# Source separation

Li Su

# Source separation problems

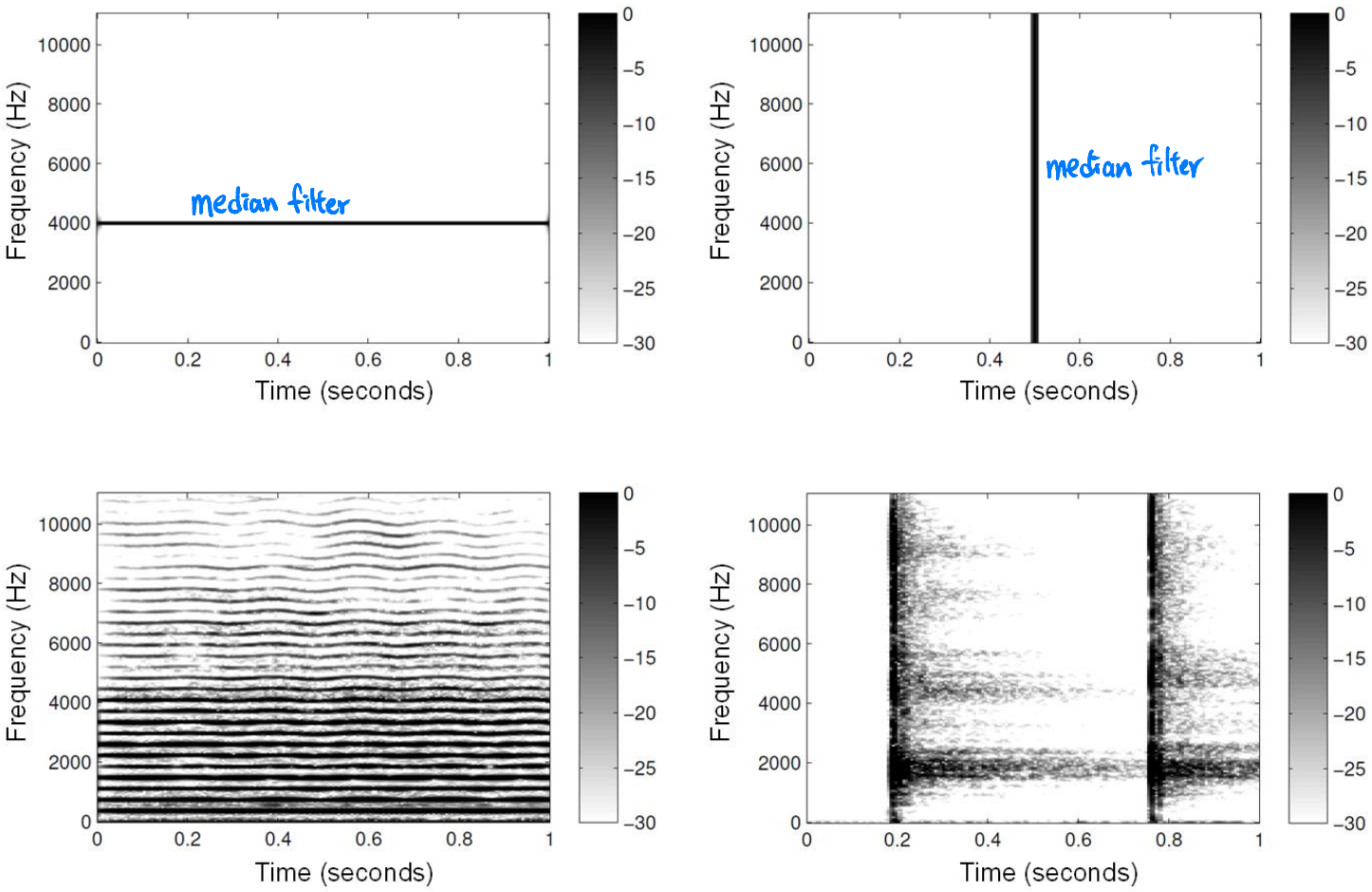
- Blind source separation (BSS) Uniformed：不知道有幾個 source
- Voice separation
  - (Typically) separate voice from accompaniment
- Binaural or multi-aural source separation
  - Independent component analysis (ICA) problems
- Monaural source separation 只有一個聲道，含有多個訊號 (本身為 ill-defined problem)
- Informed source separation (ISS)
  - Score-informed source separation (SISS)
  - User-guided source separation

# Sources and mixtures

- Sound mixtures are superposition of waves
  - No opacity (unlike vision)
  - Interference: is a silent sound composed of two sources with opposite phase?
  - Source: one kind of auditory scene
- Perception of source
  - A subjective construction
- Ill-posed problem

# Harmonic–Percussive Separation (HPSS)

Blind source seperation 最簡單的一種問題




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

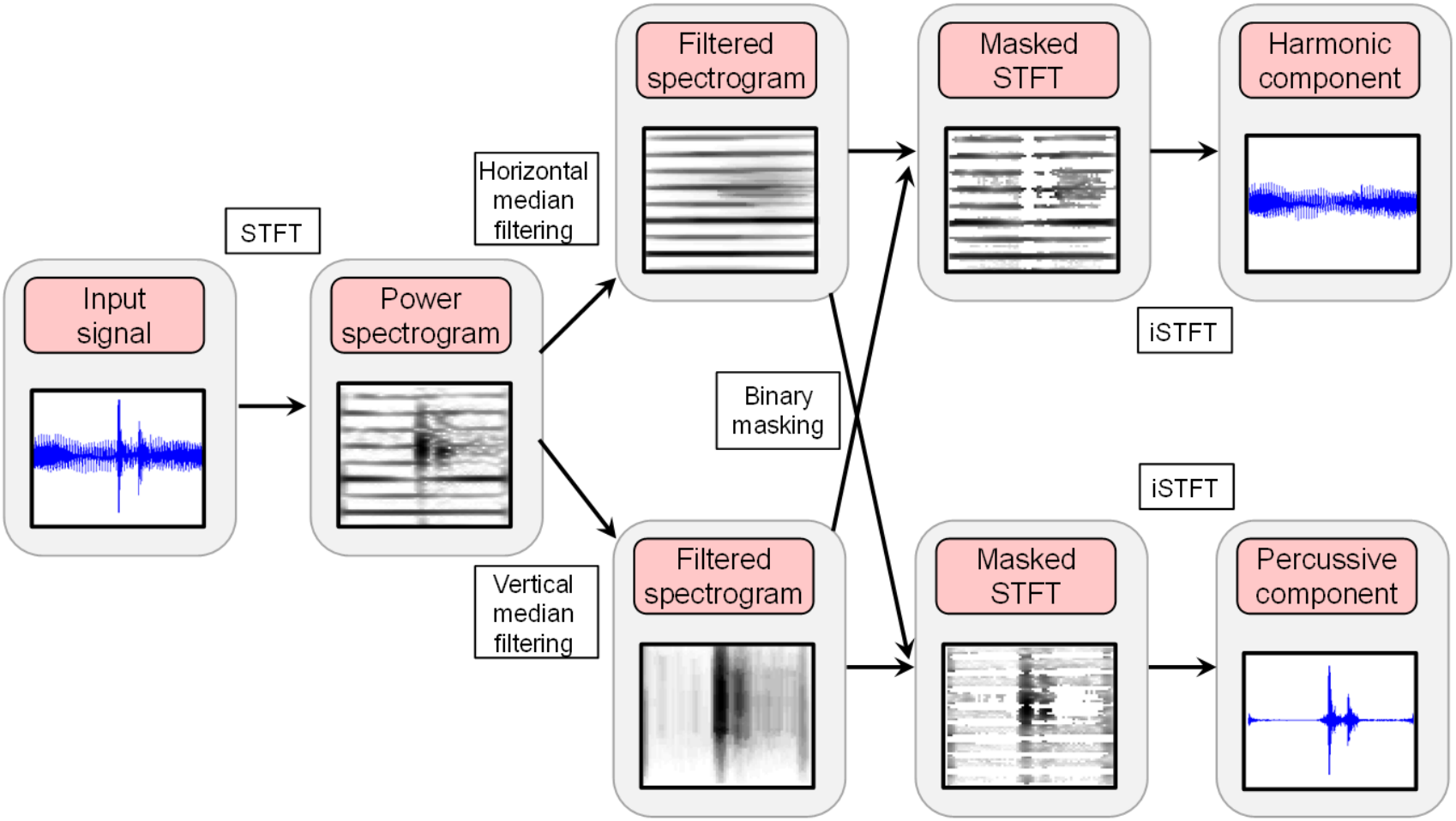
# HPSS with median-filtering

- Simple DSP techniques; no ML
- Intuition
  - stable **harmonic** or stationary components form horizontal ridges on the spectrogram
  - **percussive** components form vertical ridges with a broadband frequency response
- Related papers
  - Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram, EUSIPCO 2008
  - Harmonic/percussive separation using median filtering, DAFx 2010

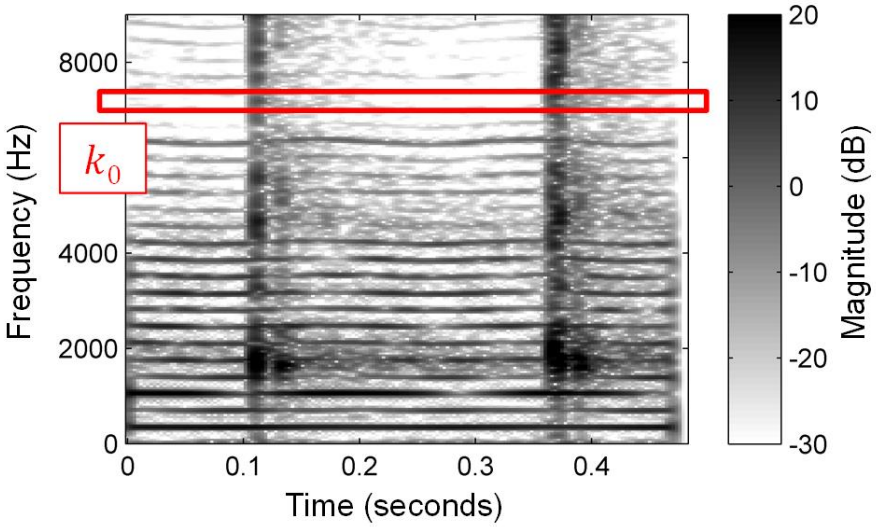
# Some simple filtering techniques

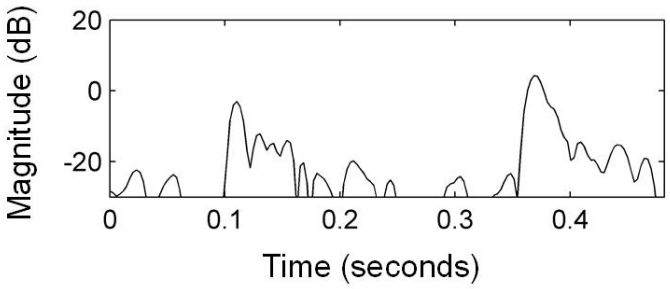
- Filtering: mean, max, median
- *Finite* filter length, e.g. filter of length 3
  - Input: [0 0 1 0 1 0 1 1 0 1 1 1 0]
  - Median: [ 0 0 1 0 0 1 1 1 1 1 1 ]
  - Max: [ 1 1 1 1 1 1 1 1 1 1 1 ]
- Pooling
- Filter length = number of samples
  - e.g. [9 0 1 3 1 2 5] -> {mean=3, max=9, median=**?**}
  - (note: to calculate median you need to ***sort*** the values)

