Lecture 5: Pitch detection

Li Su

2019/04/09

What is pitch?

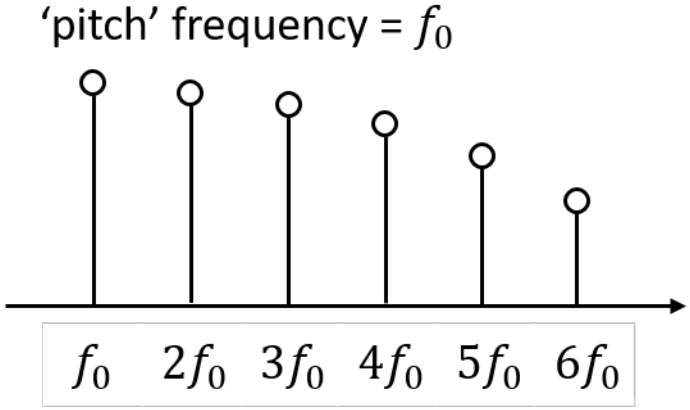
- “That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high” (ANSI, 1994)
- The perception of pitch forms the basis of musical melody and harmony

What is pitch?

大部份的PAPER這樣講，方便討論

- Physical dimension: fundamental frequency (F0) → Always true? Any limitation?
- Perceptual dimension: missing fundamentals, masking effects, pitch shifts, virtual pitch, dichotic pitch, and the pitches of things that are not there at all ...
- In MIR, 'pitch' and 'F0' are usually used interchangeably
- How easy/hard finding the F0 of a signal from its spectrum?

↓ 去看看這些



Pitch detection in terms of frequency resolution

很多的bins可以記錄細緻的抖音和滑音

- Semitone-level
  - In terms of semitone; the natural way representing note
  - Insufficient for expressive performance (F0-to-note)
  - Classification problem
- Bin-level
  - The direct outcome from signal processing
  - Classification problem
- Instantaneous F0 (real number output)
  - Time derivative of phase or interpolation
  - Catch the fundamental definition of F0 but unstable

Pitch contour and instantaneous F0

- Frequency as a function of time!
- Limitation of DFT: spectral leakage and finite bin space
- A general model of instantaneous frequency: phase?

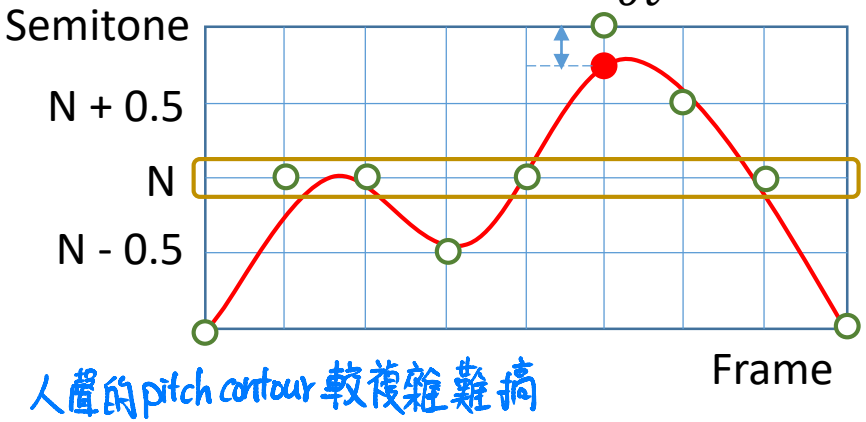
相角

Signal phase

$$x(t) = A(t)\cos\big(2\pi\phi(t)\big),\ \omega = \frac{\partial\phi(t)}{\partial t}$$

