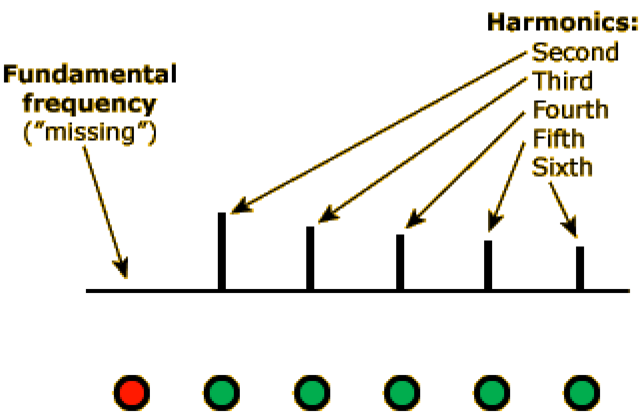
Lecture 8: Pitch and Chord (3) pitch detection and music transcription

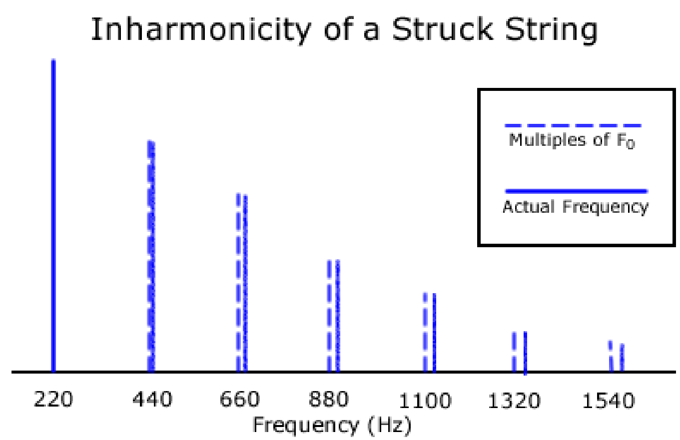
Li Su

2016/03/31

Pitch detection

- Pitch detection from the spectrum
  - Problem 1: missing fundamental
  - Problem 2: inharmonicity
- Periodicity-based pitch detection?





$$f_n = n f_0 \sqrt{1 + Bn^2}$$

“Periodicity” detection

- 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
  - $x(t) = x(t + T_0), \ \forall\, t$
- What is the definition of frequency?
- Find the fundamental frequency/period
- Application: pitch detection, transcription, beat tracking …

Basic idea of periodicity detection

- Formally, a periodic signal is defined as
  - $x(t) = x(t + T_0), \ \forall \, t$
- 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

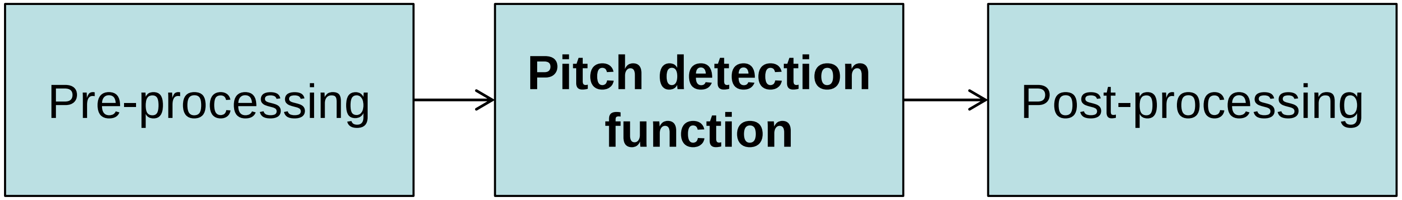
- Autocorrelation function (ACF)
- Average magnitude difference function (AMDF)
- YIN and its periodicity detector
- Generalized ACF and Cepstrum

Autocorrelation function (ACF)

- Cross product 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:
  - $R_{xx}(\tau) = \frac{1}{N-1}\sum_{t=0}^{N-1-\tau} x(t)x(t+\tau)$
- t: time-domain
- τ: lag-domain

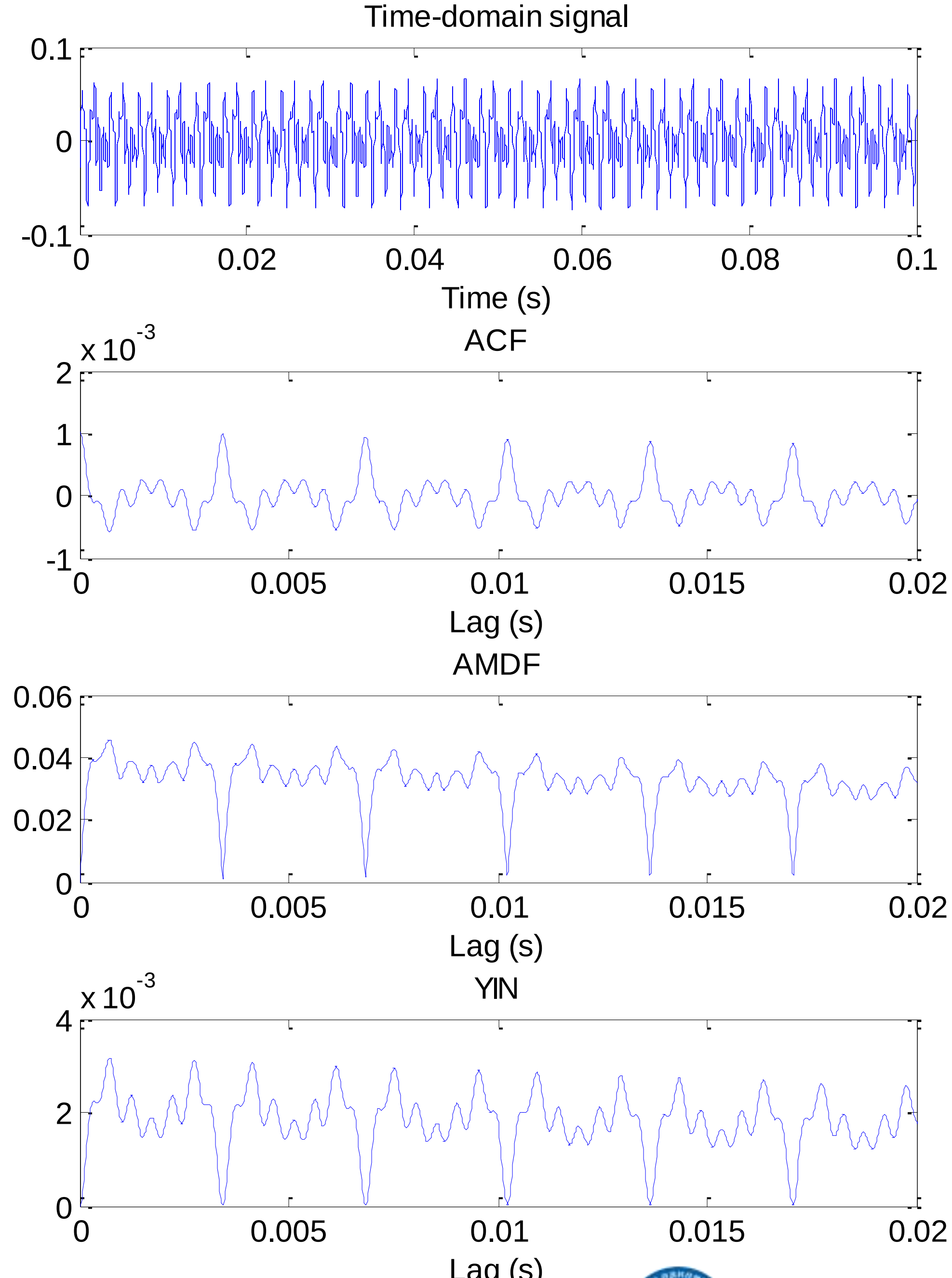
Other relevant pitch detection functions

- 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?