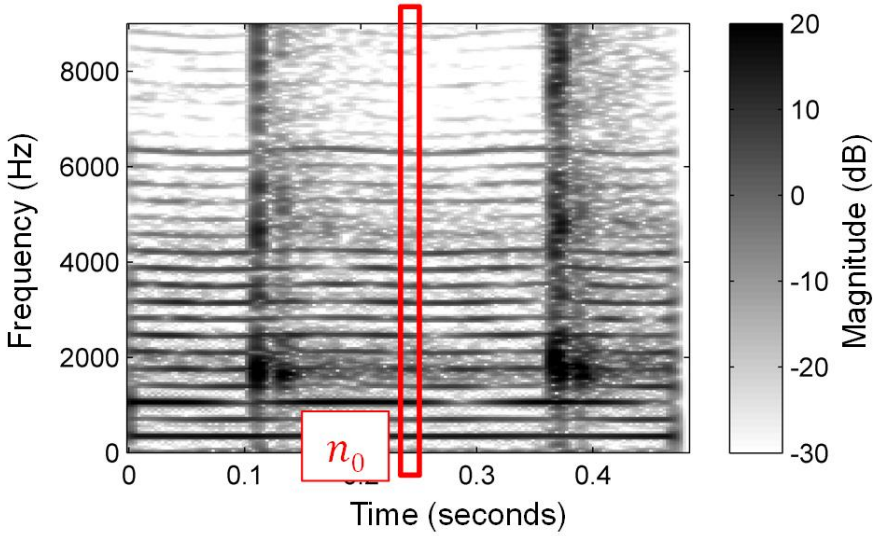
# Flow chart

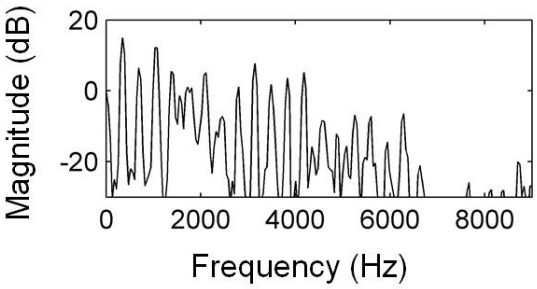



From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015



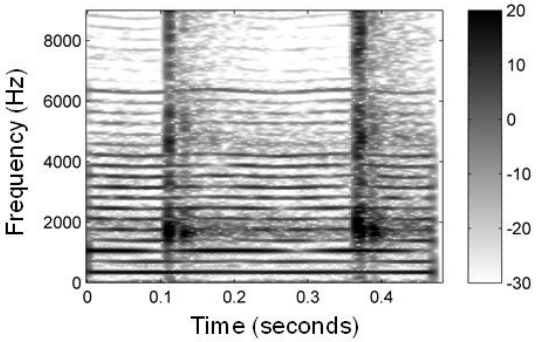



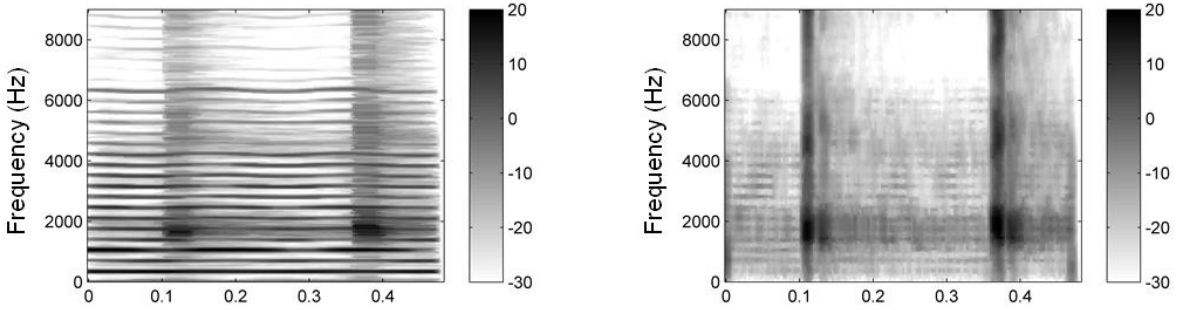

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Different filter sizes

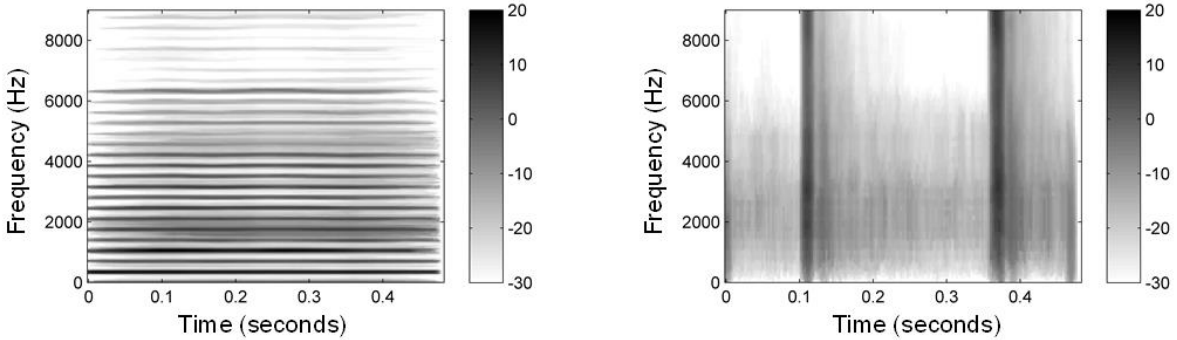
$$\tilde{Y}_h(t,k) := \text{median}(Y(t-\ell_h,k),\ldots,Y(t+\ell_h,k))$$
$$\tilde{Y}_p(t,k) := \text{median}(Y(t,k-\ell_p),\ldots,Y(t,k+\ell_p))$$




Smaller filter size




Larger filter size



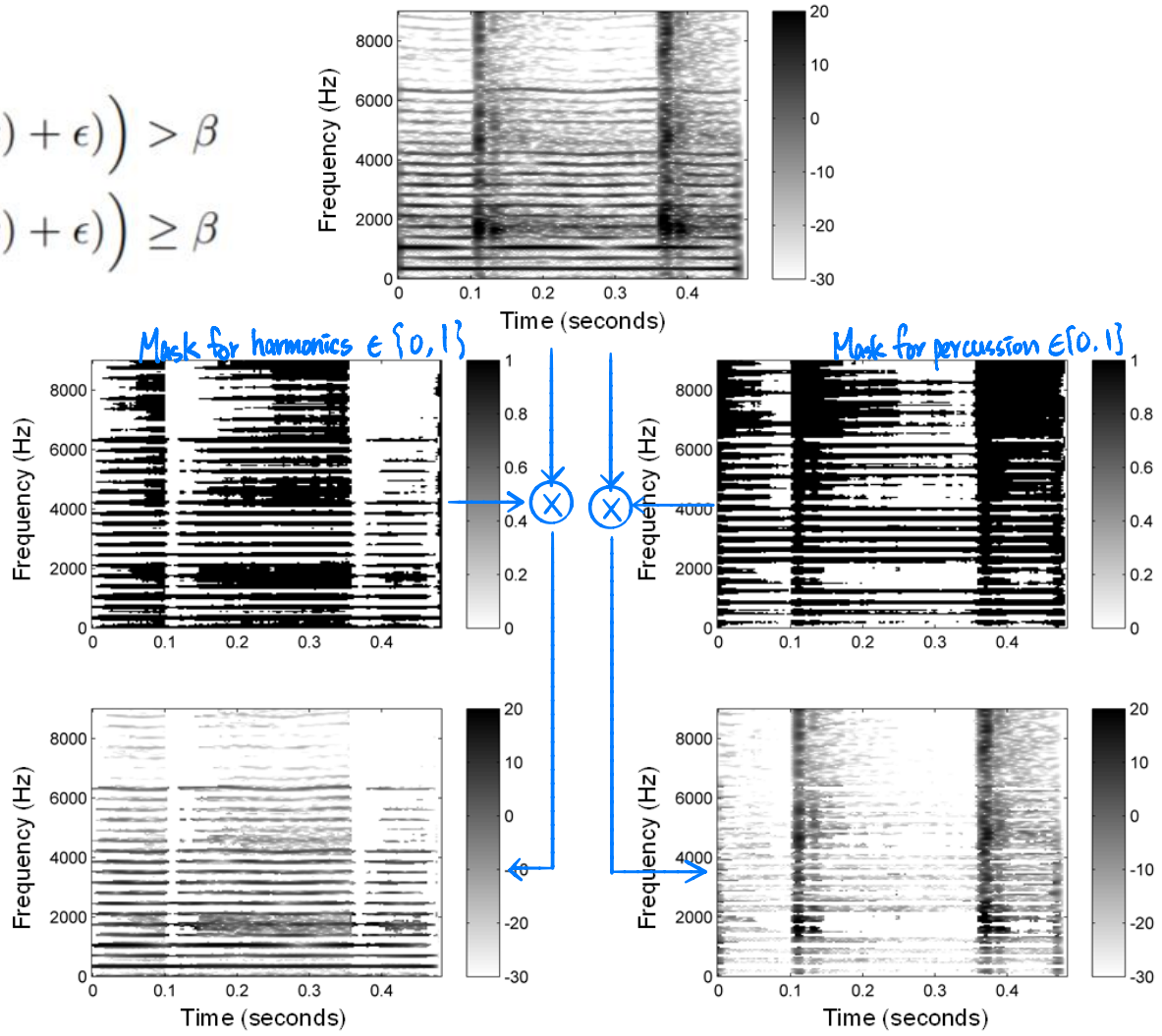

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Masking 重要觀念

$$M_h(t,k) := \left(\tilde{Y}_h(t,k)/(\tilde{Y}_p(t,k)+\epsilon)\right) > \beta$$
$$M_p(t,k) := \left(\tilde{Y}_p(t,k)/(\tilde{Y}_h(t,k)+\epsilon)\right) \geq \beta$$

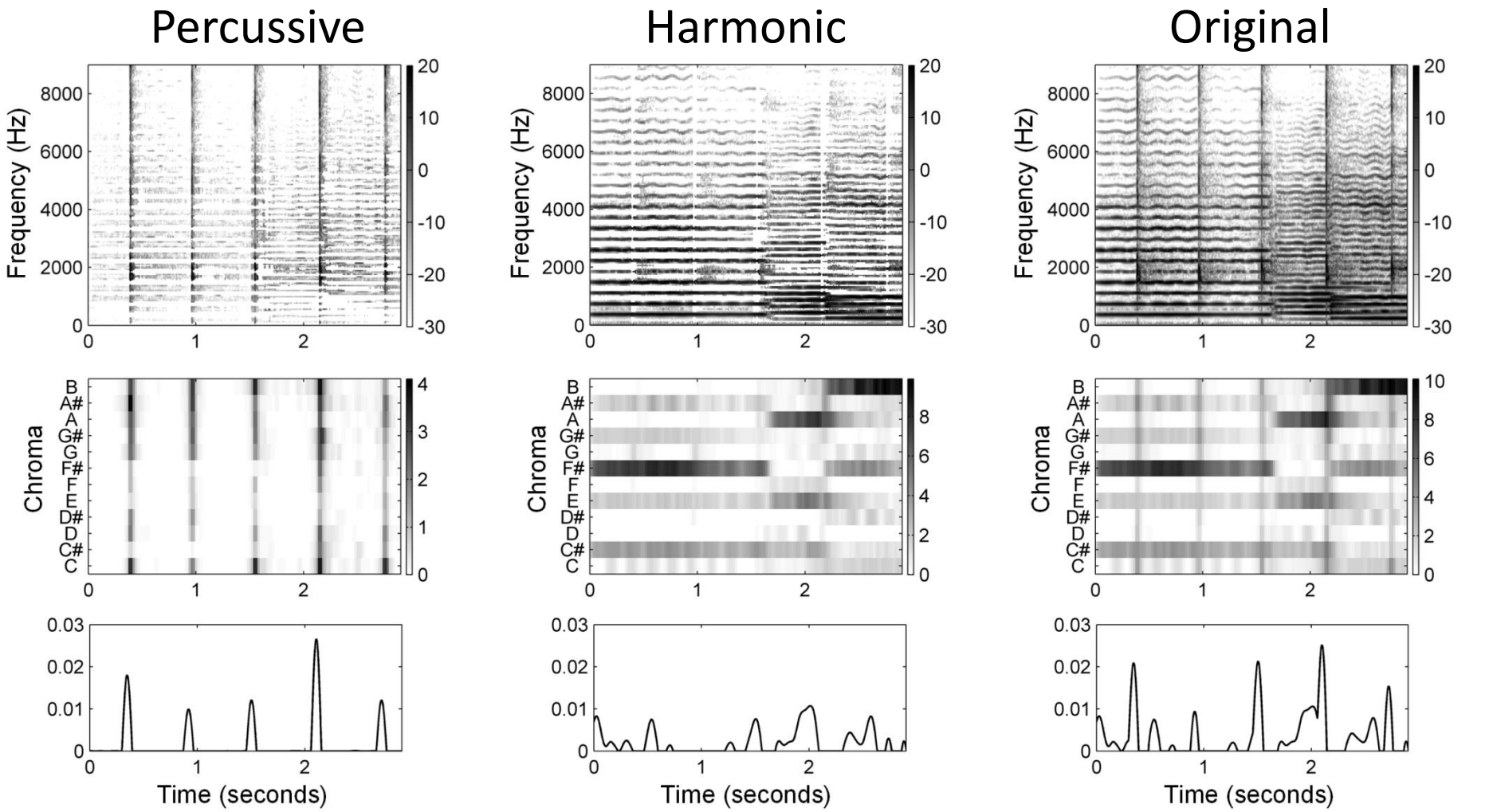
$$X_h(t,k) := X(t,k)\cdot M_h(t,k)$$
$$X_p(t,k) := X(t,k)\cdot M_p(t,k)$$