STFT phase

$$\Delta\omega = \frac{\partial\Phi(\omega,t)}{\partial t}$$

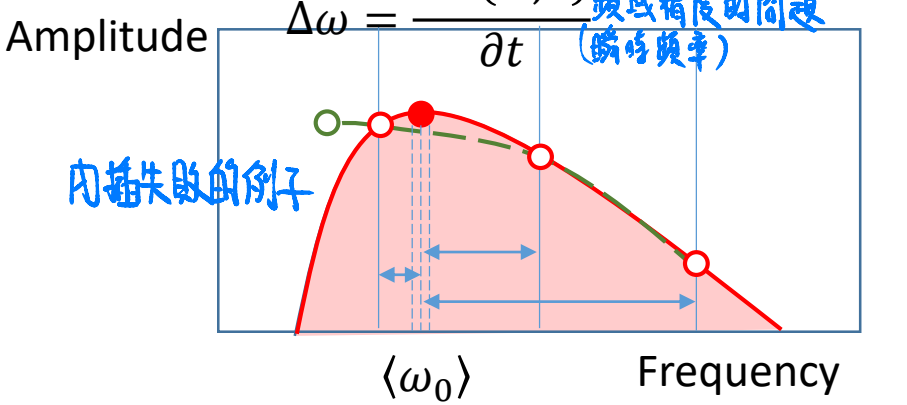


人聲的pitch contour較複雜難搞

Instantaneous F0 deviation

$$\Delta\omega = \frac{\partial\Phi(\omega,t)}{\partial t}$$

可以每3點做一次拋物線內插來簡單處理頻域精度的問題（瞬時頻率）



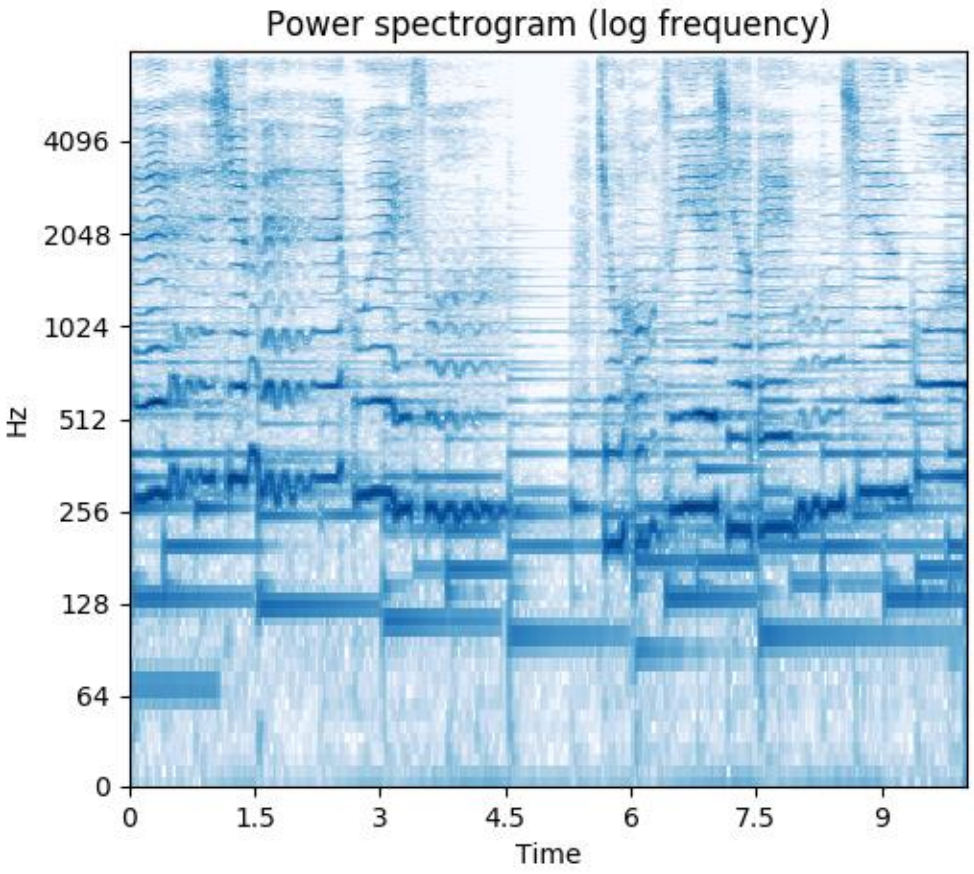
Pitch detection in terms of time scale

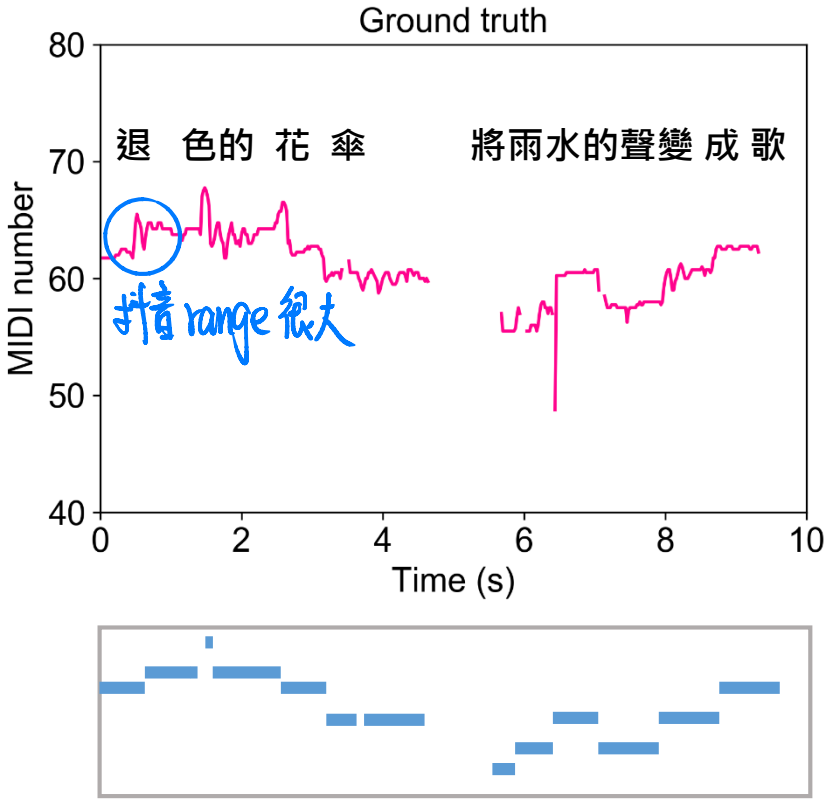
Context info.

- Frame-level 每個時間點的音高
  - The time-pitch profile
  - Example: piano roll representation with time resolution of 10 ms
- Note-level 音何時開始何時結束
  - The onset time, offset time, and pitch of each note
- Global (phrase or a full music piece)
  - Also know as “pitch streaming” 聲部的音高，Ex: 主旋律、人聲、鋼琴、Bass
  - The onset time, offset time, pitch and voice of each note
  - Example: key, mode, chord function, ...

The semantic gap in pitch detection

- Example: singing voice





Mono-pitch and multi-pitch detection

同時多音高

- Music is usually ``polyphonic'': there are usually more than one pitches at every moment of a music piece
- Note: in MIR, the definition of ``polyphony'' is somewhat different from the one in music theory
- Challenge 1: high number of polyphony -- can your ear make it?
- Challenge 2: overlapped harmonics -- consider a note combination (C4-C5-G5) -- how to identify the harmonic peaks of each note?

和諧音程難 separation

- Challenge 3: diverse timbre information -- is there enough training capacity?

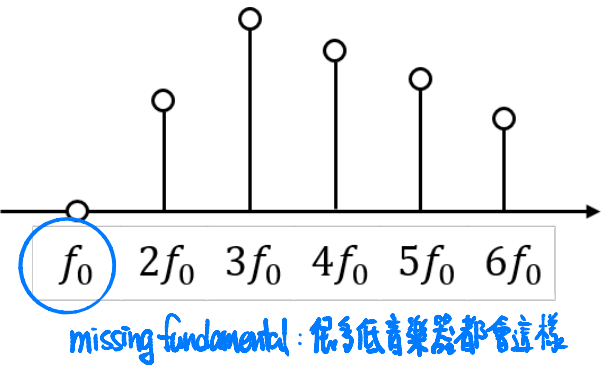
現在仍是難題

- Multi-class vs. multi-label classification problems

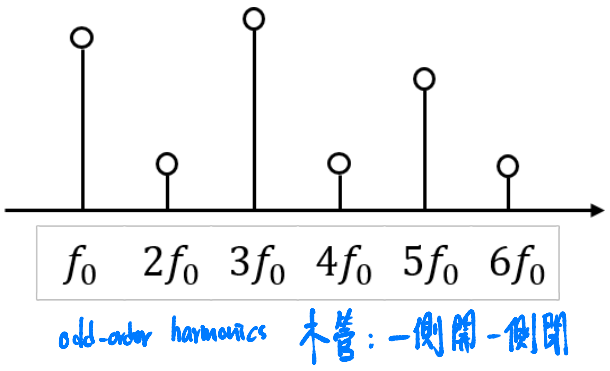
Some issue from spectral analysis

- Phenomenon 1: missing fundamental
  - Low-pitch parts of piano, low-pitch instruments, male voices, ...
- Phenomenon 2: odd-order harmonics
  - Clarinet and some woodwind instruments
- Phenomenon 3: inharmonicity $f_n = nf_0\sqrt{1+\beta n^2}$ 倍頻有偏移
  - Piano, guitar, and other stuck-string or pluck-string instruments ...
- Solution: period detection?