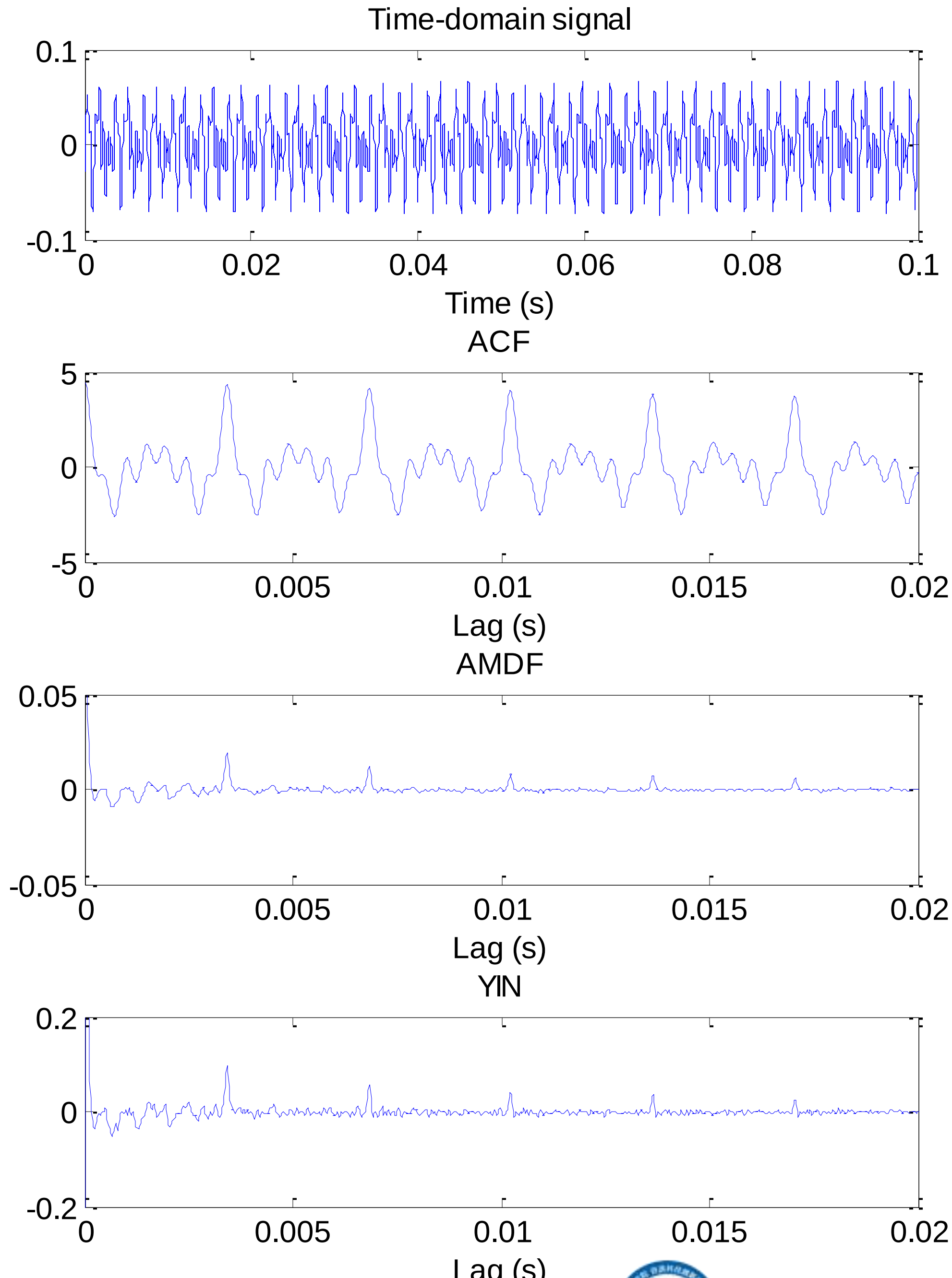
- Wiener-Khinchin theorem: the ACF is the inverse Fourier transform of the power spectrum
  - $R_{xx}(\tau) = IFFT\left(\left|FFT\left(x(t)\right)\right|^2\right)$
  - Complexity: $O(N \log N)$

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:
  - 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.

Preliminary result

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



Cepstrum

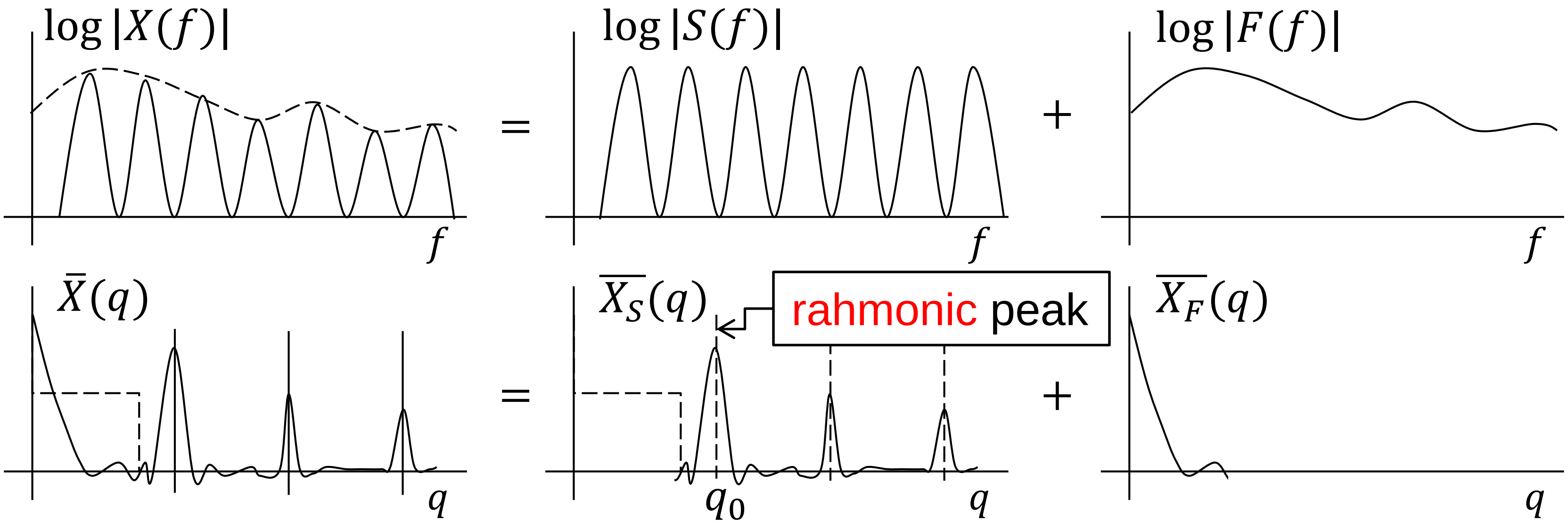
- From spectrum to cepstrum (倒頻譜)
- Spectrum computed by fast Fourier transform (FFT): $X(f) = FFT(x(t))$
- Cepstrum: $\bar{X}(q) = IFFT(\log|X(f)|)$
  - q: quefrency (倒頻率) (not frequency)
  - Quefrency in the cepstrum, and lag in the ACF are both measured in time (but not in the time domain)