- Thresholding
- Binarization
- Masking




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

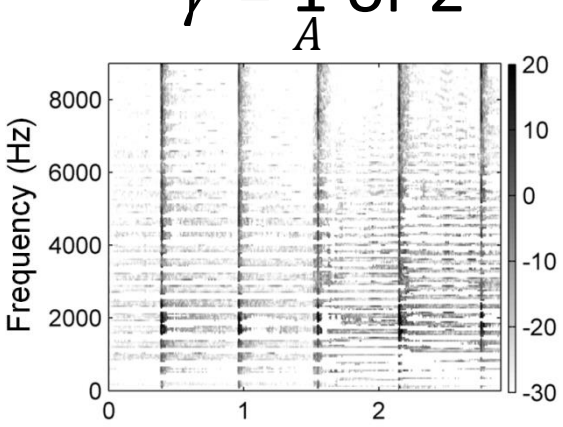
# Separation result

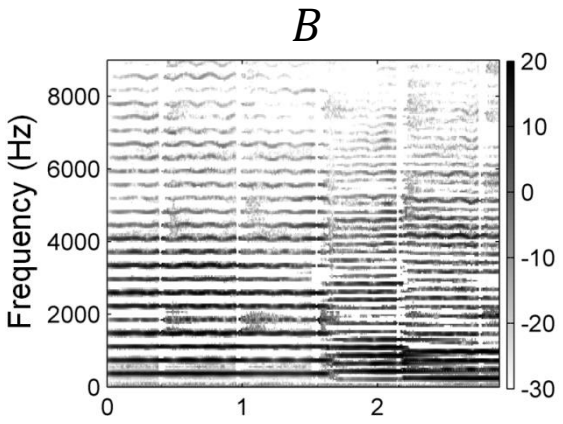


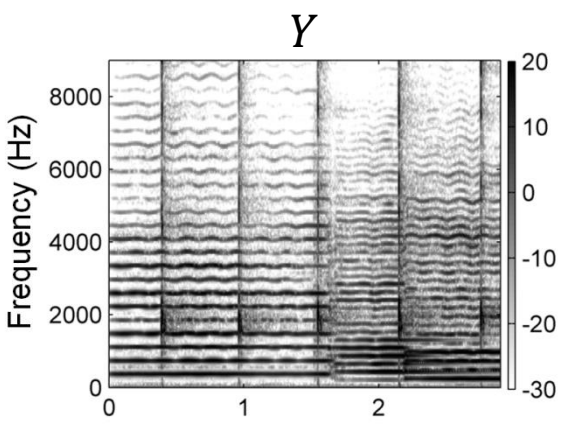

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Wiener Filter (Soft)

- $M_A(f,t) = \frac{A(f,t)^\gamma}{A(f,t)^\gamma + B(f,t)^\gamma}$ mask 亦可為 soft ($\in [0,1]$)
- $\hat{A} = M_A \odot Y$
- Use $\hat{A}$ instead of $A$ in the ISTFT → 未必 always 可逆 把分好的 spectrogram 轉回訊號
- $M_A$ is referred to as a *soft mask*
- $\gamma$ = 1 or 2

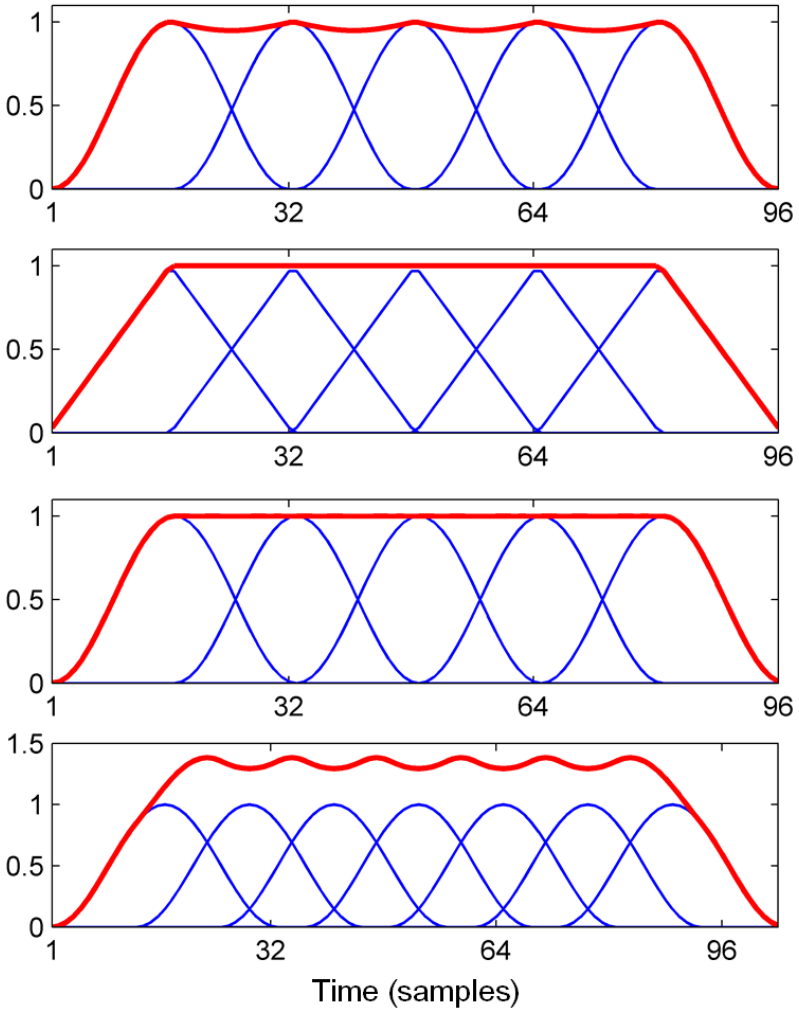



From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Overlap-add (OLA) method

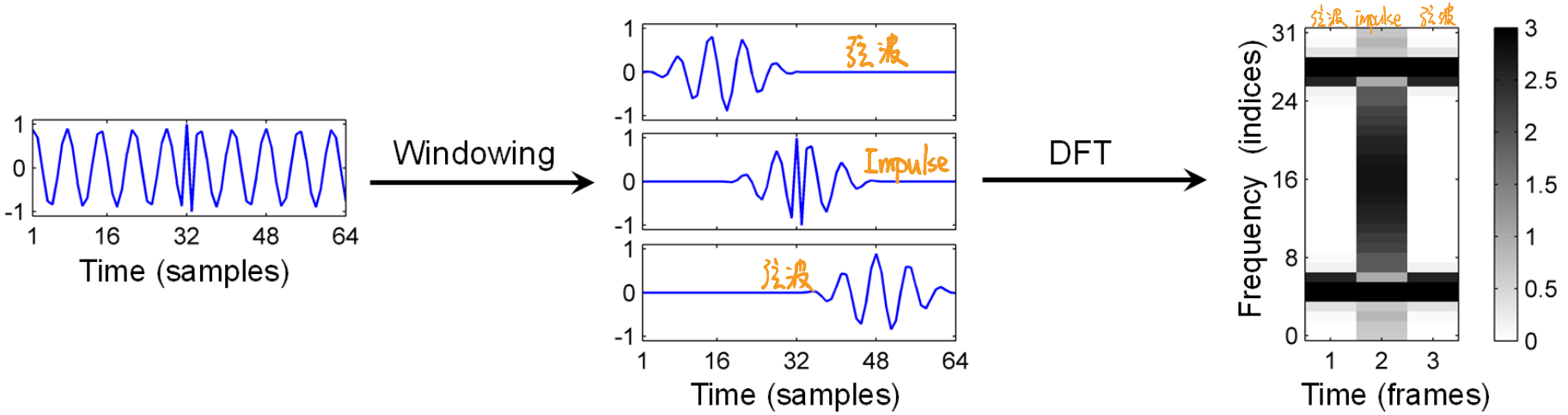



Perfect Reconstruction 條件
→ Overlap 約多 (75% ↑)
& 用好的 window function

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Overlap-add (OLA) method (2)

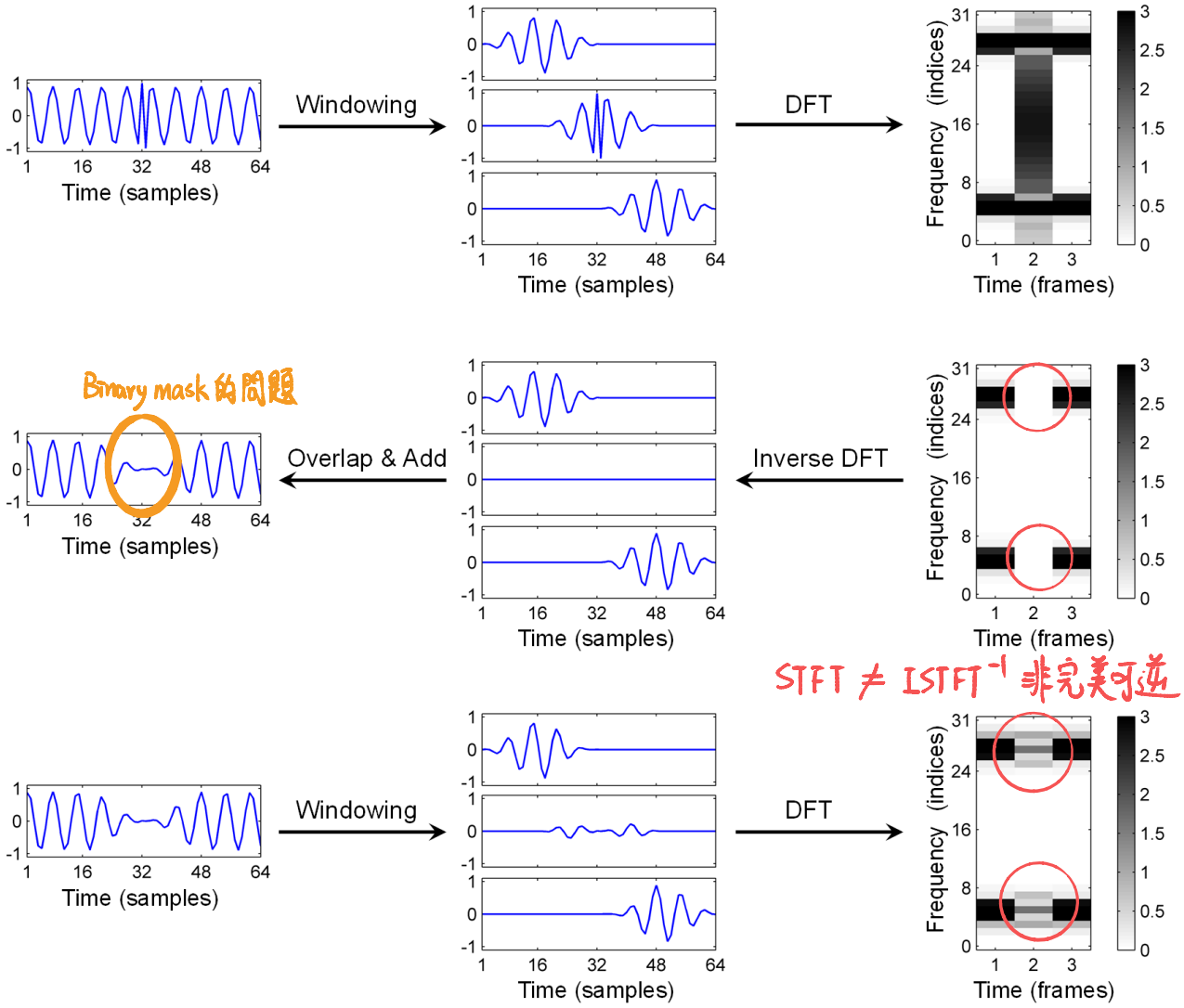
- STFT




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Overlap-add (OLA) method (3)

- Inverse STFT




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Parameters

- Window size, hop size
  - Reconstruction method
  - Filter length

直接影響

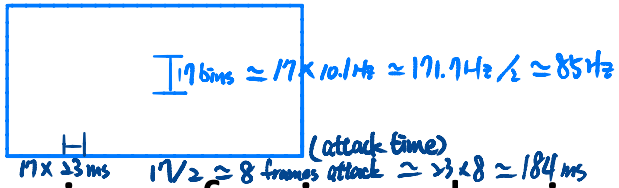
- Q1: Given sampling rate = 44.1 kHz, FFT window size = 4096 samples, hopsize = 1024 samples, what's the physical meaning of using a vertical (percussive) median filter length=17?