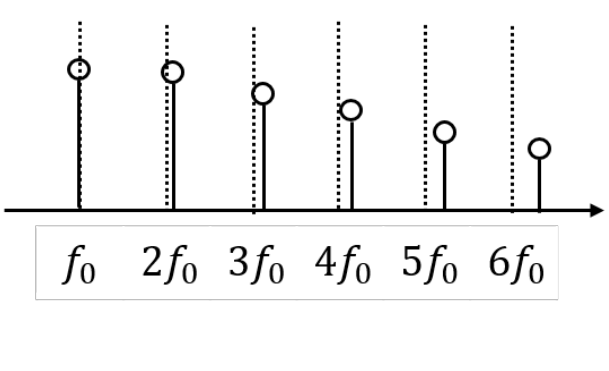
找週期



missing-fundamental：很多低音樂器都會這樣



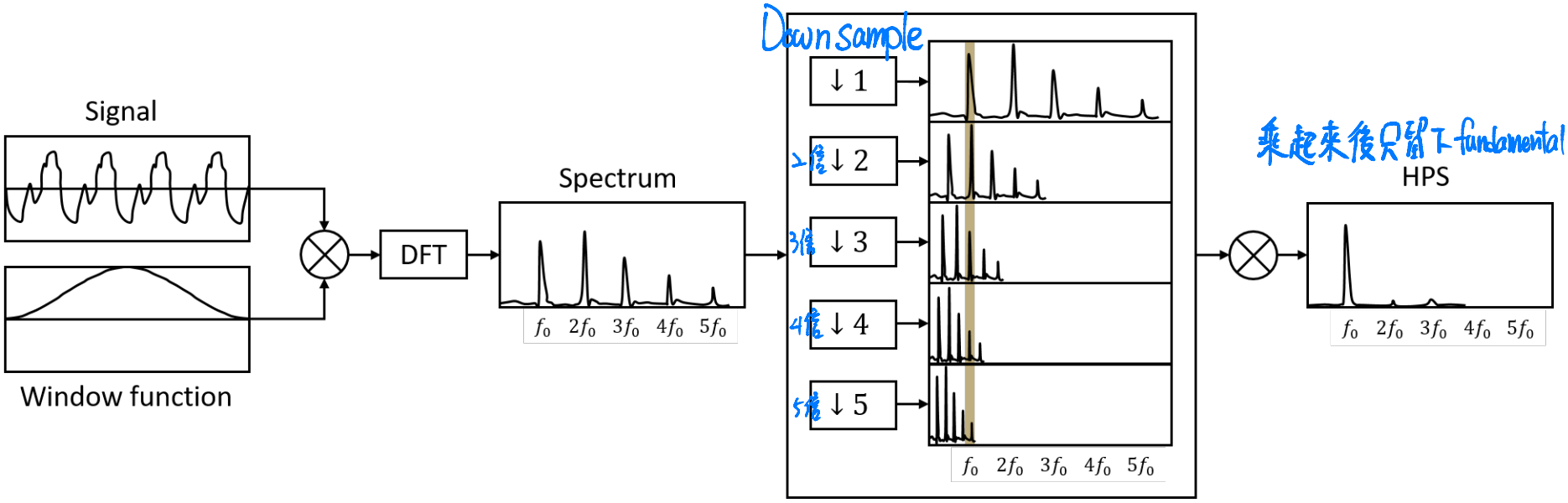
odd-order harmonics 木管：一側開一側閉



Harmonic product spectrum (HPS) [Noll, 1969]

- A nonlinear spectrum; address the weak fundamental issue
- Given the Fourier spectrum $X[k]$ of the input signal $x[n]$, the HPS is the geometric mean of amplitudes of the harmonics in $X[k]$

$$\mathrm{HPS}_x[k] = \left(\prod_{m=1}^{M} X[mk] \right)^{\frac{1}{M}}$$



“Periodicity” detection

音高=频率？Maybe=周期

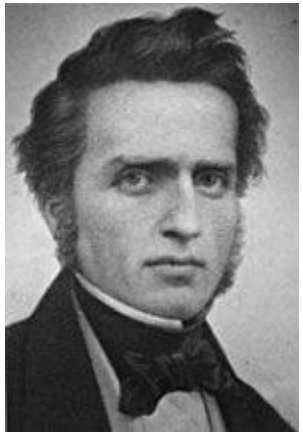
- We have discussed some techniques in spectrum estimation / frequency detection
- What is the difference between frequency and periodicity?
- Formally, a periodic signal is defined as

$$x(t) = x(t + T_0), \qquad \forall\, t$$

- What is the difference in the definition of frequency and periodicity?
- Find the fundamental frequency/period
- Application: pitch detection, transcription, beat tracking …

Scientists on Psychoacoustic

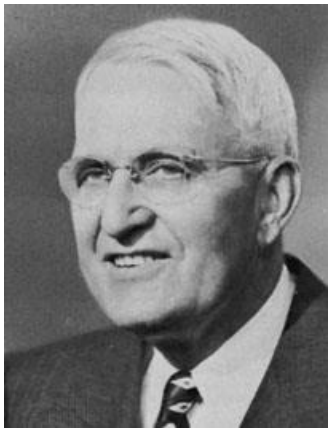
Pitch detection theory: a historical remark



August Seebeck (1805-1849)

Georg Simon Ohm (1789-1854)

Herman von Helmholtz (1821-1894)



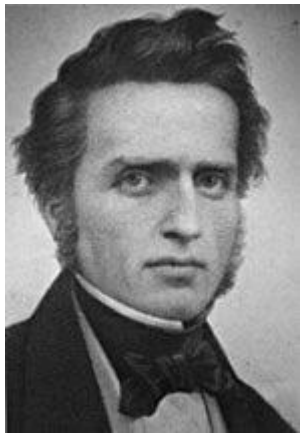
Harvey Fletcher (1884-1981)

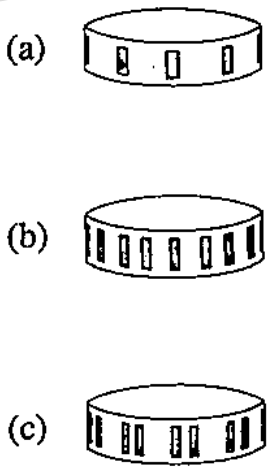
Jan Frederik Schouten (1910-1980)

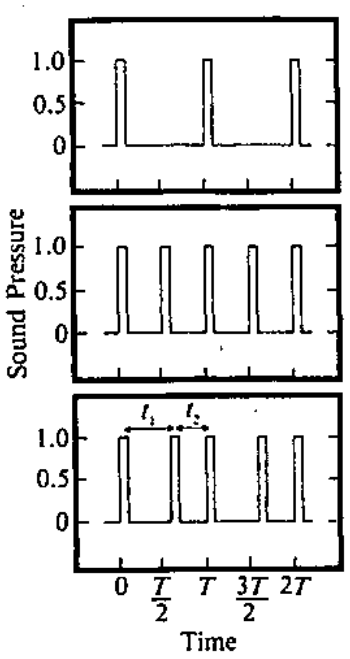
Seebeck's experiment (1841) and Ohm's second law

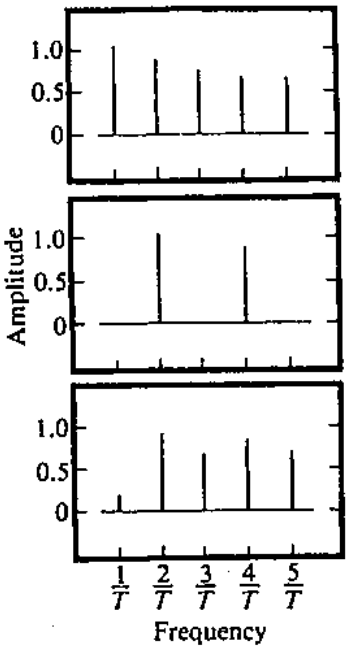
- Ohm's second law: a pitch could be heard only if the wave contains power at the frequency ("Fourierism" perspective)
- Ohm: Seebeck's finding is just an illusion

pitch is periodicity!