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

$s(t) * f(t)$	$S(f)F(f)$	$\log\lvert S(f)\rvert + \log\lvert F(f)\rvert$	$\log\lvert \bar{S}(f)\rvert + \log\lvert \bar{F}(f)\rvert$
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!



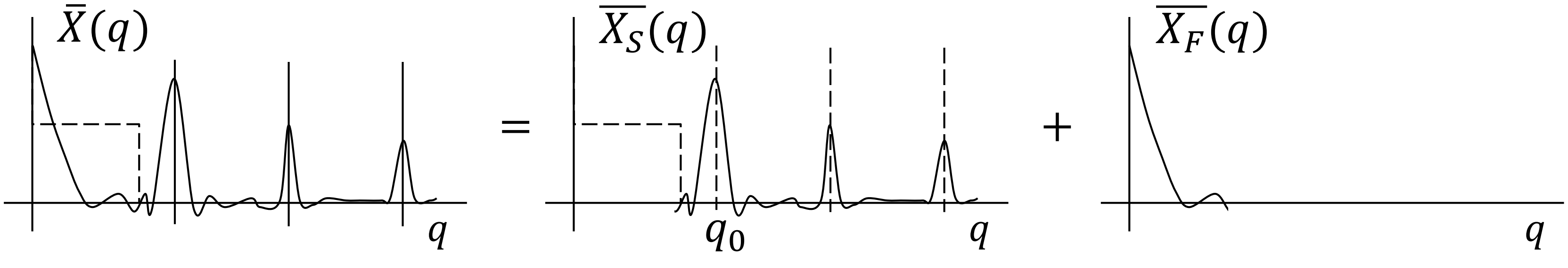
A closer look to the cepstrum

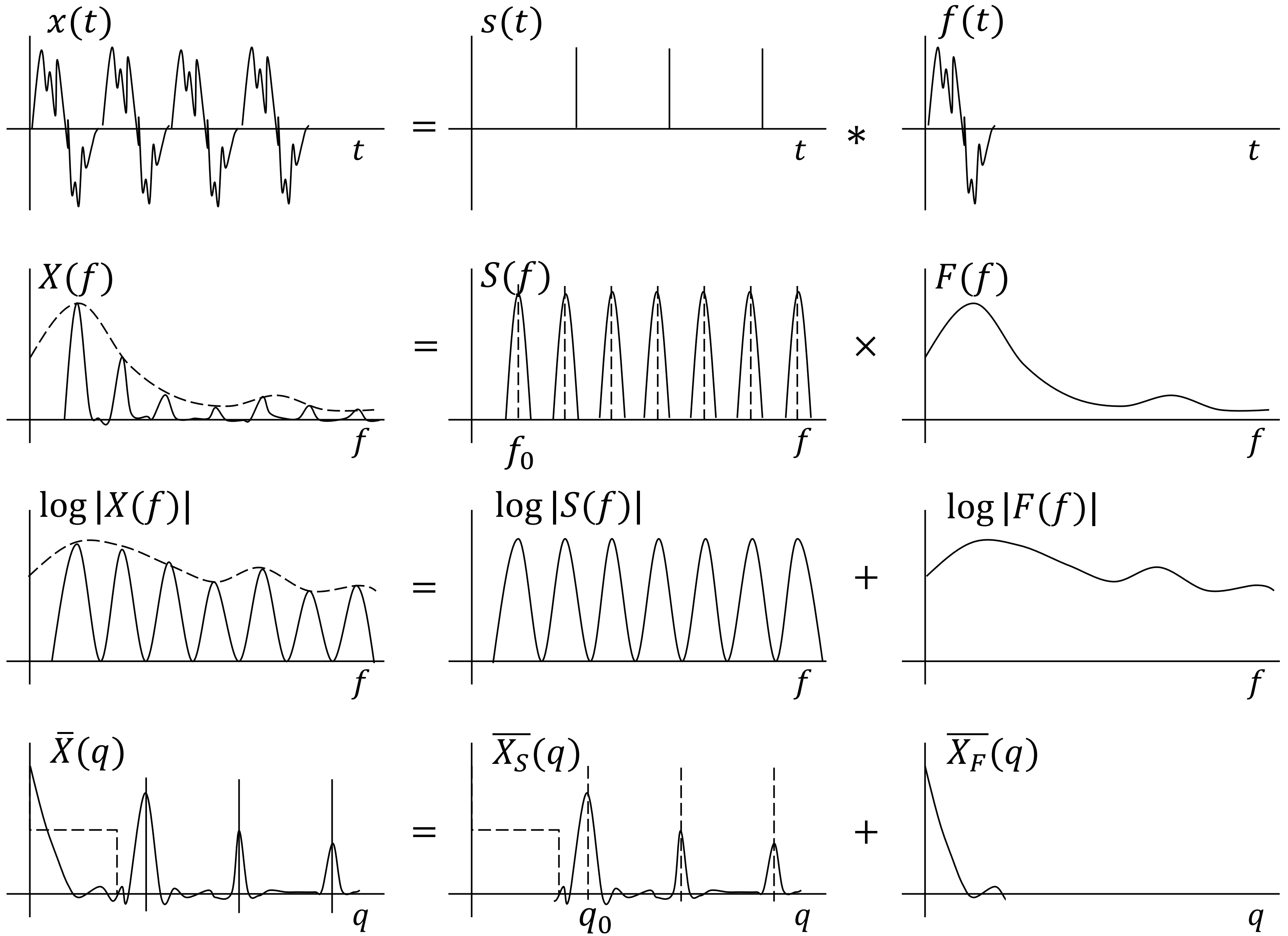
- A physical system: convolution of the excitation and the impulse response
  - Excitation: “fast” spectral variation
  - Impulse response: “slow” spectral variation
- How to do “deconvolution”?
  - Homomorphic signal processing
  - Homomorphism: to “carry over” operations from one algebra system to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

$$\boxed{\text{Convolution}} \xrightarrow{\text{Fourier transform}} \boxed{\text{Multiplication}} \xrightarrow{\text{logarithm}} \boxed{\text{Addition}}$$

