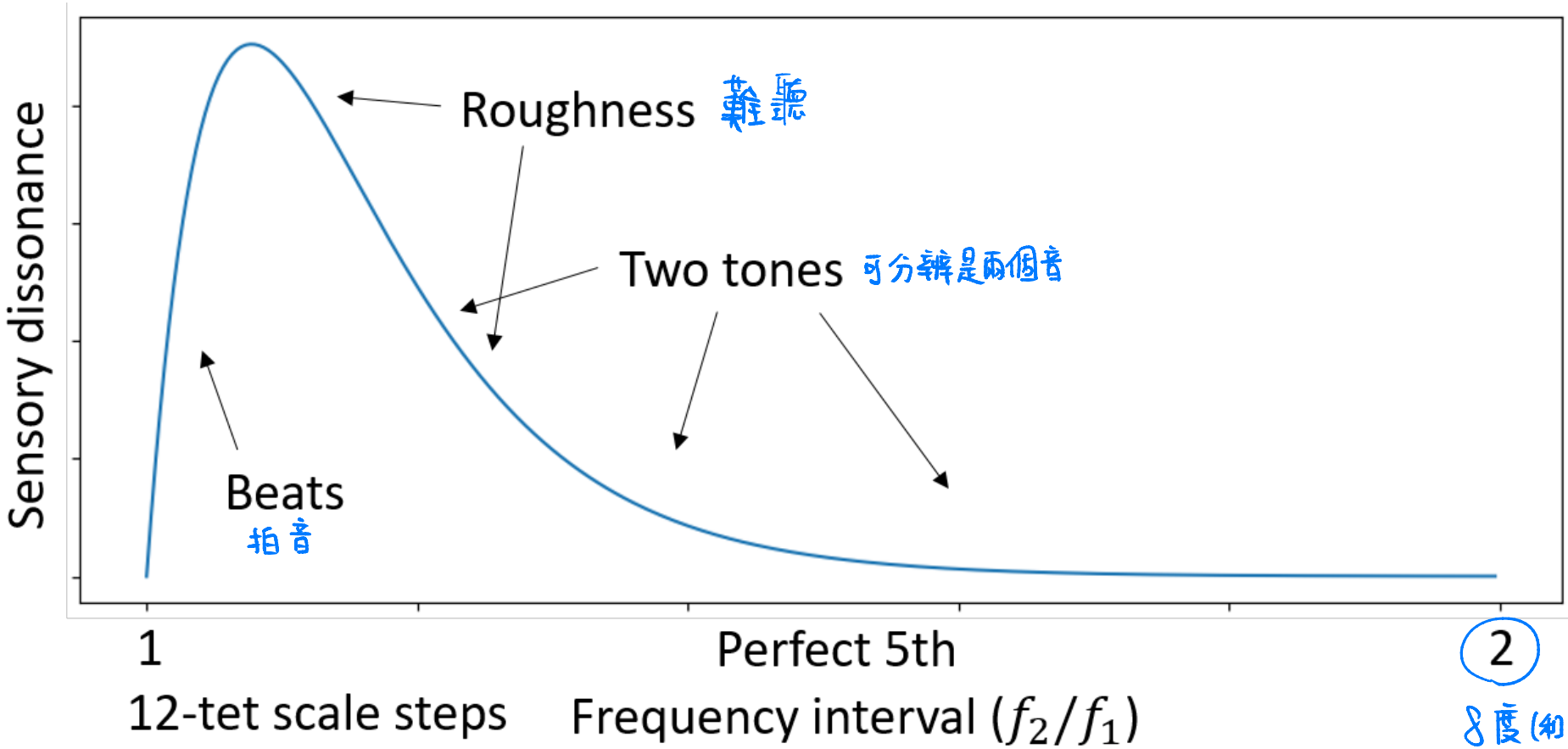
$$d(f_1, f_2, a_1, a_2) := a_1 a_2 (e^{-ax} - e^{-bx}), \tag{10}$$

$$x = \frac{0.24|f_1 - f_2|}{0.021 \min(f_1, f_2) + 19}, \tag{11}$$

$$a = 3.5, \ \ b = 5.75 \tag{12}$$

- Total roughness: the summation of roughness of all pairs of spectral peaks in a spectrum

Timbre features: roughness



Timbre features

Spectral irregularity

- The degree of variation of the successive peaks of the spectrum

$$SI(n) = \frac{\sum_{k=1}^{N-1}(\mathcal{X}(n,k+1) - \mathcal{X}(n,k))^2}{\sum_{k=1}^{N}\mathcal{X}(n,k)} \tag{13}$$