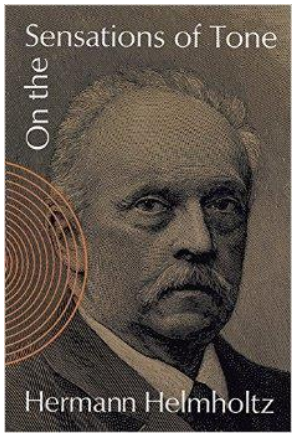




pitch is frequency!

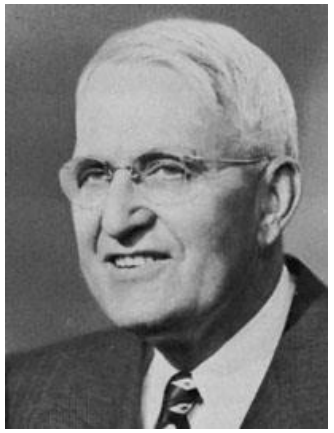
Helmholtz's theory

- 《On the Sensations of Tone as a Physiological Basis for the Theory of Music》(1877)
- "Fourierism" perspective: distortion products generated in the ear so we can hear that weak fundamental
- Fletcher: discover "missing fundamental" using high-pass filter on audio signal



I support Ohm's position, and I have a beautiful explanation

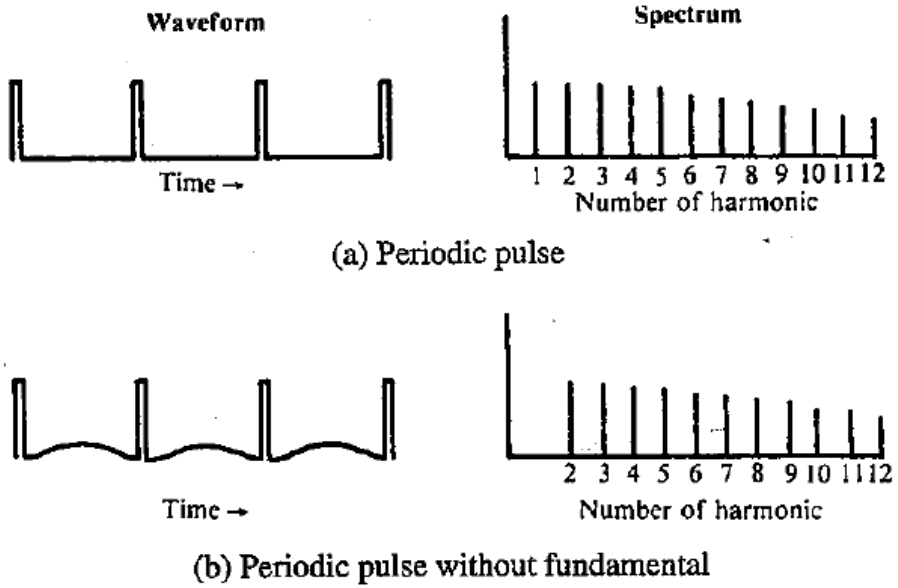
I support Helmholtz's position!



Schouten's experiment I (1938)

- Input signal: 400Hz, 600Hz, 800Hz, …, with distortion product at 200Hz (Helmholtz's theory)
- Add a pure tone of 206 Hz, beats should be heard
  - No beats were heard

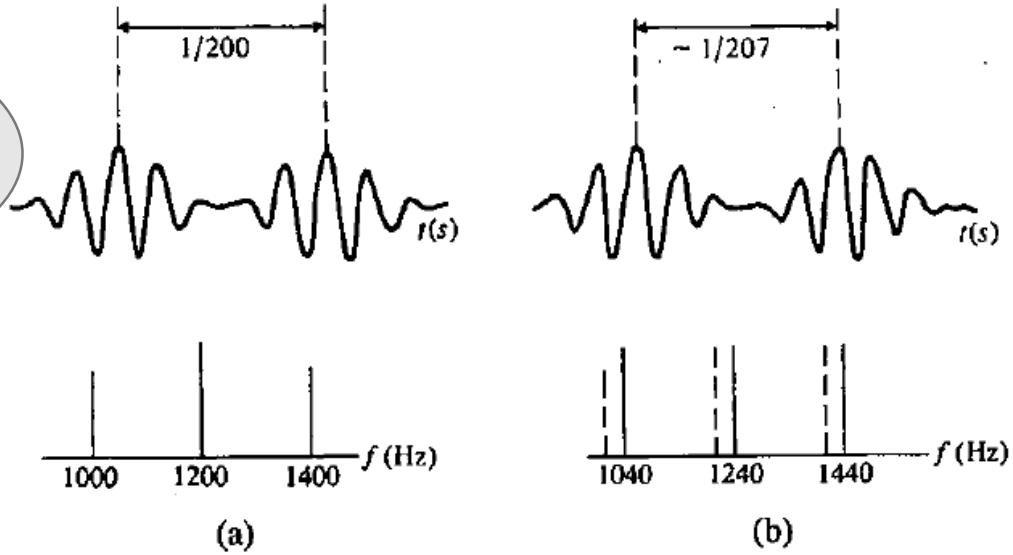
Things are not that simple…



Schouten's experiment II (1938)

- Input signal: 1000Hz, 1200Hz, 1400Hz
  - A clear pitch at 200 Hz should be heard (Helmholtz's theory)
- Input signal: 1040Hz, 1240Hz, 1440Hz
  - Also a clear pitch at 200 Hz should be heard (Helmholtz's theory)
- Experiment: ~207 Hz

Things are not that simple…



Basic idea of periodicity detection

- Formally, a periodic signal is defined as

$$x(t) = x(t + T_0), \qquad \forall\, t$$

- Formally, the frequency spectrum of a signal is defined as…
- Frequency analysis: the relationship between the signal and the sinusoidal basis
- Periodicity analysis: the relationship between the signal and itself

Basic periodicity detection functions

- Autocorrelation function (ACF)
- Average magnitude difference function (AMDF)
- YIN and its periodicity detector
- Generalized ACF and Cepstrum

Autocorrelation function (ACF)

算訊號週期

- Measures similarity across time
- Cross correlation:

$$R_{xy}(\tau) = \frac{1}{N-1} \sum_{t=0}^{N-1-\tau} x(t)y(t+\tau)$$

- Autocorrelation:

和自己shift過的訊號內積

$$R_{xx}(\tau) = \frac{1}{N-1} \sum_{t=0}^{N-1-\tau} x(t)x(t+\tau)$$

- t: time-domain
- τ: lag-domain

Other relevant pitch detection functions

- Average magnitude difference function (AMDF)

$$AMDF_{xx}(\tau) = \frac{1}{N-1} \sum_{t=0}^{N-1-\tau} |x(t) - x(t+\tau)|$$

- The pitch detection function used in YIN

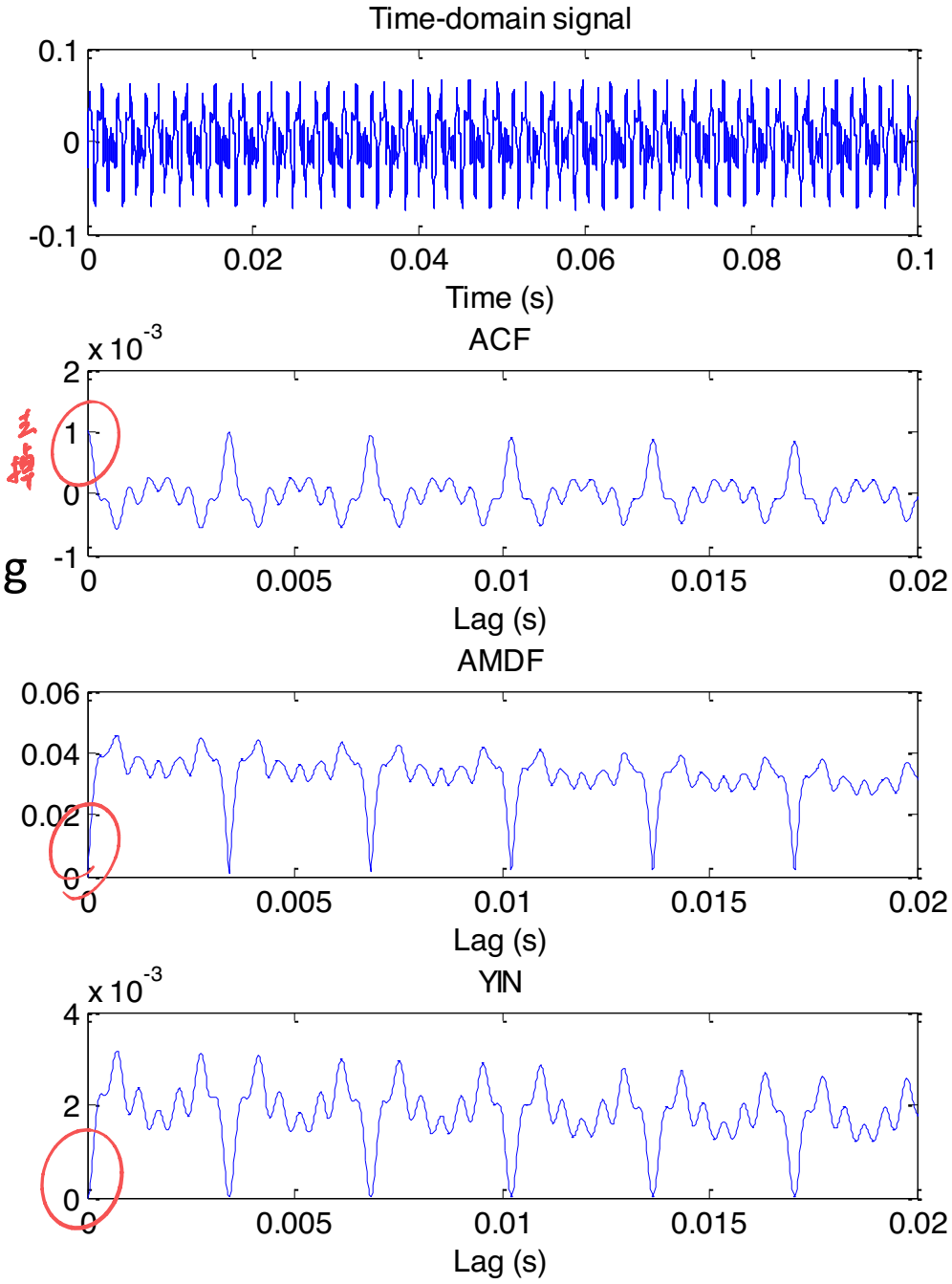
$$YIN_{xx}(\tau) = \frac{1}{N-1} \sum_{t=0}^{N-1-\tau} \left(x(t) - x(t+\tau)\right)^2$$

Ref: Alain de Cheveigné et al, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. 111 (4), April 2002

Result

- A violin D4
- $f_0 = 293$ Hz
- $T = 3.41$ msec
- Pitch indicator: discarding zero-lag term (for zero lag the signal matches the signal itself)

- $p^* = \text{argmax}_p \; ACF(p)$
- $p^* = \text{argmin}_p \; AMDF(p)$



Wiener-Khinchin Theorem

- The computational complexity of a N-point ACF:
  - $O(N \times N)$
  - Is there any way to accelerate it?

- **Wiener-Khinchin theorem**: the ACF is the inverse Fourier transform of the power spectrum
  - Complexity: $O(N \log N)$

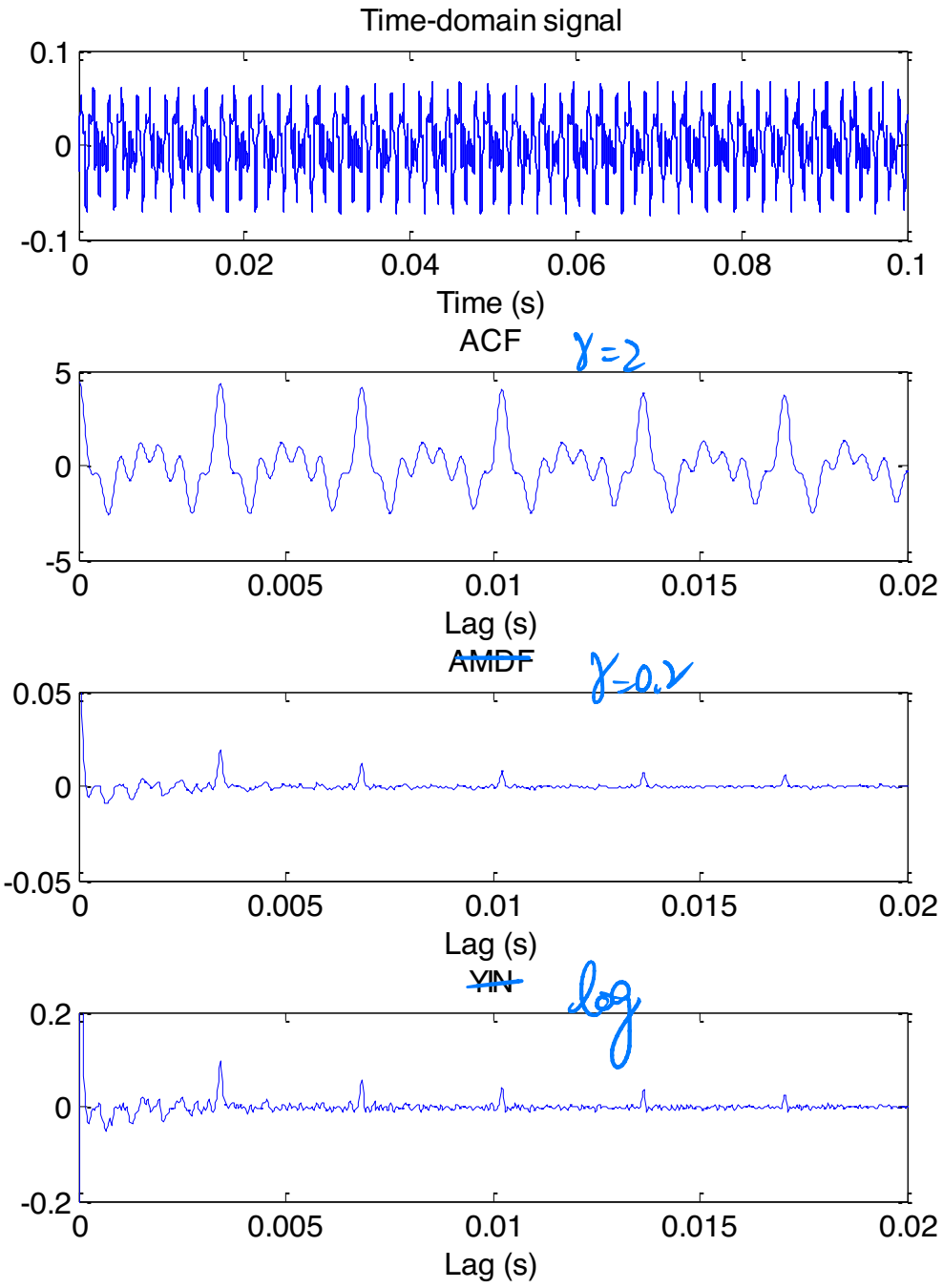
ACF = $R_{xx}(\tau) = IFFT(\left|FFT\left(x(t)\right)\right|^2)$