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
=
$S(f)$
f_0
f
×
$F(f)$
f
$\log|X(f)|$
f
=
$\log|S(f)|$
f
+
$\log|F(f)|$
f
$\bar{X}(q)$
q
=
$\overline{X_S}(q)$
q_0
q
+
$\overline{X_F}(q)$
q

Generalized logarithm and cepstrum

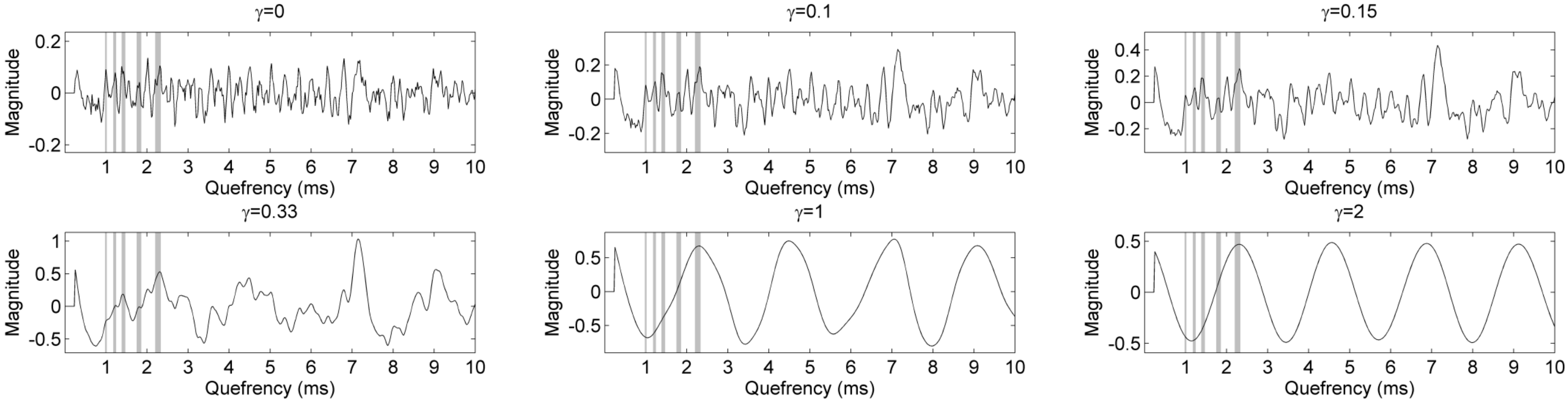
- If we just care about the effect of "nonlinear scaling" (i.e., logarithm) when computing cepstrum
  - Pros: simulate human's perception by compression
  - Cons: sensitive to noise and zeros in the spectrum
- Generalized logarithm:

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

- 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

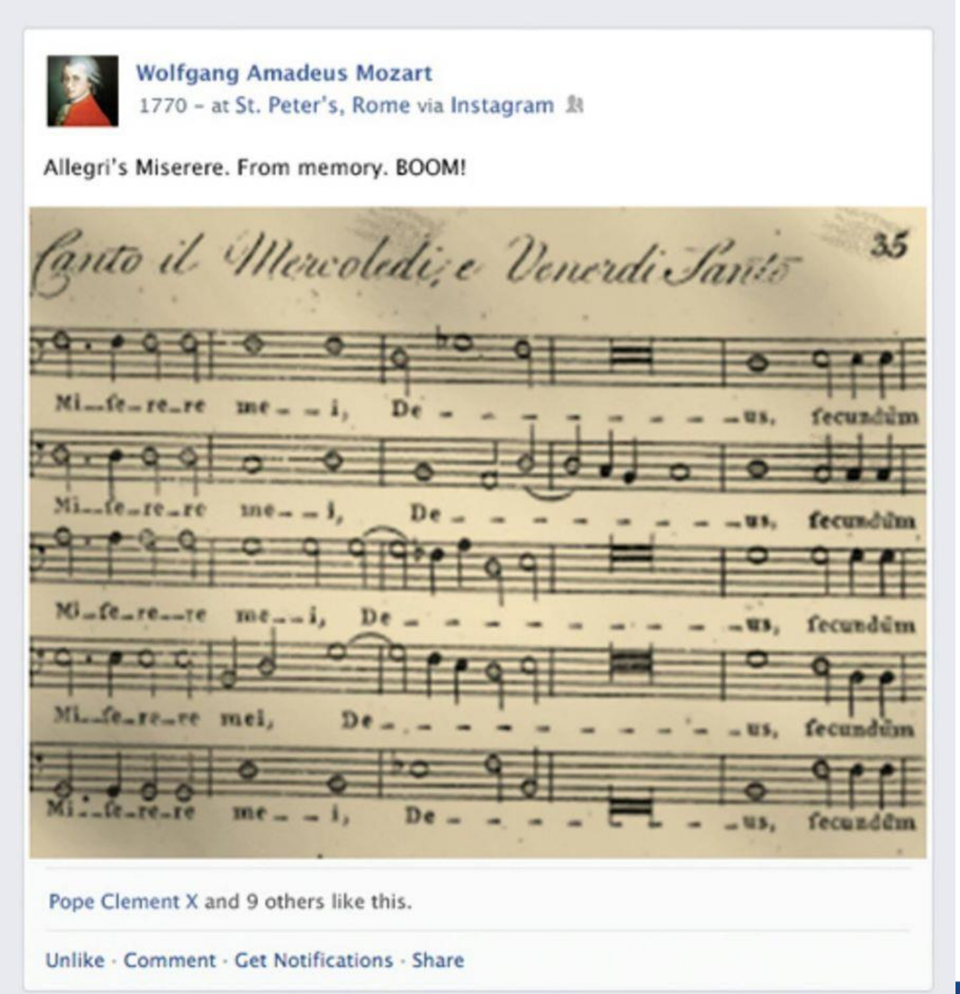
- A complicated example: 5-polyphony piano sample A4+C#5+F5+G#5+B5



- 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.