Timbre features: derived from $\mathcal{X}(n,k)$

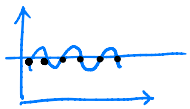
- Spectral centroid: $\mu_1 = \sum_{k=1}^{N} k\mathcal{X}(n,k) / \sum_{k=1}^{N} \mathcal{X}(n,k)$
- Spectral spread: $\mu_2 = \sum_{k=1}^{N} (k-\mu_1)^2 \mathcal{X}(n,k) / \sum_{k=1}^{N} \mathcal{X}(n,k)$
- Spectral skewness: $\mu_3 = \sum_{k=1}^{N} (k-\mu_1)^3 \mathcal{X}(n,k) / \sum_{k=1}^{N} \mathcal{X}(n,k)$
- Spectral kurtosis: $\mu_4 = \sum_{k=1}^{N} (k-\mu_1)^4 \mathcal{X}(n,k) / \sum_{k=1}^{N} \mathcal{X}(n,k)$
- Entropy: $H(n) = -\sum_{k=1}^{N} \left(k \log \mathcal{X}(n,k)\right) / \log N$
- Spectral flatness: the ratio between the geometric mean and the arithmetic mean

$$SFM(n) = 10 \log_{10} \left(\frac{\left(\prod_{k=1}^{N} \mathcal{X}(n,k)\right)^{1/N}}{\frac{1}{N} \sum_{k=1}^{N} \mathcal{X}(n,k)} \right) \tag{14}$$

Zero-crossing rate (ZCR)

- Zero-crossing rate (ZCR): 一個訊號在單位時間內經過幾次0

常用來分辨子音母音，亦可算音高 但不可有noise



$$ZCR(n) = \frac{1}{2N}\sum_{i=1}^{N}|\text{sign}(x(n+i)) - \text{sign}(x(n+i-1))| \tag{15}$$

$$\text{sign}(x) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0. \end{cases} \tag{16}$$

- ZCR of noise is usually larger than ZCR of the sound with identifiable pitches
- ZCR is the fastest pitch detector – why? any drawbacks?
- More reference: Roger Jang's website ▸ Website

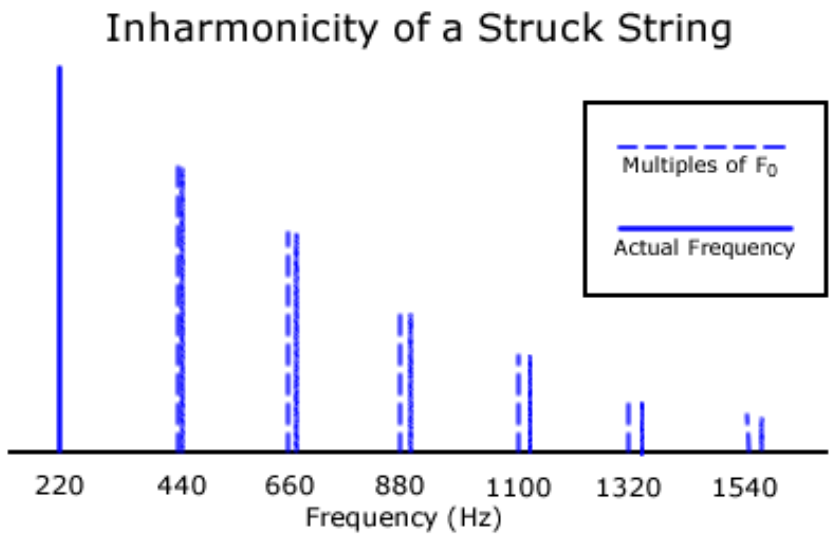
Prof. 張智星

Inharmonicity

一些樂器泛音可能不是 fundamental frequency 的整數倍

- For some pitched instrument, the overtones of the sound are not exactly the multiples of the fundamental frequency
- Example: piano, guitar, harp, chimes, glockenspiel 自由振動常有 inharmonicity (free oscillation)
- Free oscillation vs. forced oscillation!
- Non-ideal string: Hooke's Law is no longer valid because of finite mass, cross-section area, gyration, etc. of the string

$$f_n = nf_0\sqrt{1+\beta n^2} \quad \text{n倍倍頻} \tag{17}$$



Harmonic features

- Denote the amplitude of the i-th harmonic peak as $a(i)$, $i = 0, 1, \cdots$, and Q harmonic peaks (including fundamental)
- Inharmonicity