Generalized ACF

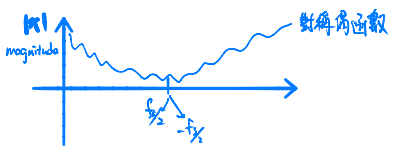
- Consider a generalization of ACF:
  - $R_{xx}(\tau) = IFFT(\left|FFT\left(x(t)\right)\right|^{\gamma}), 0 < \gamma < 2$
  - Or, $R_{xx}(\tau) = IFFT(\log\left|FFT\left(x(t)\right)\right|)$?
- What are the advantages of generalized ACF?
  - Recall the "logarithmic compression" part of the chromagram!
- Reference: data normalization
  - Helge Indefrey, Wolfgang Hess, and Günter Seeser. "Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain-preliminary results." *in Proc, ICASSP*, 1985.
  - Anssi Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model." *IEEE Transaction on Audio, Speech and Language Processing,* Vol.16, No.2, pp. 255-266, 2008.

Result

- A violin D4 ($f_0 = 293$ Hz, $T = 3.41$ msec)
- Pitch indicator:
  - $\gamma = 2$ (ACF)
  - $\gamma = 0.2$
  - Logarithm



Cepstrum



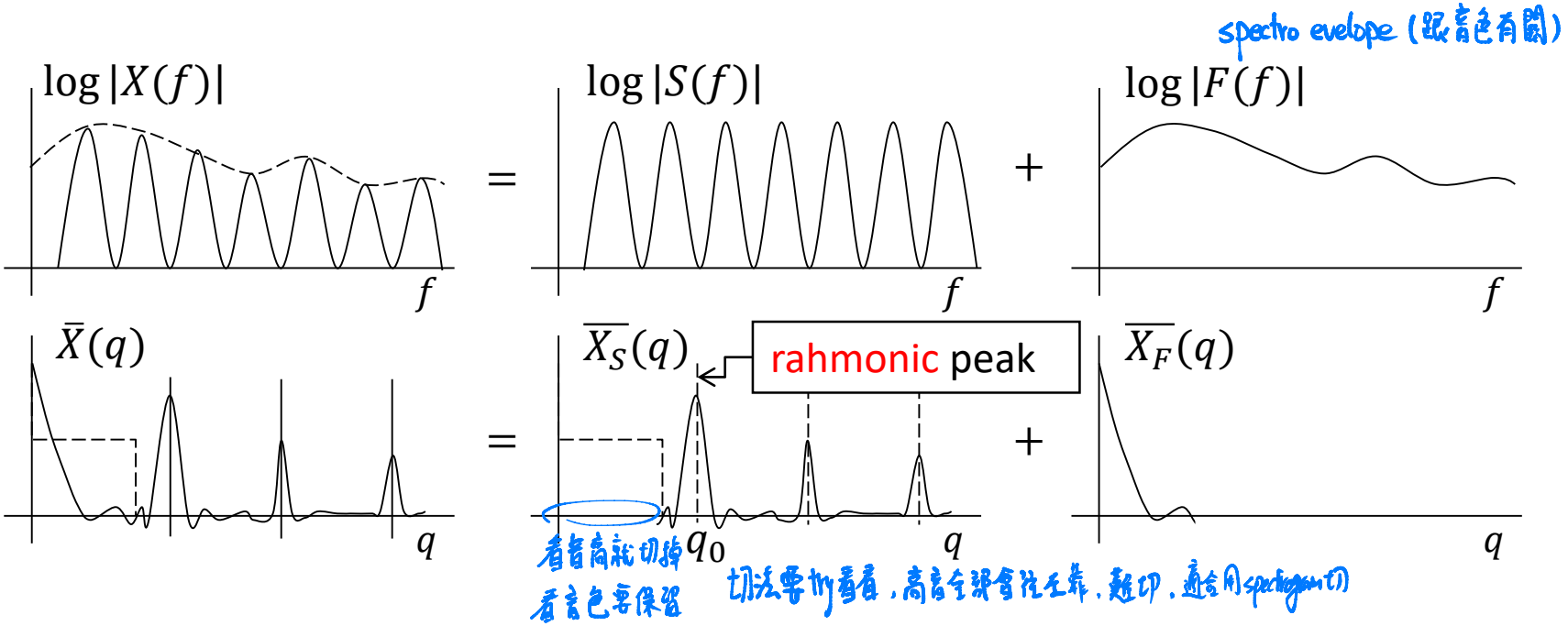
- From spectrum to cepstrum (倒頻譜)
- Spectrum computed by fast Fourier transform (FFT): $X(f) = FFT(x(t))$
- Cepstrum: $\bar{X}(q) = IFFT(\log|X(f)|)$
  - q: quefrency (倒頻率) (not frequency)
  - Quefrency in the cepstrum, and lag in the ACF are both measured in time (but not in the time domain)

$$x(t) \xrightarrow{FT[\cdot]} X(f) \xrightarrow{\log[\cdot]} \log|X(f)| \xrightarrow{FT^{-1}[\cdot]} \bar{X}(f)$$

$s(t) * f(t)$	$S(f)F(f)$	$\log\|S(f)\| + \log\|F(f)\|$	$\log\|\bar{S}(f)\| + \log\|\bar{F}(f)\|$
convolution	product	addition	addition

The meaning of the cepstrum

- What is the meaning for “the spectrum of a spectrum”?
  - It extracts the “oscillatory behaviors” of the spectrum
  - It measures “how many oscillatory shapes per frequency” -> fundamental period!
  - We can also think ACF in this way!