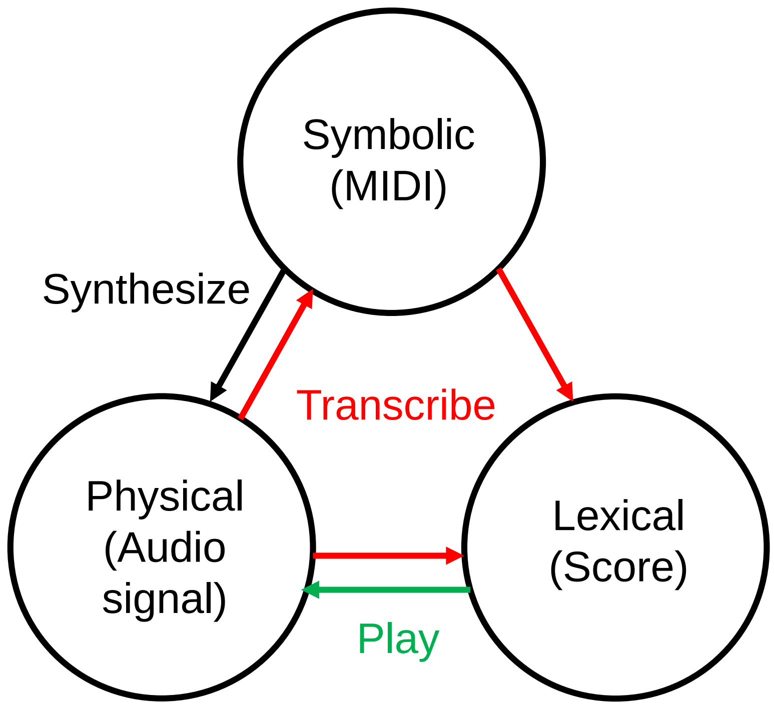
Automatic music transcription

- Can a machine beat a music genius?



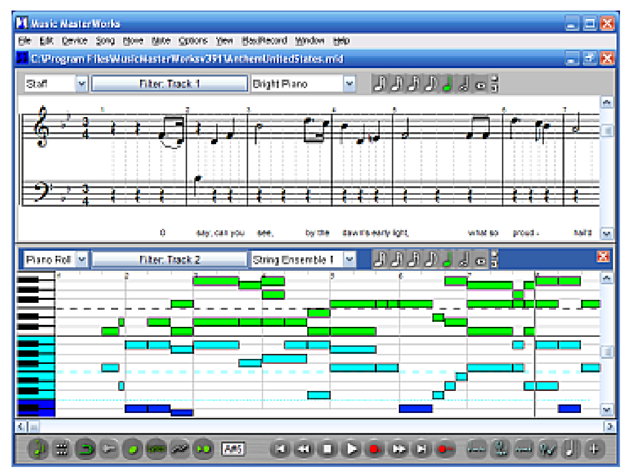
Digital music formats

- 訊號層面 (signal)
  - 原始格式：.wav
  - 壓縮格式：.mp3, .mp4, ……
- 符號層面 (symbolic)
  - 音樂數位介面 (MIDI)
- 文字層面 (lexical)
  - 原始格式：紙本、掃描成 .pdf
  - 可編輯：musicXML
- 主要的音樂轉譜問題
  - WAV to MIDI: automatic music transcription
  - MIDI to musicXML: automatic music transcription (note parsing)
  - PDF to musicXML: optical music recognition (OMR)



音樂數位介面 (MIDI)

- 實際製作、播放、辨識數位音樂時，單純給電腦樂譜資訊是不夠的
  - 舉例：如何定義漸快或漸慢？
  - 反過來問：給定一個C4，從第1.73秒到第2.24秒，它是四分音符還是八分音符？
  - 音樂數位介面(Musical Instrument Digital Interface, MIDI)於1980年代問世
- 相容於(所有的)鍵盤樂器、音效卡、合成器、電子鼓等裝置，內含以下資訊：
  - Onset
  - Duration (offset)
  - Pitch
  - Velocity (力度)



MusicXML

- 相容於多數樂譜編輯軟體如 Finale, Sibelius, MuseScore 等
- 基本格式
  - <part> <measure> <attributes> <divisions>
  - <key>
    - <fifths><mode>
  - <time>
    - <beats><beat-type>
  - <clef>
    - <sign><line> ……


```xml
<part id="P1">
  <measure number="1" width="533.09">
    <print>
      <system-layout>
        <system-margins>
          <left-margin>73.88</left-margin>
          <right-margin>0.00</right-margin>
          </system-margins>
        <top-system-distance>180.00</top-system-
        </system-layout>
      </print>
    <attributes>
      <divisions>2</divisions>
      <key>
        <fifths>1</fifths>
        <mode>major</mode>
        </key>
      <time>
        <beats>4</beats>
        <beat-type>4</beat-type>
        </time>
      <clef>
        <sign>G</sign>
        <line>2</line>
        </clef>
      </attributes>
    <note default-x="98.40" default-y="-45.00">
      <pitch>
        <step>D</step>
        <octave>4</octave>
        </pitch>
      <duration>2</duration>
      <voice>1</voice>
      <type>quarter</type>
      <stem>up</stem>
      <lyric number="1">
        <syllabic>single</syllabic>
        <text>春</text>
        </lyric>
      </note>
```

WAV to MIDI自動採譜的三個層次

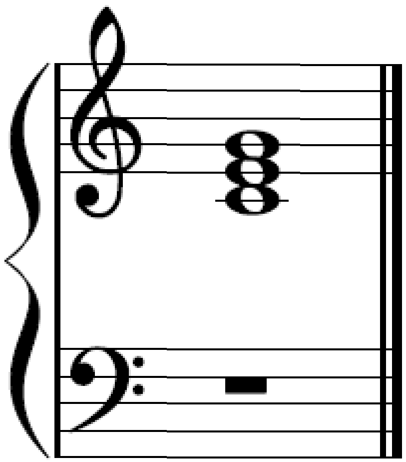
- Multi-pitch estimation (MPE):
  - collectively estimate pitch values of all concurrent sources at each individual time frame, without determining their sources
- Note tracking (NT):
  - estimate continuous segments that typically correspond to individual notes or syllables
- Timbre tracking, streaming:
  - stream pitch estimates into a single pitch trajectory over an entire conversation or music performance for each of the concurrent sources
- 其他WAV to MIDI的相關問題
  - Onset detection
  - Beat tracking / downbeat tracking
  - Meter recognition
  - Chord recognition
  - Structure segmentation