0.1 second

≥ 23 ms

44100 / 4096 ≃ 10.1 Hz (frequency resolution)

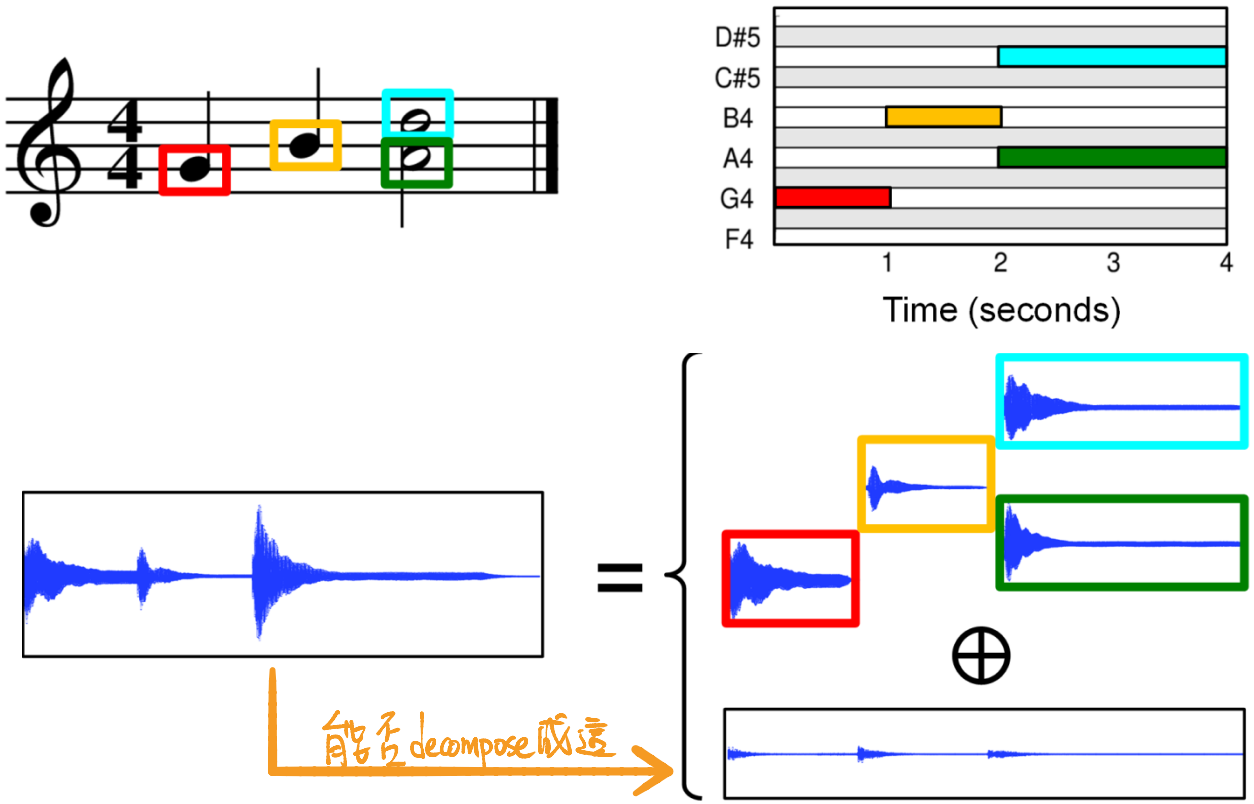
C4 = 261.6
A4 = 440
C5 = 523




- Q2: What's the physical meaning of using a horizontal (harmonic) median filter length=17?

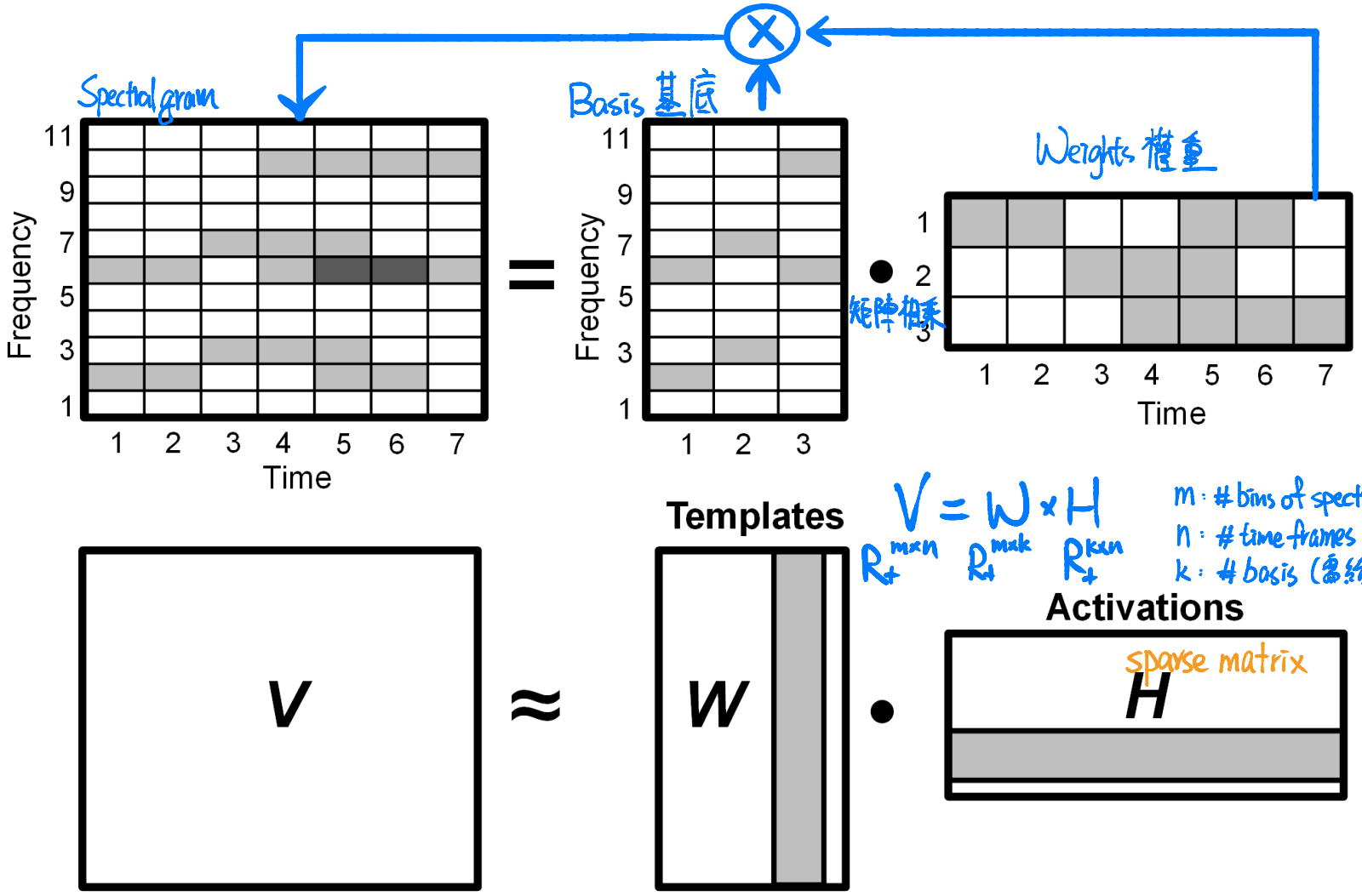
# NMF-Based Audio Decomposition

- Non-negative matrix factorization (NMF)




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Principles of NMF 下面三個 matrices 的所有 elements 皆為非負數




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# The NMF problem

- Minimize the input data $\boldsymbol{V}$ and model $\boldsymbol{WH}$, subject to non-negativity of $\boldsymbol{W}$ and $\boldsymbol{H}$

距离函数：Euclidean, KL divergence

$$\min_{W,H\geq 0} D(\boldsymbol{V}|\boldsymbol{WH})$$

通常希望H是sparse & 时间上是smooth

- $D(\boldsymbol{V}|\boldsymbol{WH})$ represents the distance between $\boldsymbol{V}$ and $\boldsymbol{WH}$
- Regularization terms are often added to $D(\boldsymbol{V}|\boldsymbol{WH})$ in order to favor sparsity or smoothness of $\boldsymbol{W}$ and $\boldsymbol{H}$
- Paper: Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788.

# Properties of NMF

- Easily preserver nonnegativity
  - Find templates and activations simultaneous
  - Easy to implement
  - Fast (of complexity O(FKN) per iteration)
  - Zeros remain zeros! 一旦有個 element 被 update 成 0，之後的 iteration 都會統統是 0
- Can be set as either an unsupervised or a supervised learning problem
  給定名樂器的W(basis)

# Non-negative matrix factorization (NMF)

- NMF: $\mathbf{V} \approx \mathbf{WH}$
- $\mathbf{V} \in R^{+M \times N}$, $\mathbf{W} \in R^{+M \times K}$, $\mathbf{H} \in R^{+K \times N}$
- Update rule (error measured by KL divergence):

Iteratively update W and H
直到收斂

$$W_{mk} \leftarrow W_{mk} \sum_n \frac{V_{mn}}{(WH)_{mn}} H_{kn}$$

$$W_{mk} \leftarrow \frac{W_{mk}}{\sum_m W_{mk}}$$

$$H_{kn} \leftarrow H_{kn} \sum_m W_{mk} \frac{V_{mn}}{(WH)_{mn}}$$

# Distance functions for NMF

- $\beta$-divergence

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^\beta\right), & \beta \neq 0,1 \\ x(\log x - \log y) + (y - x), & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1, & \beta = 0 \end{cases}$$

- $\beta = 2$: Euclidean distance; $\beta = 1$: KL divergence; $\beta = 0$: IS divergence
- Scaling property of $\beta$-divergence

$$d_\beta(\gamma x|\gamma y) = \gamma^\beta d_\beta(x|y)$$

# Beta NMF

- Beta NMF

$$W \leftarrow W \frac{\left[(WH)^{\cdot(\beta-2)} \cdot V\right] H^T}{[WH]^{\cdot(\beta-1)} H^T}$$

$$H \leftarrow H \cdot \frac{W^T\left[(WH)^{\cdot(\beta-2)} \cdot V\right]}{W^T [WH]^{\cdot(\beta-1)}}$$

- Févotte, Cédric, and Jérôme Idier. "Algorithms for nonnegative matrix factorization with the β-divergence." Neural computation 23.9 (2011): 2421-2456.

# Algorithm

**Algorithm:** NMF ($V \approx WH$)

**Input:** Nonnegative matrix $V$ of size $K \times N$
Rank parameter $R \in \mathbb{N}$
Threshold $\varepsilon$ used as stop criterion

**Output:** Nonnegative template matrix $W$ of size $K \times R$
Nonnegative activation matrix $H$ of size $R \times N$

**Procedure:** Define nonnegative matrices $W^{(0)}$ and $H^{(0)}$ by some random or informed initialization. Furthermore set $\ell = 0$. Apply the following update rules (written in matrix notation):

(1) $H^{(\ell+1)} = H^{(\ell)} \odot \left( ((W^{(\ell)})^\top V) \oslash ((W^{(\ell)})^\top W^{(\ell)} H^{(\ell)}) \right)$

(2) $W^{(\ell+1)} = W^{(\ell)} \odot \left( (V(H^{(\ell+1)})^\top) \oslash (W^{(\ell)} H^{(\ell+1)} (H^{(\ell+1)})^\top) \right)$

(3) Increase $\ell$ by one.

Repeat the steps (1) to (3) until $\|H^{(\ell)} - H^{(\ell-1)}\| \leq \varepsilon$ and $\|W^{(\ell)} - W^{(\ell-1)}\| \leq \varepsilon$ (or until some other stop criterion is fulfilled). Finally, set $H = H^{(\ell)}$ and $W = W^{(\ell)}$.