Cepstrum terminology

- Oppenheim, Alan V., and Ronald W. Schafer. "From frequency to quefrency: A history of the cepstrum." *IEEE signal processing Magazine* 21.5 (2004): 95-106.

- (Note: some terms are actually seldom used now)

Frequency domain	Cepstrum domain
Frequency	Quefrency
Spectrum	Cepstrum
Harmonic	Rahmonics
Filtering	Liftering
Low-pass filter	Short-pass lifter
High-pass filter	Long-pass lifter

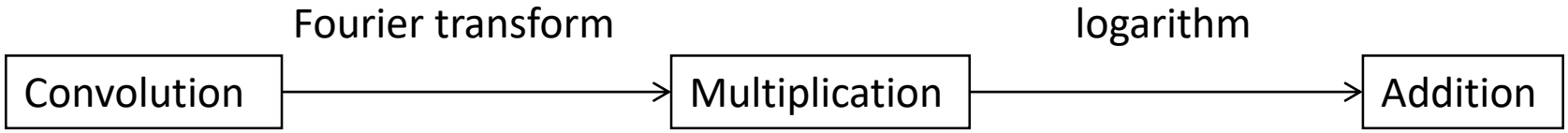
A closer look to the cepstrum

- A physical system: convolution of the excitation and the impulse response
  - Excitation: “fast” spectral variation
  - Impulse response: “slow” spectral variation
- How to do “deconvolution”?
  - Homomorphic signal processing
  - Homomorphism: to “carry over” operations from one algebra system to another
  - Convert complicated operation to simple ones
- Example:

$$y(t) = x(t) * h(t) \quad (* \text{: convolution})$$
$$Y(t) = X(t) \times H(t)$$
$$\log Y(t) = \log X(t) + \log H(t)$$
$$FFT(\log Y(t)) = FFT(\log X(t)) + FFT(\log H(t))$$

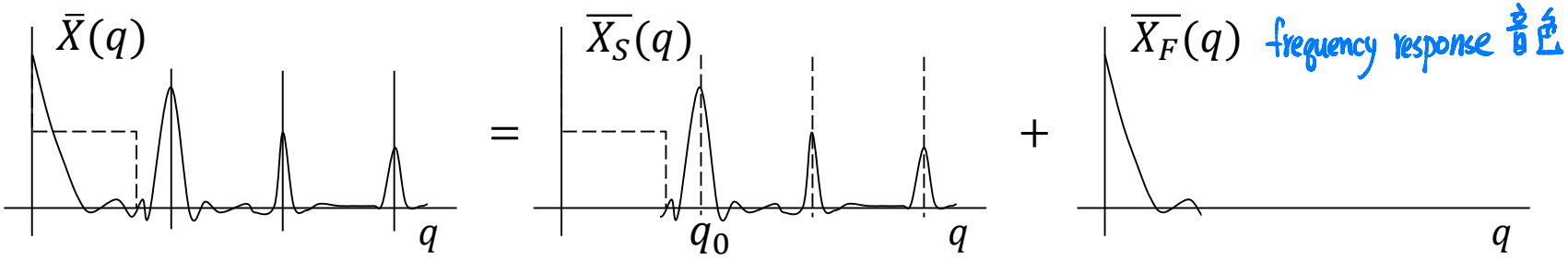
乘法到加法

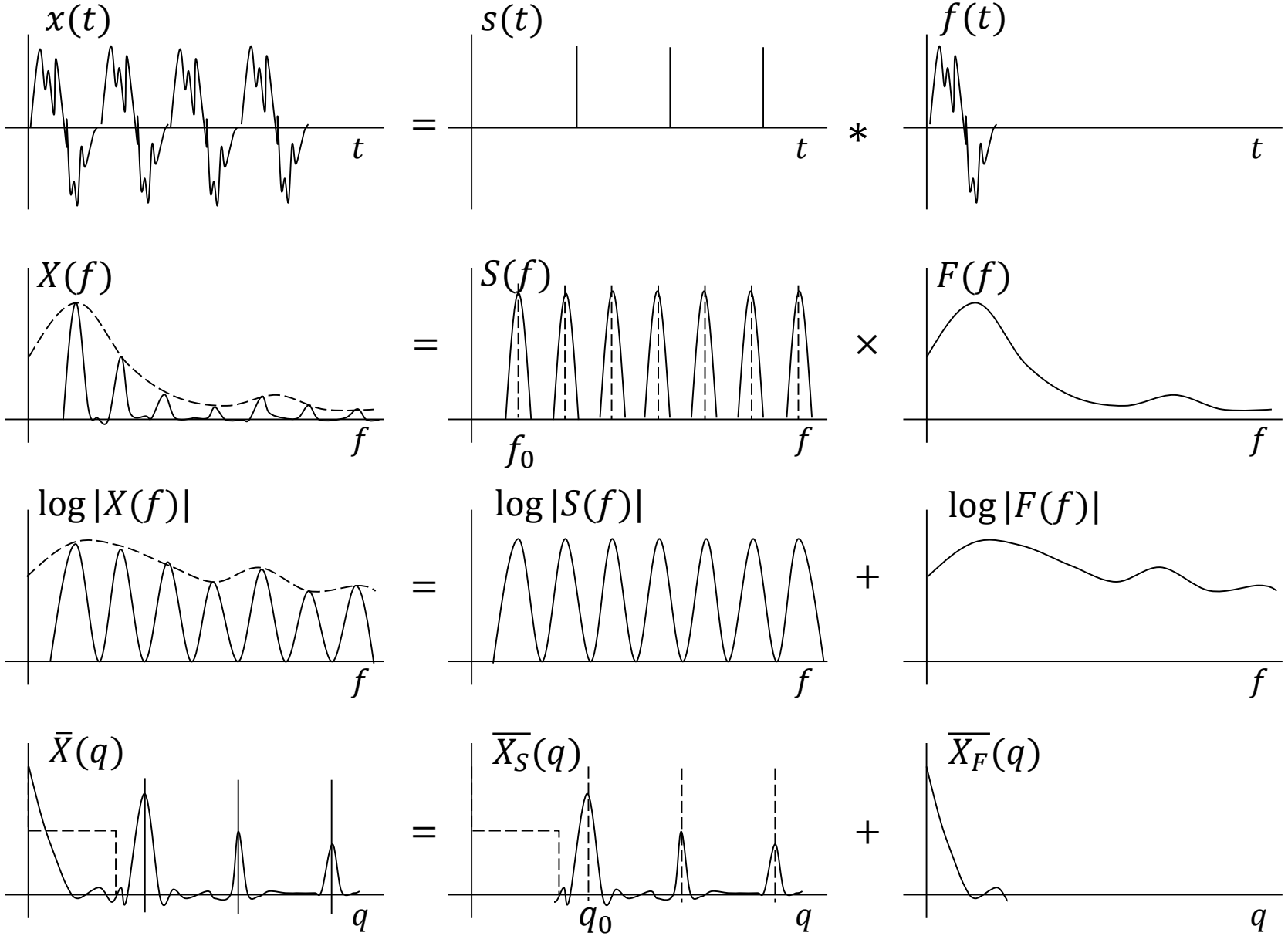
可用来抓 impulse response



Homomorphic signal processing for pitch detection

- Source-filter model: pitch signal as an impulse train convolved by an impulse response
- Separation of “oscillatory” part and the “impulse response” part
- Example of homomorphic filtering: a long-pass lifter for capturing pitch information
  - High-pass vs. long-pass
  - Filter vs. lifter