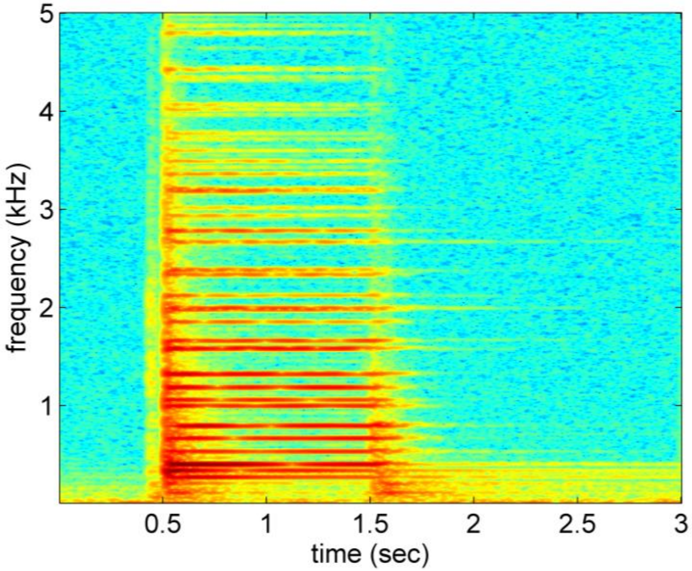
Important papers

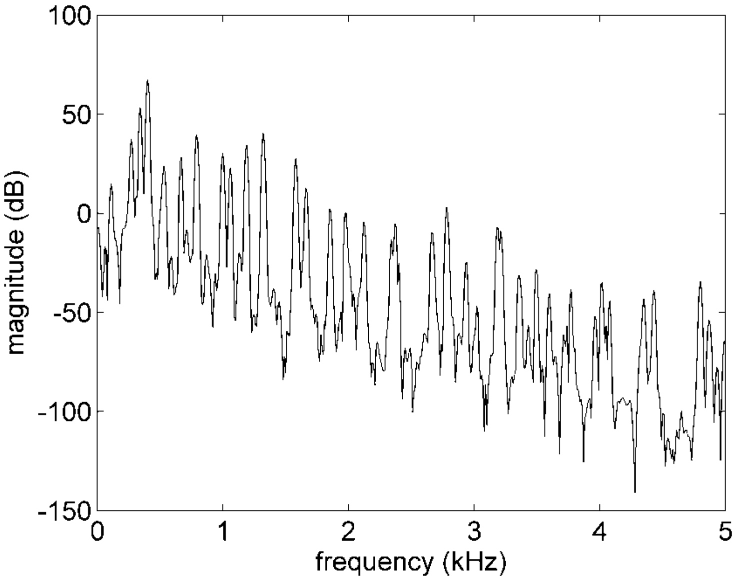
- M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, “Signal processing for music analysis,” *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, Dec. 2011.
- E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: Challenges and future directions,” *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 407–434, 2013.
- Z. Duan, J. Han, and B. Pardo, “Multi-pitch streaming of harmonic sound mixtures,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 138–150, Jan. 2014.

Challenges of multipitch signal (1)

- Example: C major triad (C4+E4+G4)

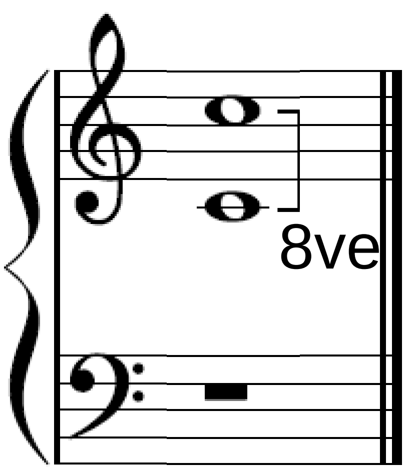


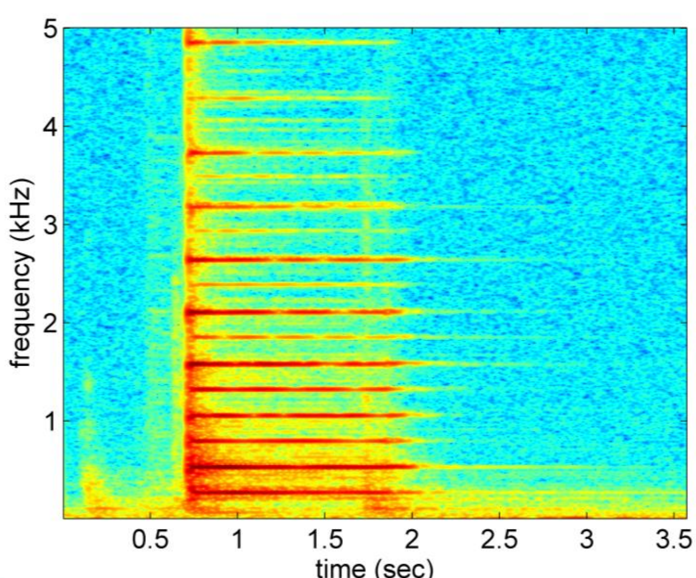


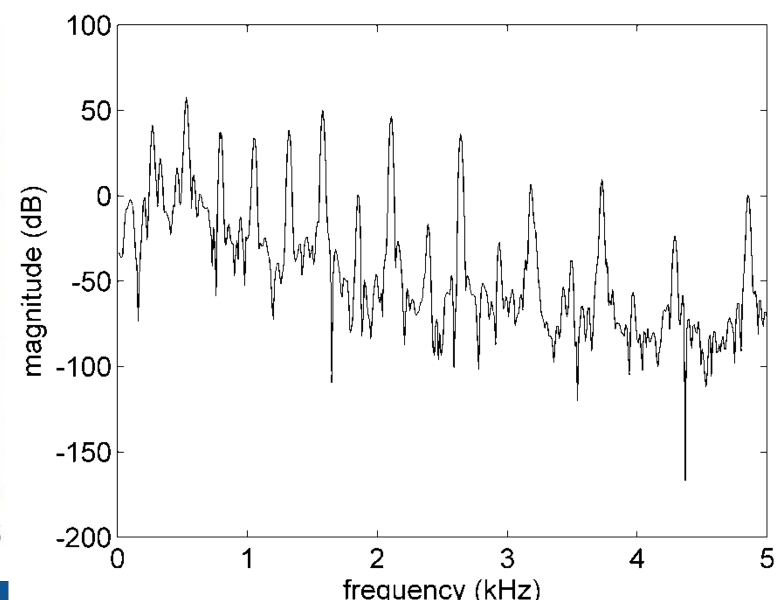


Challenges of multipitch signal (2)

- Octave dual-tone (C4+C5)
- The harmonic series of C5 is fully overlapped with C4
- Ill-posed problem: hard to determine whether C4 or C4+C5
- Can be found by the ratio of even and odd harmonics, but not very efficient in experiment