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Types of NMF

- Unsupervised NMF: decompose the matrix itself

$$\min_{W,H\geq 0} D(V|WH)$$

- Supervised NMF: use pre-trained templates
- Training phase

①给定W，即给乐器的spectralgram
②给定H，即给谱

$$\min_{W_A H_A} D(V_A|W_A H_A),\ \min_{W_B H_B} D(V_B|W_B H_B)$$

很多钢琴曲　很多小提琴曲

- Testing phase

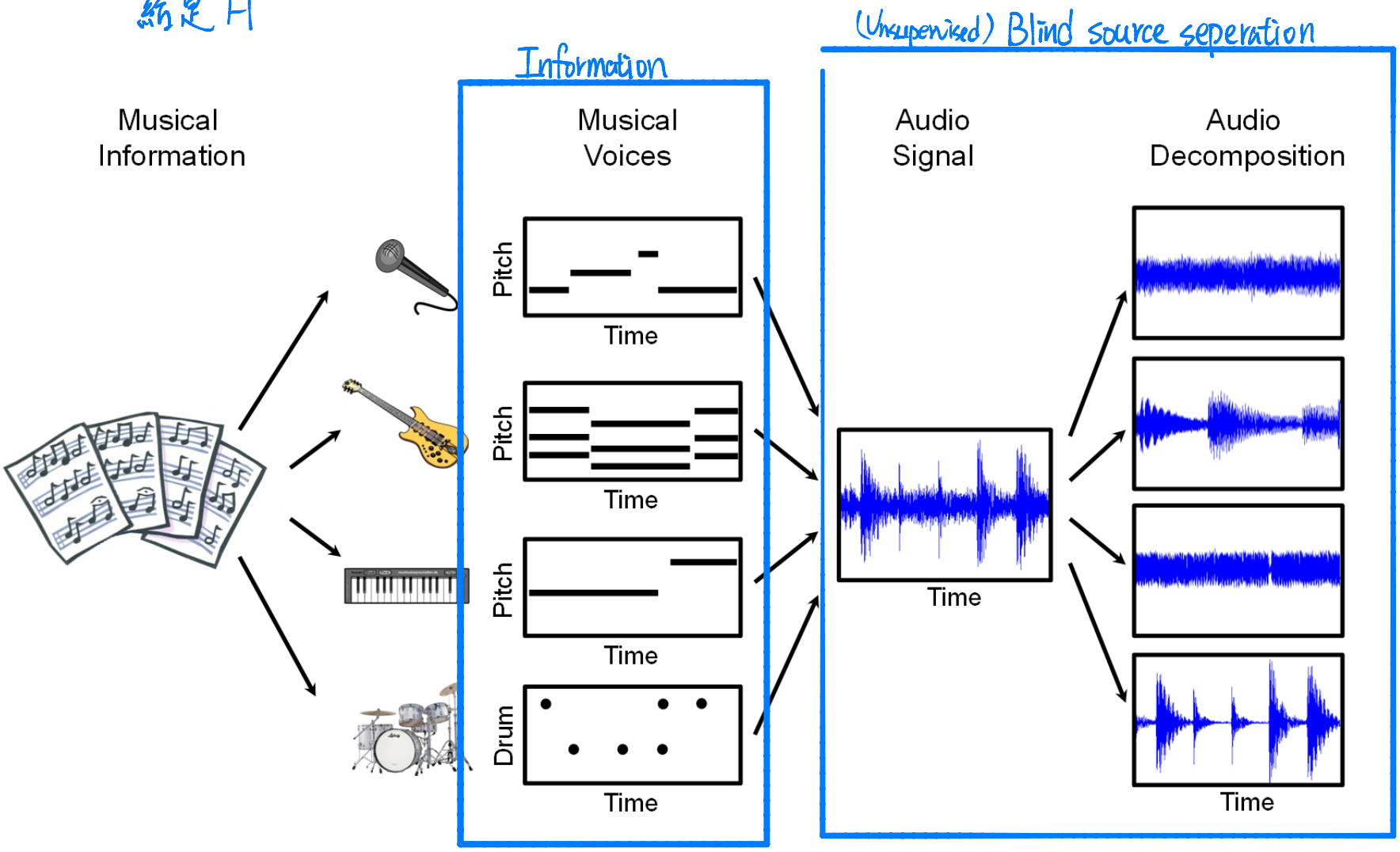
concatenate

$$\min_{\mathrm{H}} D(V_{mix}|[W_A, W_B]H)$$

可以不再 update [WA, WB]

- Informed NMF: initialize $W$ or $H$ with prior information

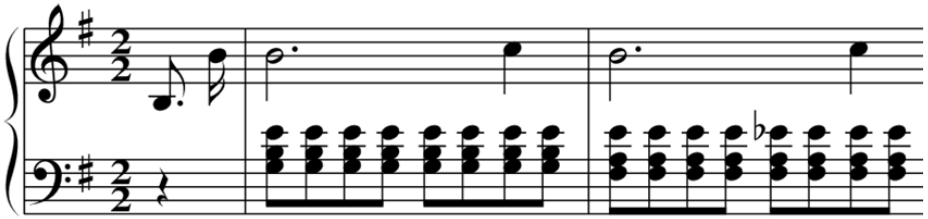
# Score-informed source separation

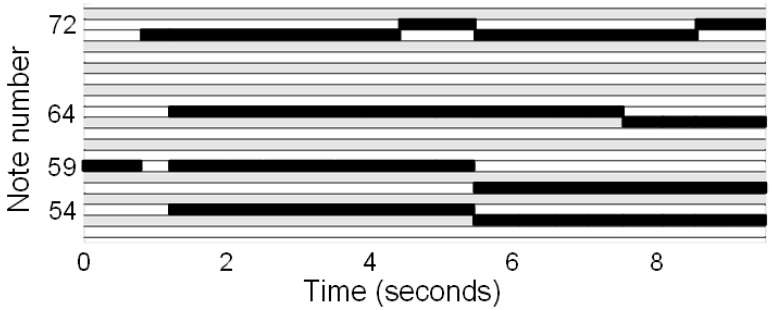


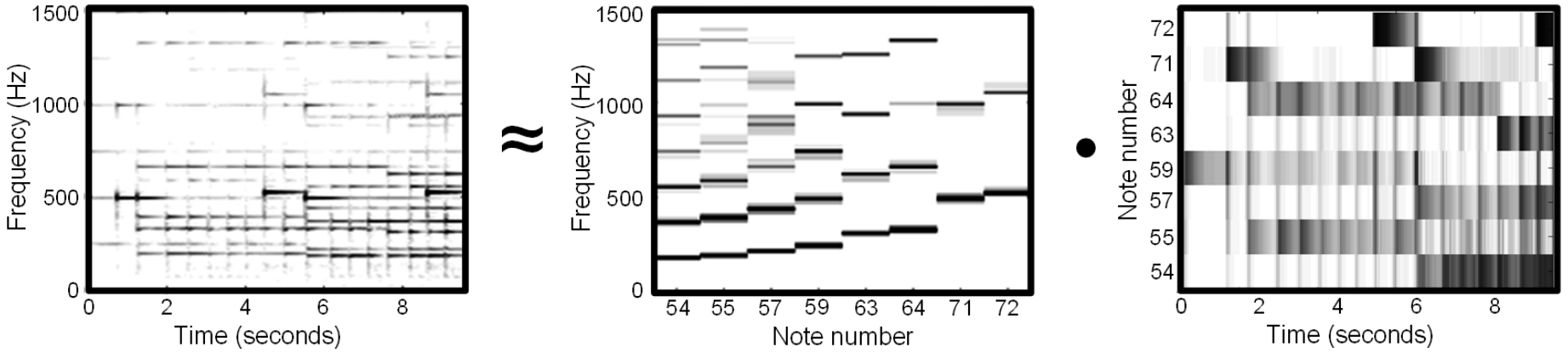

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Score-informed NMF

- The score of the music is known



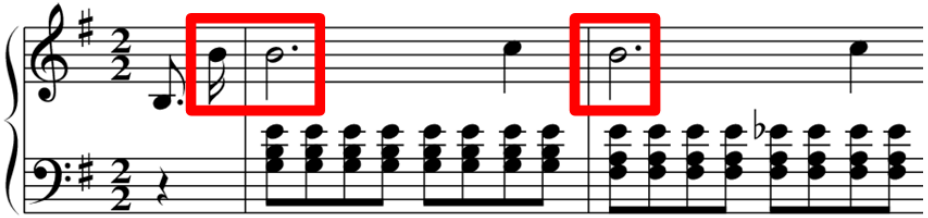


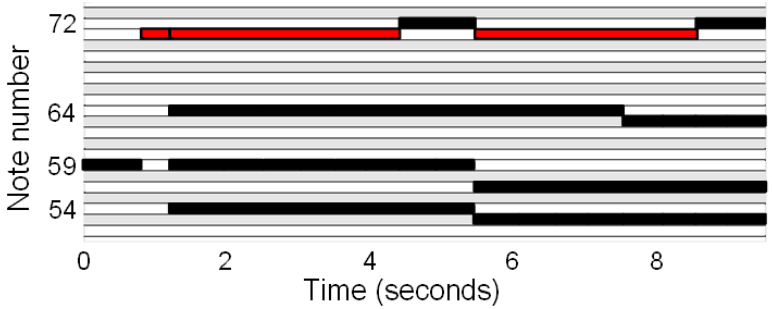



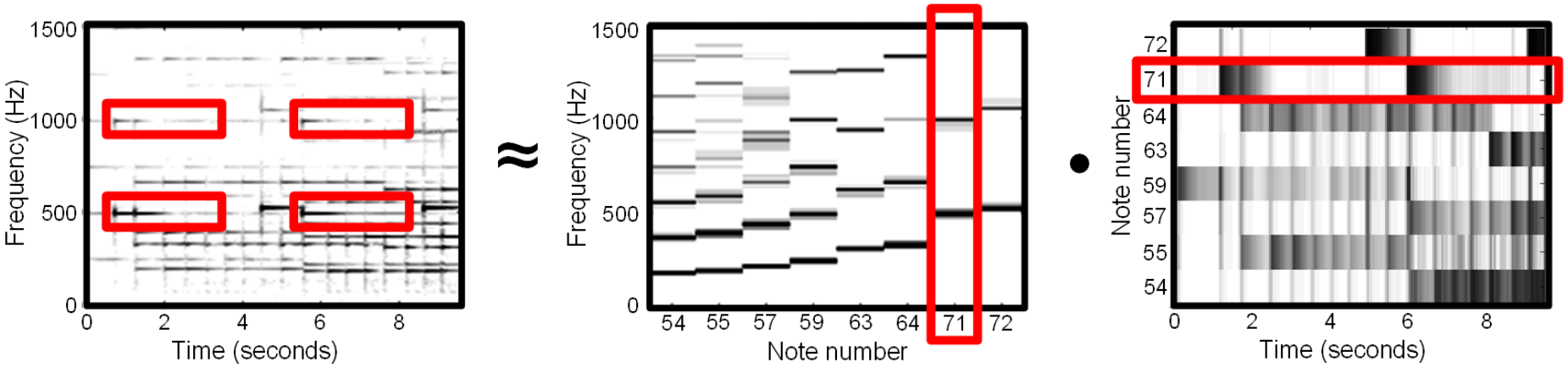

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Score-informed NMF

- The score of the music is known
- How to utilize such information?