x(t)
t
=
s(t)
t
*
f(t)
t
X(f)
f
=
S(f)
f0
f
×
F(f)
f
log |X(f)|
f
=
log |S(f)|
f
+
log |F(f)|
f
X̄(q)
q
=
X̄S(q)
q0
q
+
X̄F(q)
q

Generalized logarithm and cepstrum

- If we just care about the effect of "nonlinear scaling" (i.e., logarithm) when computing cepstrum
  - Pros: simulate human's perception by compression
  - Cons: sensitive to noise and zeros in the spectrum
- Generalized logarithm:

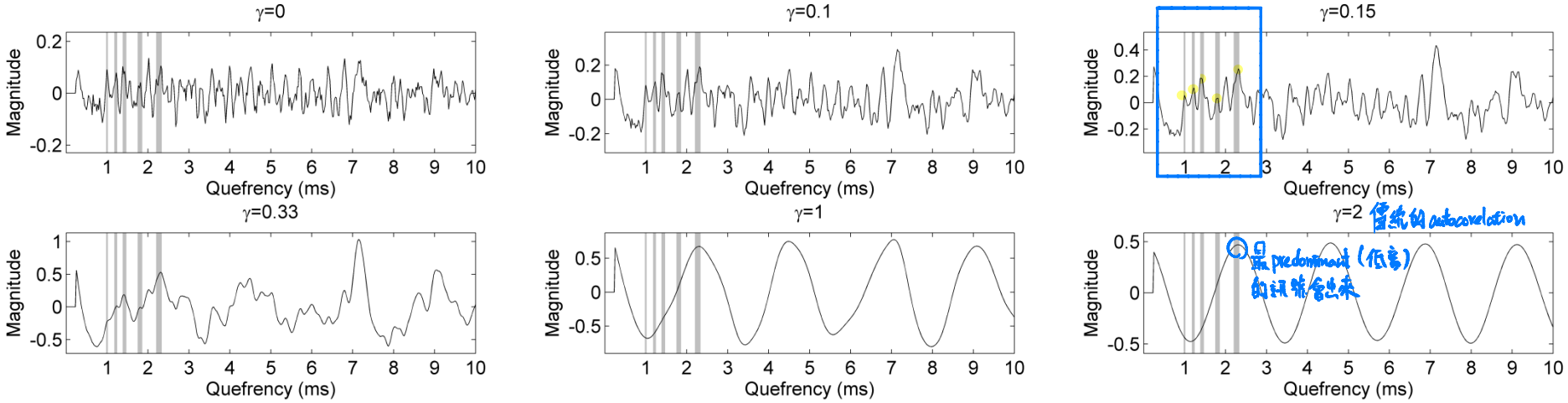
$$g_\gamma(x) = \begin{cases} \dfrac{|x|^\gamma - 1}{\gamma} & ,0 < \gamma < 2 \\ \ln x & ,\gamma = 0 \end{cases}$$

$\gamma = 0.3, 0.5$ (常用)

- Generalized cepstrum: $\bar{X}_\gamma(q) = IFFT(g_\gamma(X(f)))$
  - Similar to the generalized ACF: $R_{xx}(\tau) = IFFT(|X(f)|^2)$
  - Useful when there are multiple pitches
  - Implication: our perception may be neither linear scale (ACF) nor log scale (cepstrum)

Example

- A complicated example: 5-polyphony piano sample A4+C#5+F5+G#5+B5 multi pitch estimation



- T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.
- L. Su and Y.-H. Yang, "Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music", *IEEE/ACM Speech Audio Language Process.*, vol. 23, no. 10, pp. 1600—1612, Oct. 2015.

Dictionary-based pitch detection: basic

近期常用的技巧

- From frequency representation
- A "dictionary" $\mathbf{D} \in R^{m \times n}$ be a set of spectral features (from data or predefined)
- $\mathbf{D} = [d_1, d_2, \cdots, d_n]$, column $d_k \in R^m$ called an "atom" or "template"
- Input feature vector: $\mathbf{x} \in R^m$
- Encoding process: template matching
- Solve linear equations / linear approximation, $\boldsymbol{\alpha} \in R^m$

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$$

or

$$\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$$

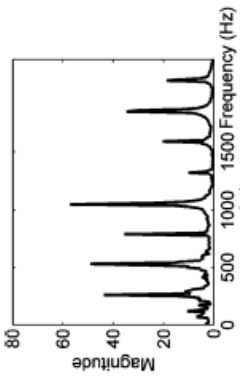
Basic template matching: single pitch detection

已知樂器時都算非難

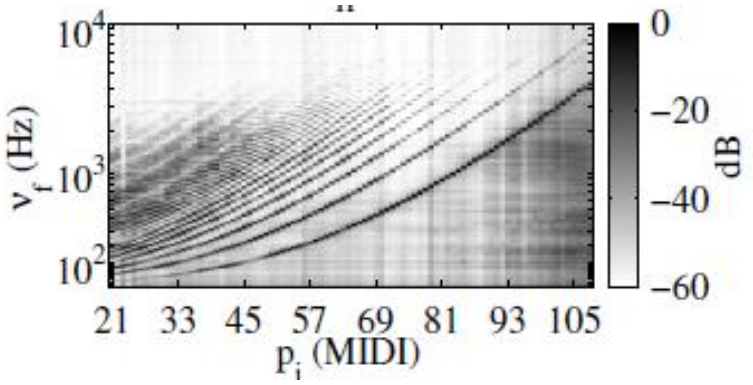
- Input $\mathbf{x}$, dictionary $\mathbf{D} = [d_1, d_2, \cdots, d_{88}]$, each d_k represents one pitch (e.g., d_1 is the spectral pattern of A0, d_{40} is the spectral pattern of C4)
- Find a d_k such that $\mathbf{x} \cdot d_k$ is maximum
- Vector quantization (VQ): "sparsest" approximation

$$\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha} \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_0 = 1$$

非0的個數 (vector quantization)



C4



Templates from A0 to C8

Deep learning-based pitch detection

- Kim, Jong Wook, et al. "CREPE: A convolutional representation for pitch estimation." IEEE ICASSP 2018