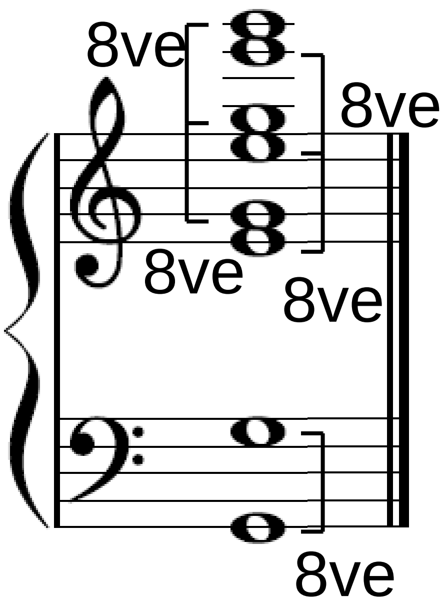


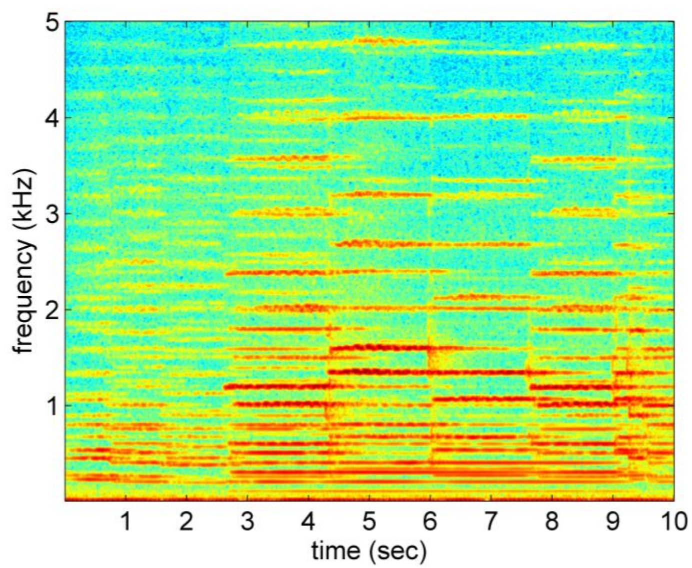


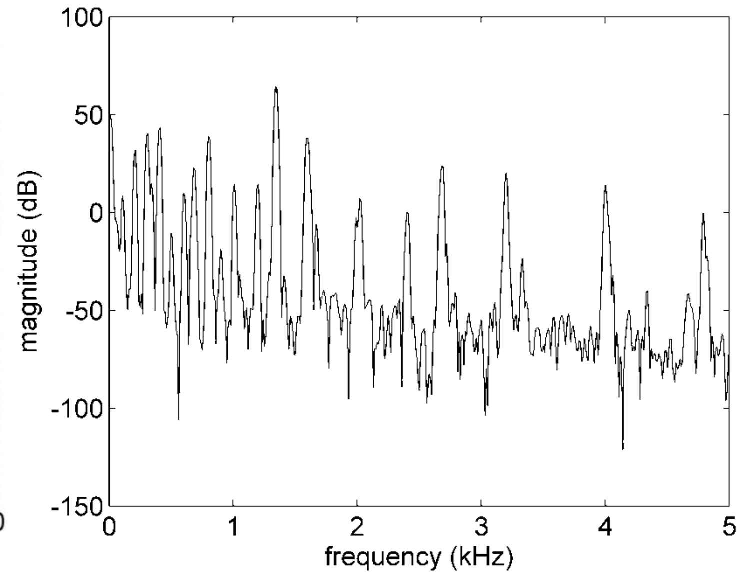


Challenges of multipitch signal (3)

- 複雜的音型 (海頓/驚愕交響曲第一樂章/第25.5秒)
- Who is who？







Challenges in multipitch estimation

- The best grade in Music Information Retrieval Evaluation eXchange (MIREX) Multi-F0 challenges:
  - Multi-pitch estimation: 72.3 % accuracy (Anders Elowsson et. al, 2014)
  - Note tracking: 58.2 % accuracy (Anders Elowsson et. al, 2014)
  - Evaluated on only 30 woodwind quintets and 10 piano clips
- Challenge (1)：overlapped harmonics (octaves, fifths)
- Challenge (2)：noise
- Challenge (3)：threshold of detection
- Challenge (4)：labeled dataset
  - We have not enough labeled data!!
- Challenge (5):timbre complexity
  - One pitch played by multiple instruments
- Challenge (6)：efficiency

State of the art

- Feature-based (expert knowledge-based)
  - Using audio features derived from the input time-frequency representation (e.g., spectrum, autocorrelation,…), and designing *pitch salience function*
- Statistical model-based
  - Maximum a posteriori (MAP) estimation problem
- Matrix factorization-based
  - Non-negative matrix factorization (NMF)
  - Dictionary-based methods

Dictionary-based pitch detection: basic

- From frequency representation
- A "dictionary" $\mathbf{D} \in R^{m \times n}$ be a set of spectral features
- $\mathbf{D} = [d_1, d_2, \cdots, d_n]$, 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, $\boldsymbol{\alpha} \in R^m$

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

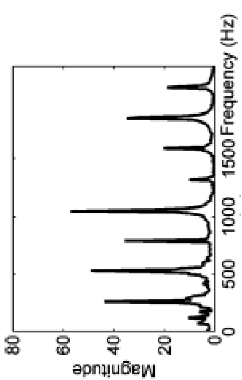
or

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

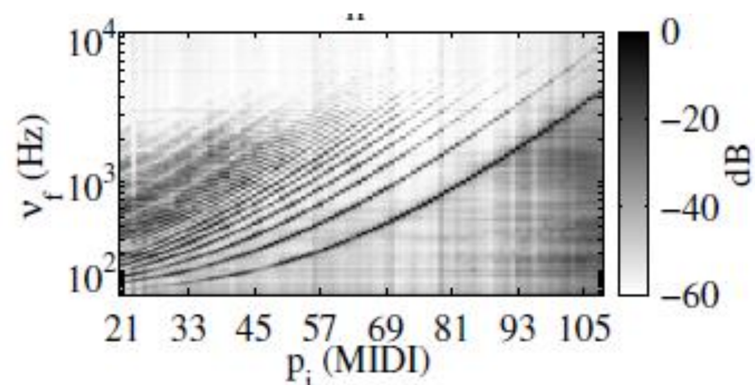
Basic template matching: single pitch detection

- 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

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



C4

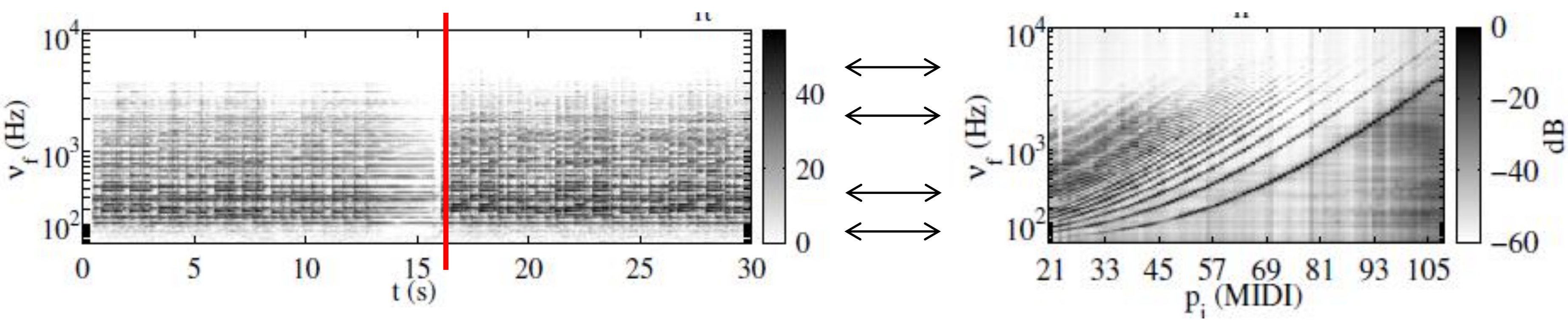


Templates from A0 to C8

How about polyphonic signals?