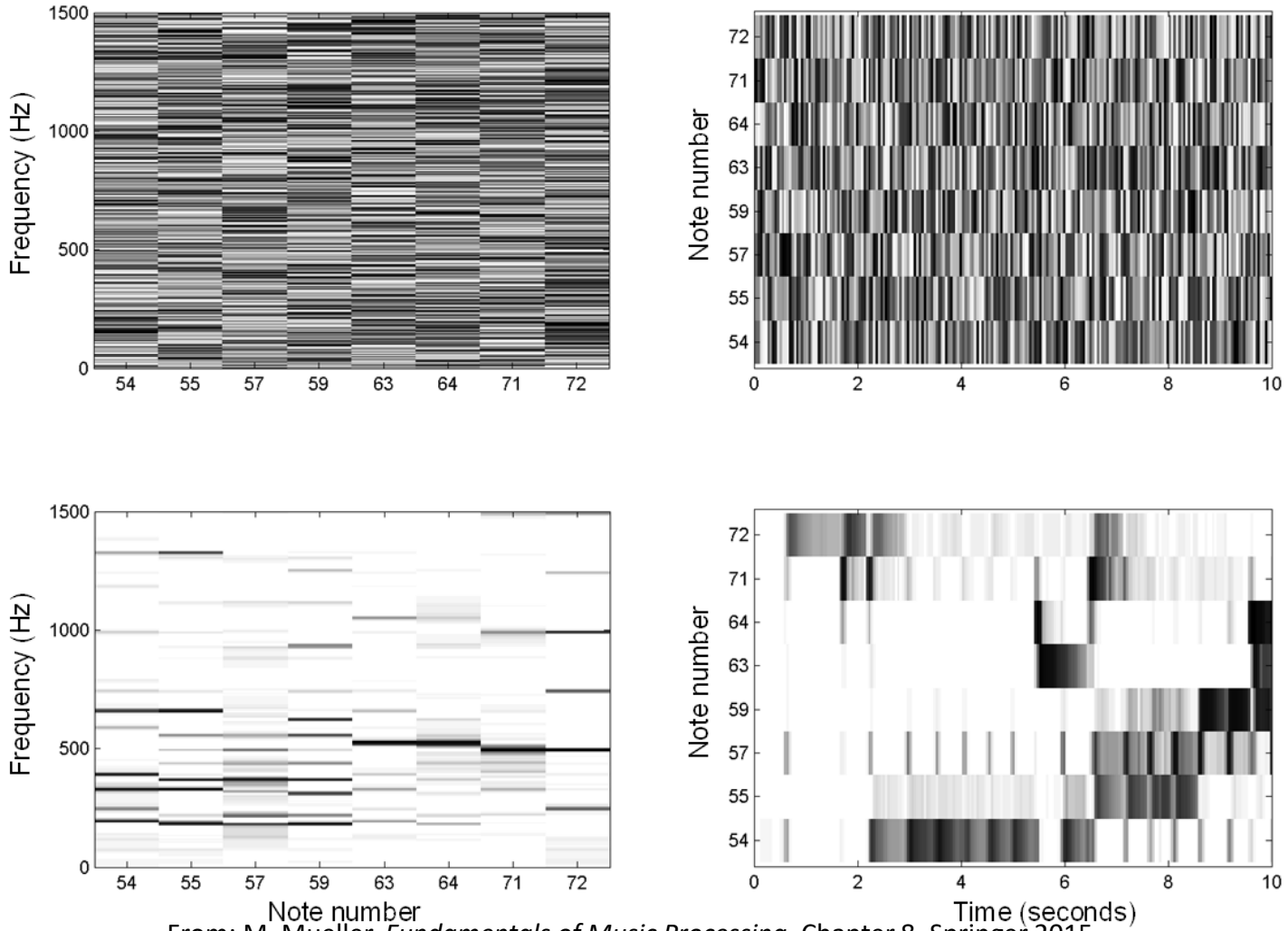




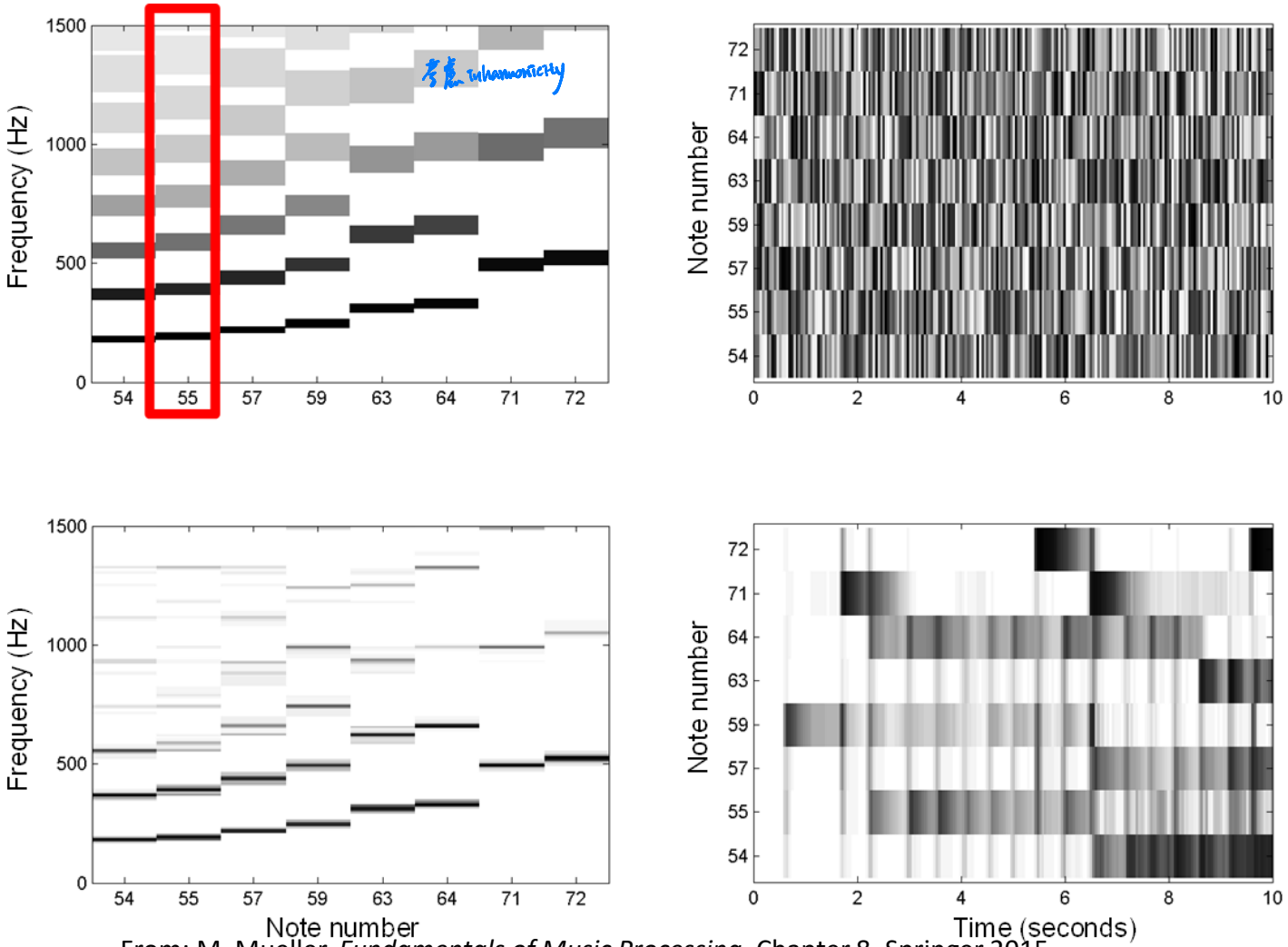

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# NMF: Random Initialization

- 

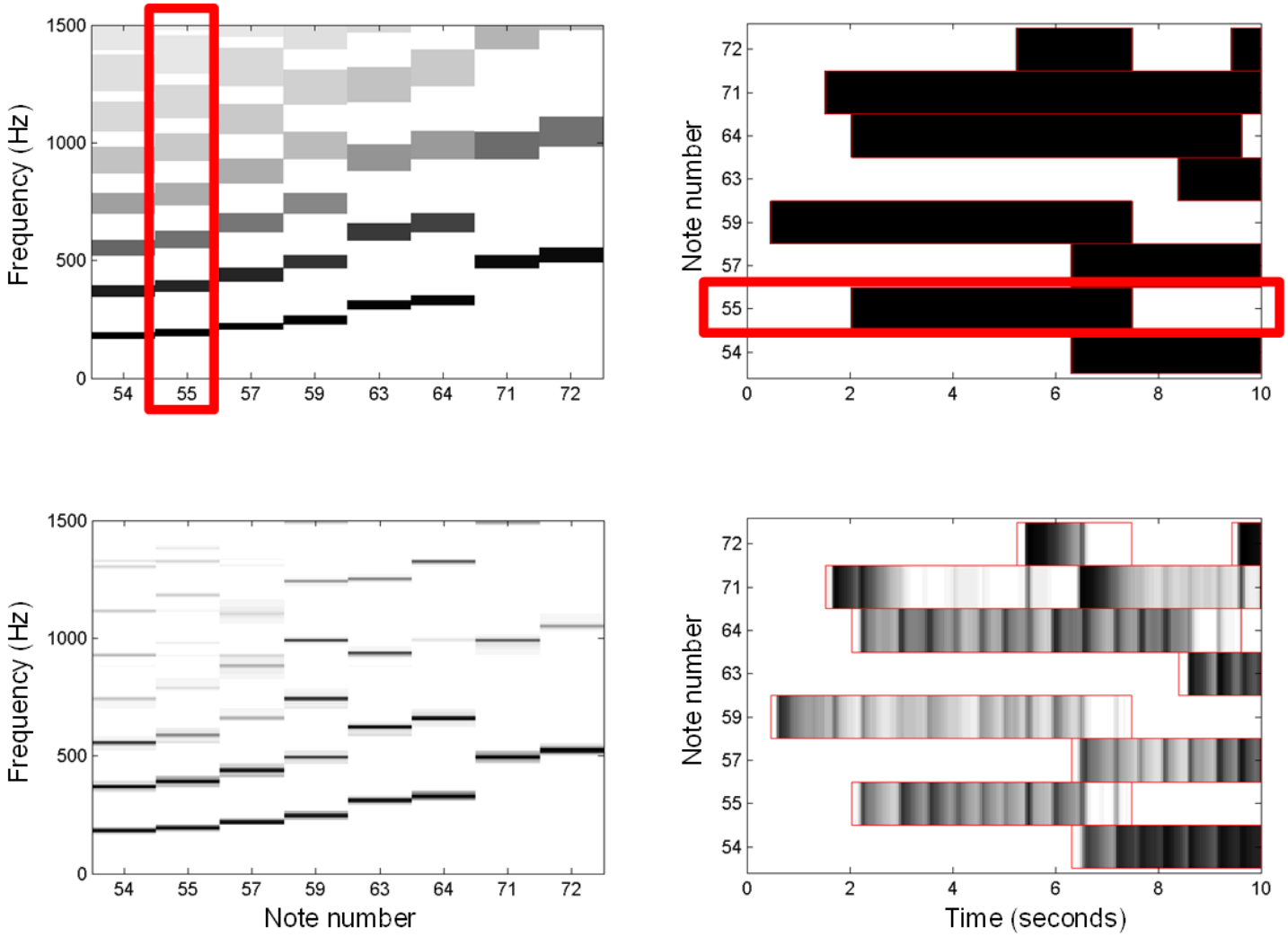

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# NMF: Harmonic Template Initialization