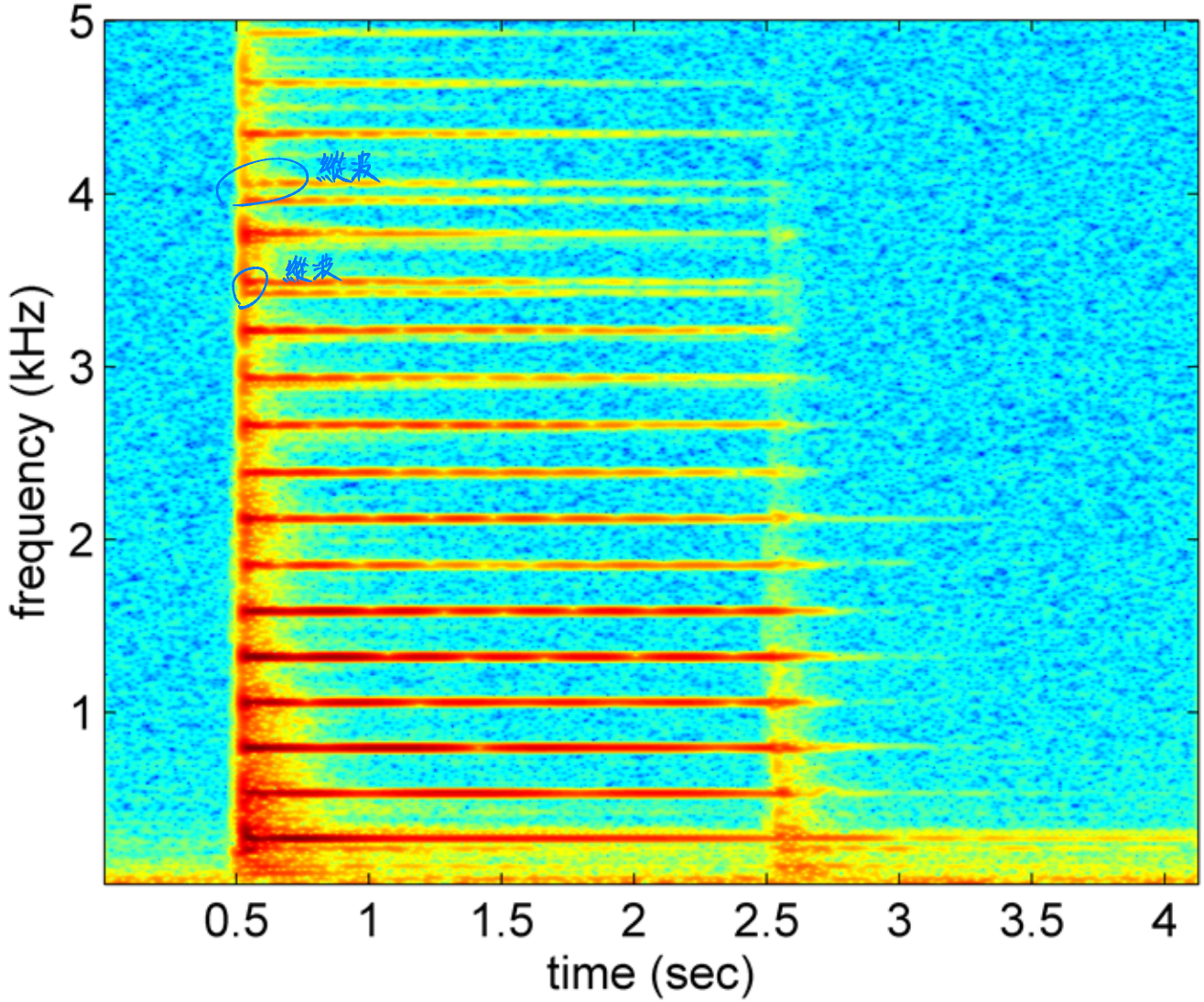
$$IH = \frac{2}{f_0} \times \frac{\sum_{i=1}^{Q} |f_i - if_0| \times a^2(i)}{\sum_{i=1}^{Q} a^2(i)} \tag{18}$$

- Odd-to-even ratio

$$OER = \frac{\sum_{q\,\text{odd}} a^2(i)}{\sum_{q\,\text{even}} a^2(i)} \tag{19}$$

Example: piano

What do you see?



The application of timbre feature

- Classification
- Similarity estimation
- Sound quality assessment
- Transcription
- Voice conversion
- Sound synthesis
- And more...

Useful tools for music processing and feature extraction

- librosa ▸ Website
- MIRtoolbox ▸ Website
- Essentia ▸ Website
- CQT toolbox ▸ Website
- And others...