- Find the atoms having the k-th largest $\mathbf{x} \cdot d_k$
- k-nearest neighbor (kNN)
- Or, how about the following formulation?

$$\min \quad \|\boldsymbol{\alpha}\|_0 \ \text{ s.t. } \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2 < \epsilon$$



From: E. Vincent et. al, "Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation," IEEE TASLP 2010

Sparse coding

- Perfect reconstruction

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

- Approximation, hard constraint

$$\text{minimize} \quad \|\boldsymbol{\alpha}\|_0 \;\; \text{s.t.} \quad \|\mathbf{x} - \mathbf{D\alpha}\|_{\mathbf{2}} < \epsilon$$

- Approximation, soft constraint

$$\text{minimize} \quad \|\mathbf{x} - \mathbf{D\alpha}\|_{\mathbf{2}} + \lambda\|\boldsymbol{\alpha}\|_0$$

Some concepts revisited

- Assume $\mathbf{D}$ full rank,

Condition	Example Solution	Application
• m>n: "skinny" $\mathbf{D}$ • $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$: an over-determined system • $\mathbf{D}$: an under-complete dictionary	Least square error: $\boldsymbol{\alpha} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\boldsymbol{x}$ …	Regression Curve fitting …
• m<n: "fat" $\mathbf{D}$ • $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$: an under-determined system • $\mathbf{D}$: an over-complete dictionary	Least norm solution: $\boldsymbol{\alpha} = \mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}\boldsymbol{x}$ Sparse solution: min $\|\boldsymbol{\alpha}\|_0$ s.t. $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ …	Signal recovery Feature selection …

Some points in sparse coding

- Overcompleteness
  - For $n > 2m$, sparse solution is guaranteed
- L1-norm regularization
  - L0-norm is non-convex (no guarantee of global optimal solution)
  - Use L1-norm instead of L0-norm (a compromise between convexity and sparsity)

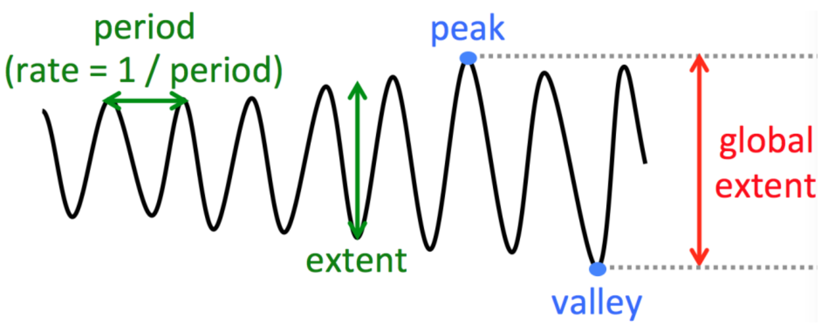
$$\mathrm{argmin}_{\boldsymbol{\alpha}} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_{\mathbf{2}} + \lambda\|\boldsymbol{\alpha}\|_1$$

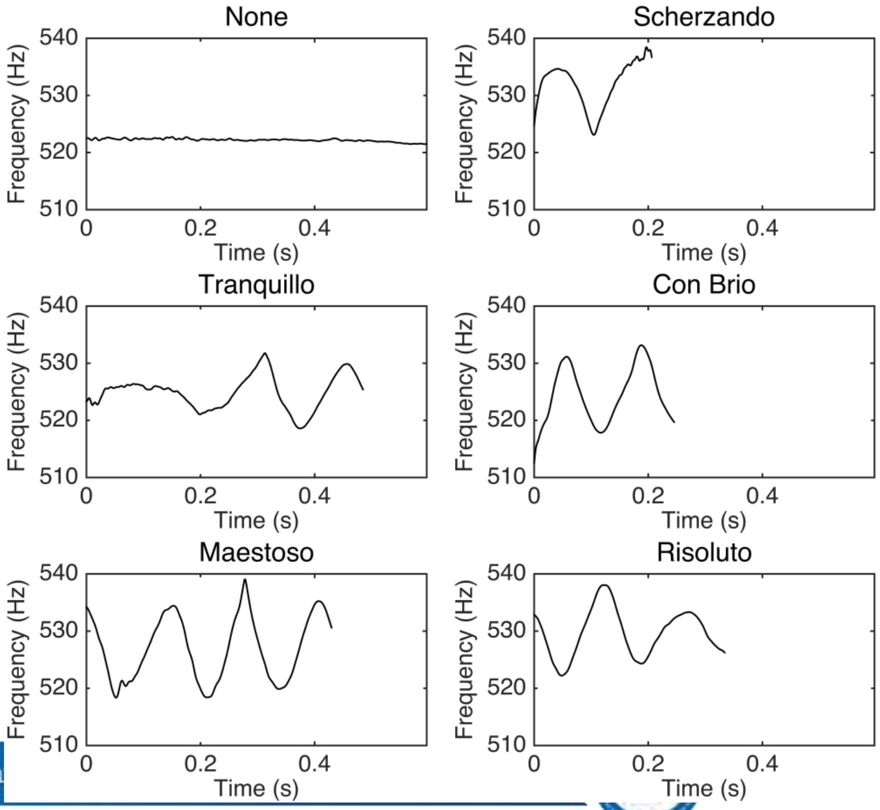
Pitch tracking and streaming

- 偵測音符的起始點(onset)、音高(pitch)和終止點(offset)
- 完成pitch tracking，才算是具備實際用途的自動採譜演算法
- 比MPE更精細複雜的問題
  - Repeating note
  - Tie, fermata
  - Legato, portato, staccato
  - Trill, mordant, grace note
  - Vibrato
  - Tremolo
  - Slide
- The challenge of streaming
  - Clustering of timbre feature
  - Multiple instrument recognition

Example: vibrato

- Pitch contours of the first note of Mozart's Variationen (C5) interpreted in 6 different musical terms: None, Scherzando, Tranquillo, Con Brio, Maestoso, and Risoluto





Instantaneous frequency estimation

- How accurate can we estimate the fundamental frequency?
- The "finest" grid in frequency (Δf) and periodicity ($\Delta\tau$) representation
  - Frequency-based method: $\Delta f = f_s/N$
  - Periodicity-based method: $\Delta\tau = 1/f_s$
- Super-resolution methods
  - Interpolation
  - Higher-order physical quantities
- Example: use the temporal difference of STFT phase to calibrate the instantaneous frequency
- Reference: Justin Salamon and Emilia Gómez. "Melody extraction from polyphonic music signals using pitch contour characteristics." *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 6, pp. 1759-1770, 2012.