- 


From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

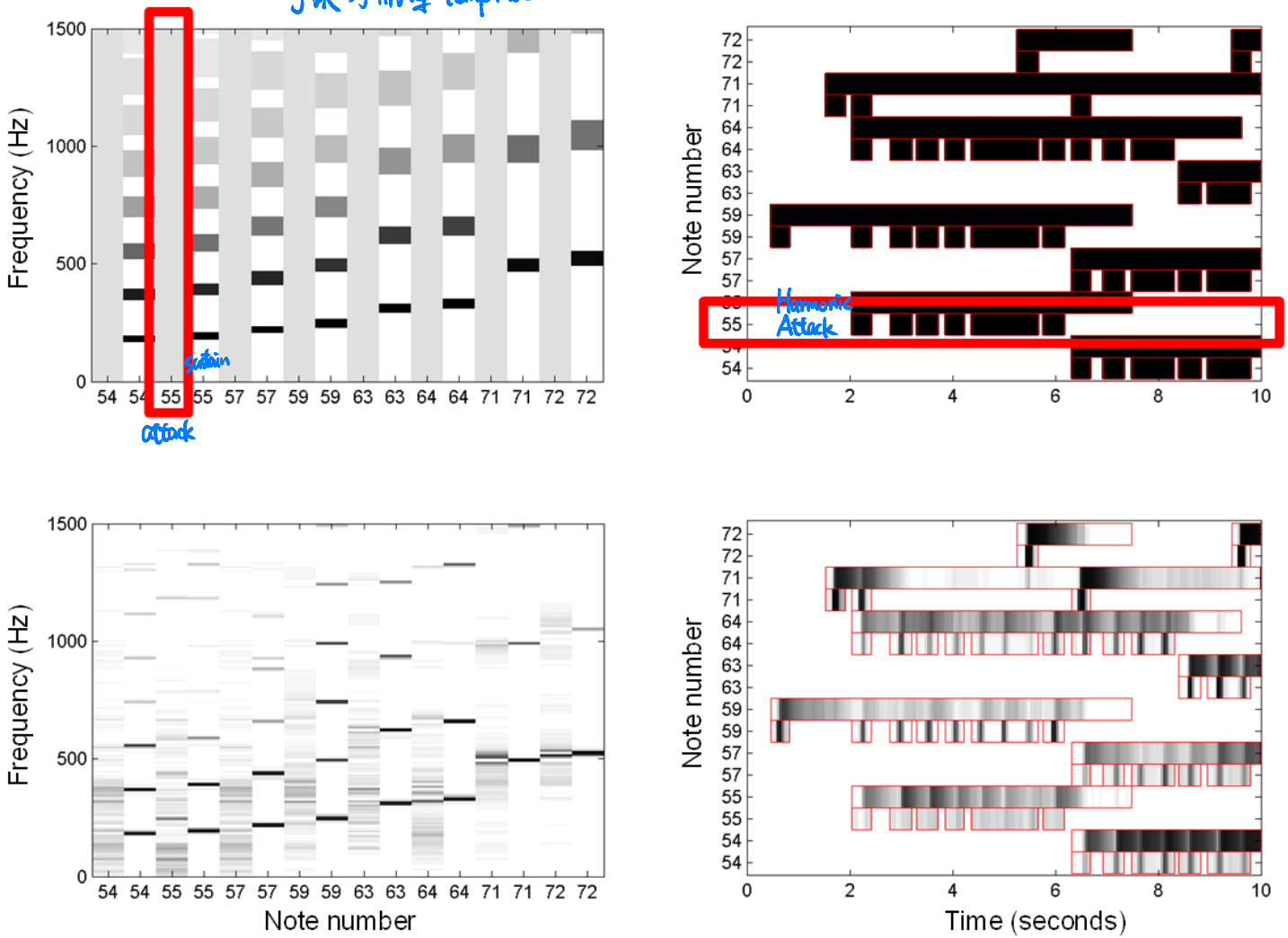
# NMF: Score-Informed Initialization

- 


From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

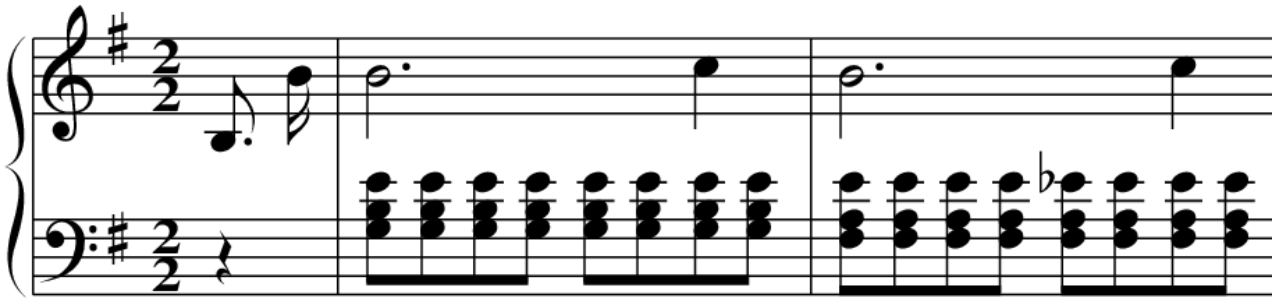
# NMF: Score-Informed Initialization + Onset

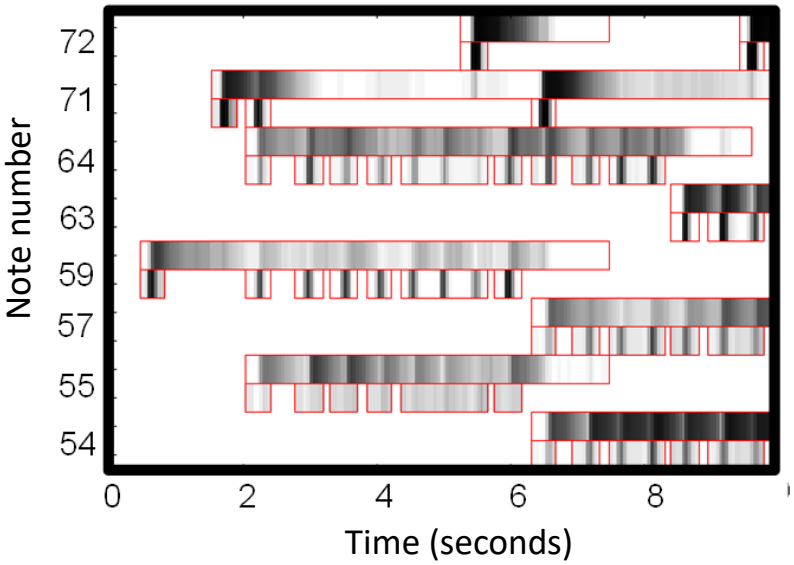
attack和sustain的频谱长不像
可以分开学template




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Score and activation matrix

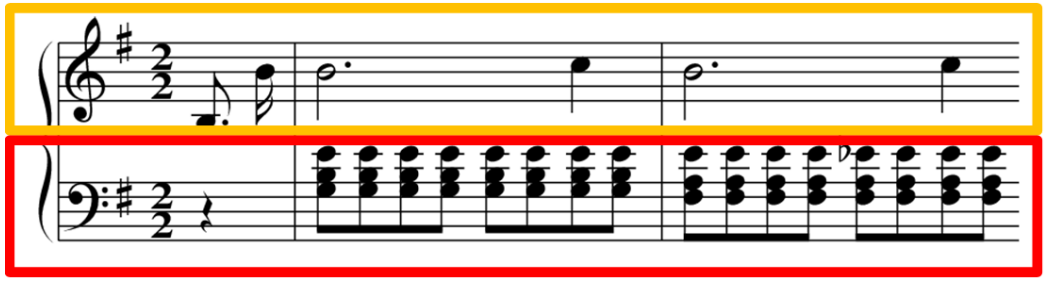
- 

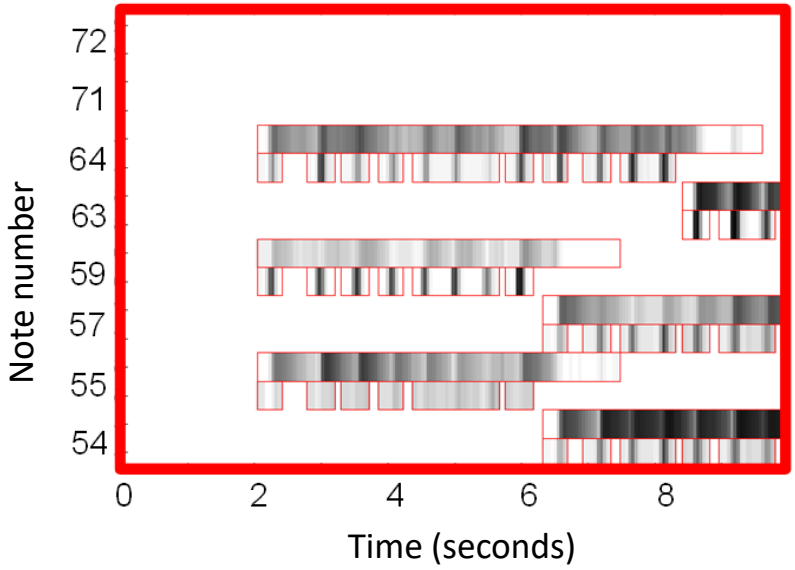


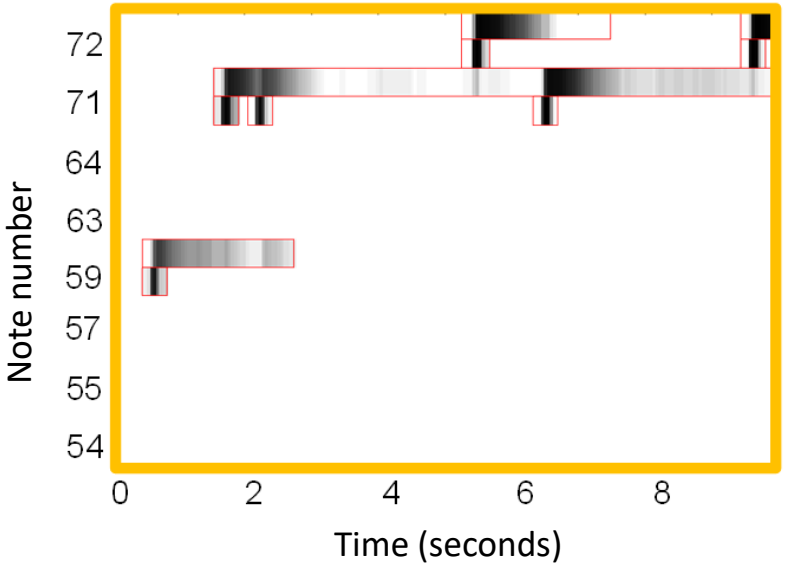

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Import more score information

- Left-hand and right-hand parts

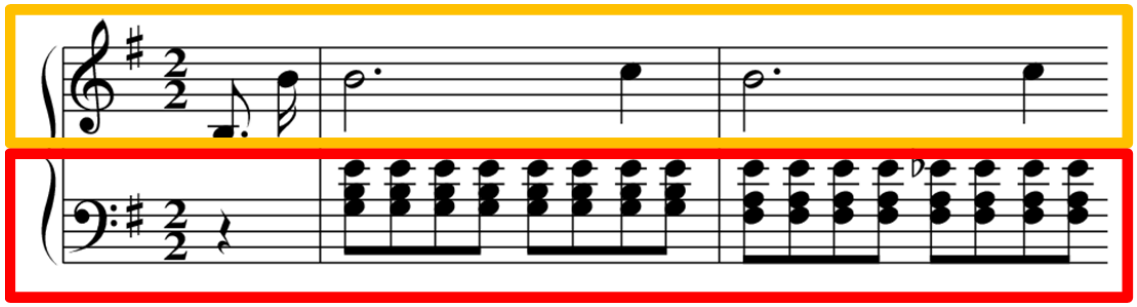


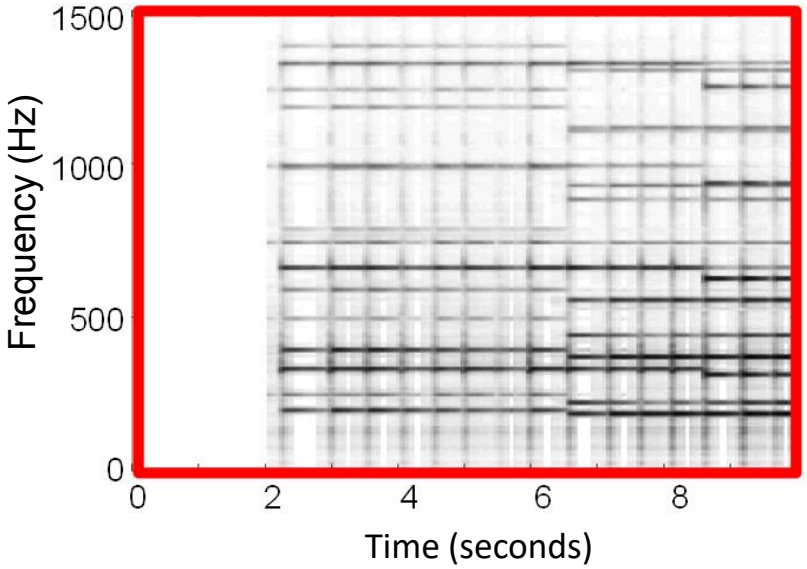


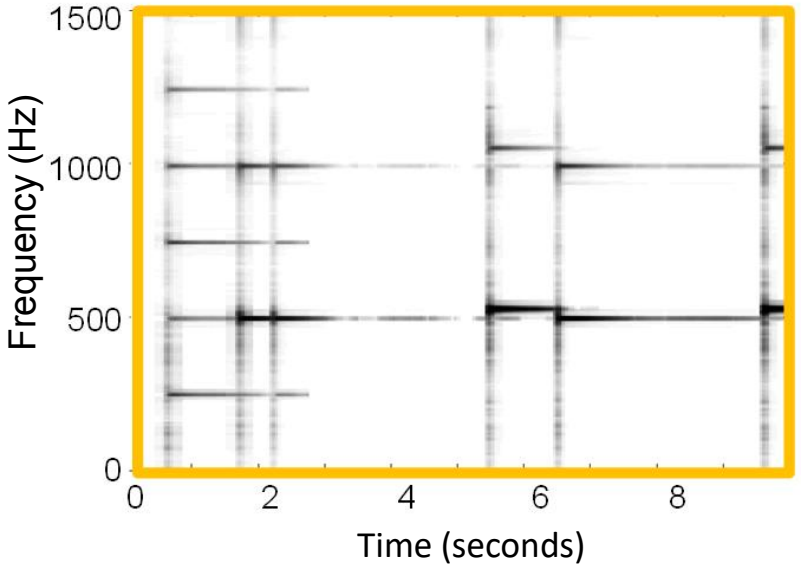




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Construct the spectrogram

- 

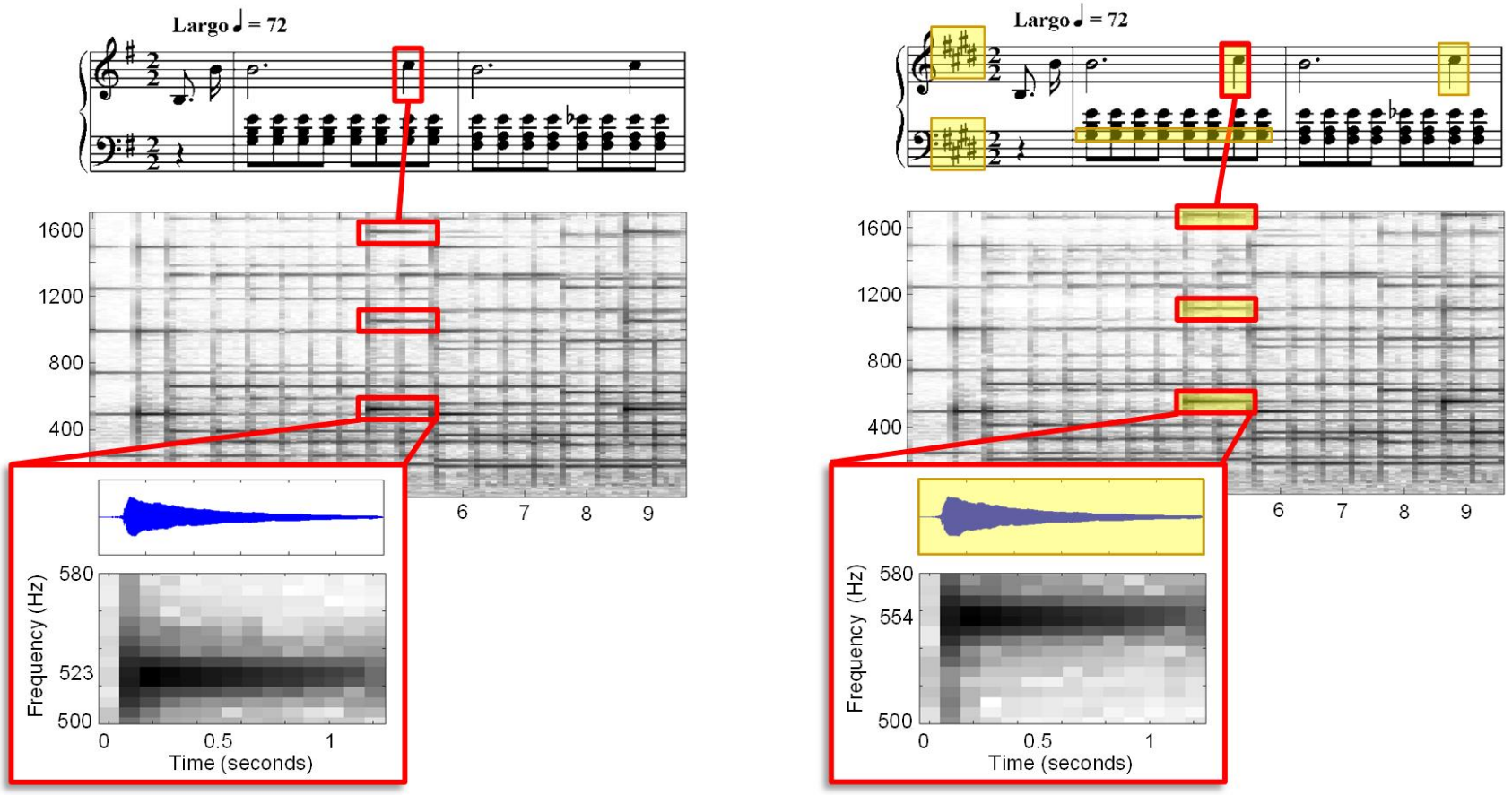



From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# With transposition

移調



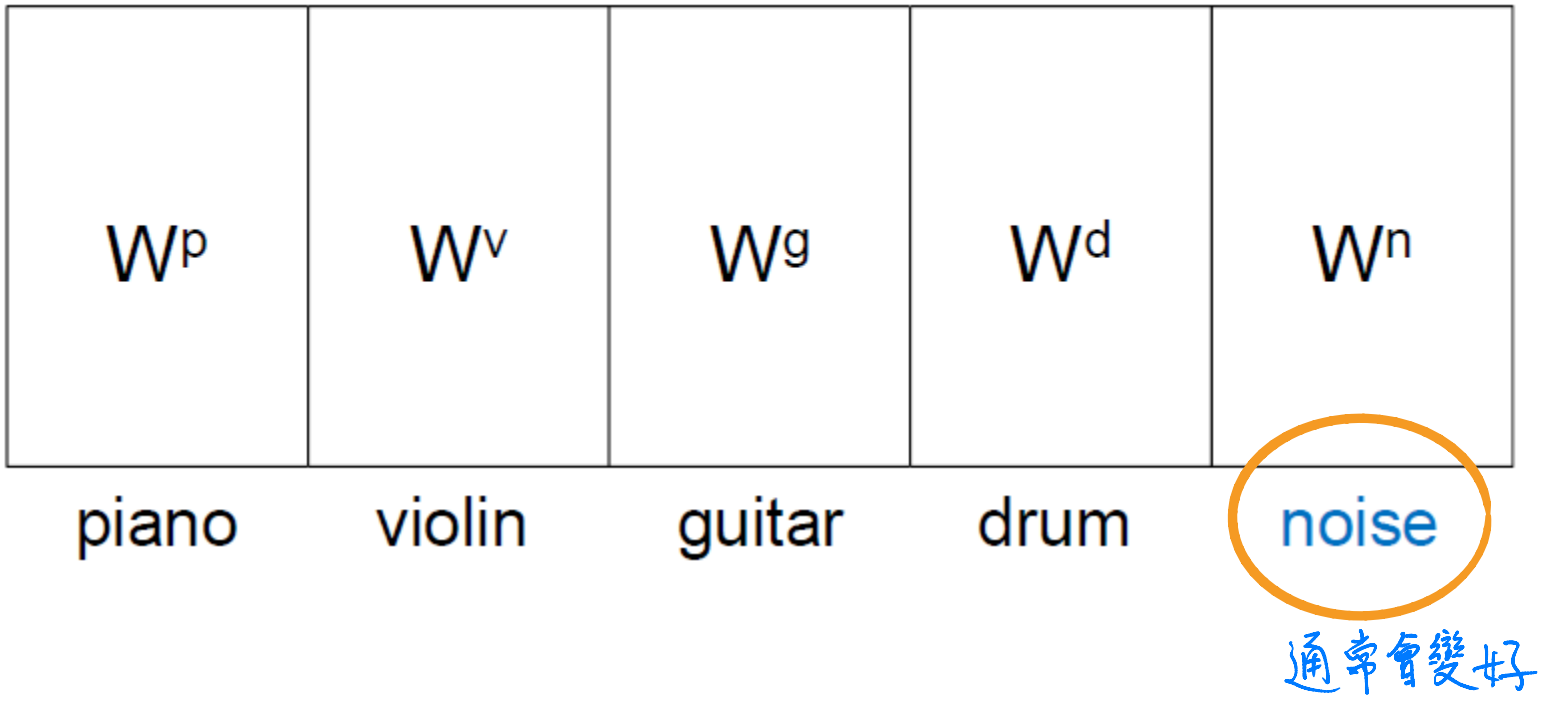

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Evaluation

- Tool: BSS toolbox in Matlab, mir_eval 也有实现
- Source-to-distortion ratio (SDR)
- Source-to-interference ratio (SIR)
- Source-to-artifact ratio (SAR)
  - true sources: **a**, **b**
  - estimated sources: **ae**, **be** (predicted)
  - SDR(a): how **ae** is similar to **a**
  - SIR(a): how **ae** is similar to **b**
  - SAR(a): how **ae** is not similar to either **a or b**
  - we can also compute SDR(b), SIR(b), SAR(b)

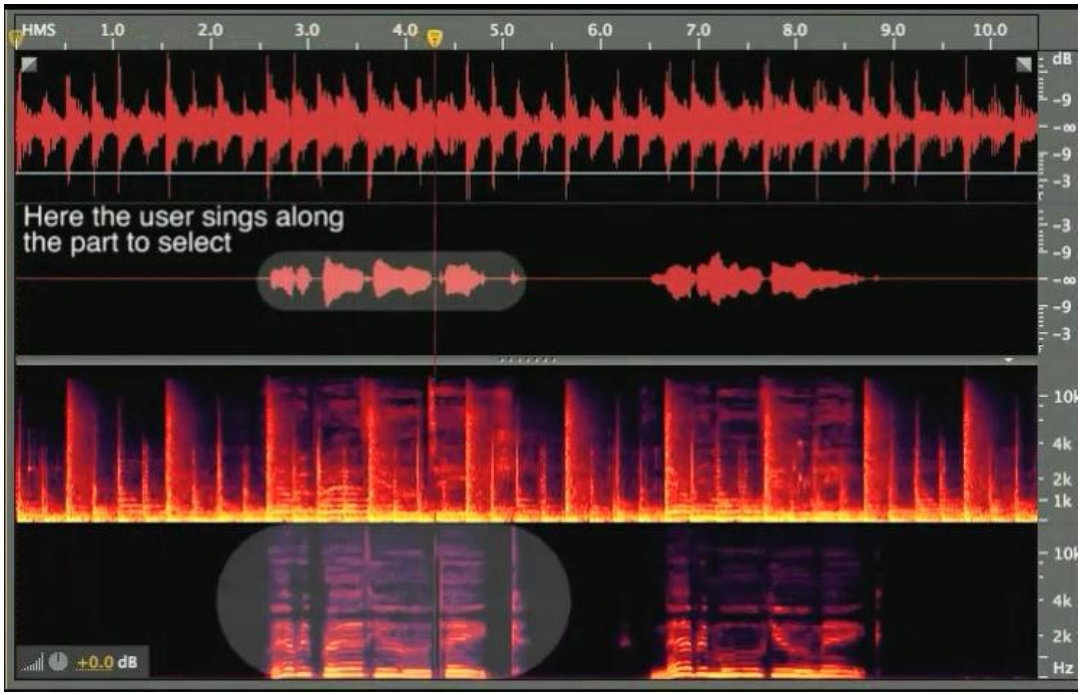
# Adding a Noise Dictionary

- To account for the possible noises in the signal



# User-guided separation

- Example: separation by humming



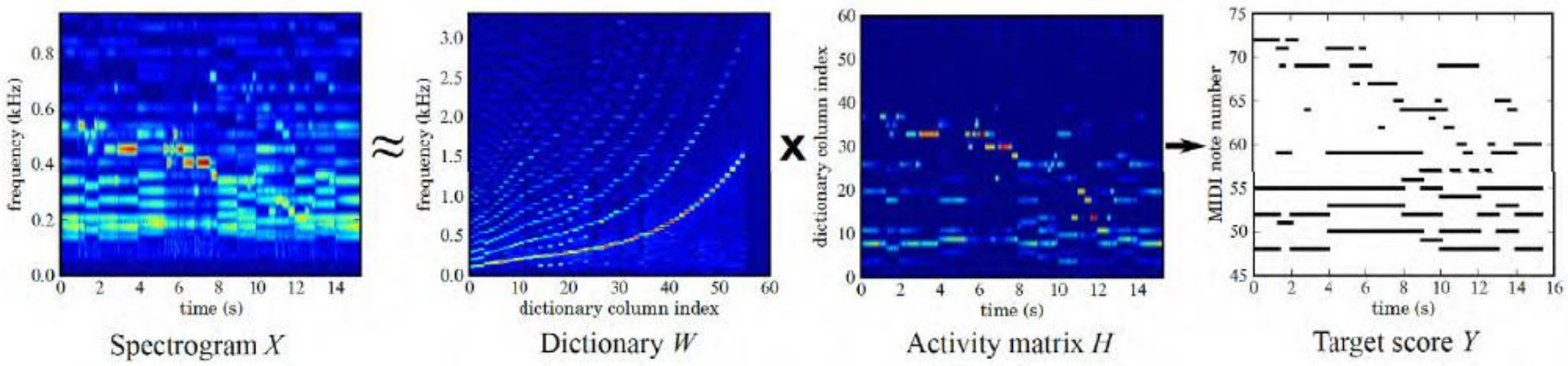

哼完即得!

Video from P. Smaragdis and G. Mysore, "Separation by Humming": User Guided Sound Extraction from Monophonic Mixtures" in Proc. WASPAA, New Paltz, NY. October 2009
http://www.cs.illinois.edu/~paris/demos/ai/user-guide.mp4 去瞧瞧

# Extension: Dictionaries for pitch estimation

Source seperation 和 transcription 有點G生蛋蛋生G問題

- Decompose the input as a linear combination of individual components
  - templates of instruments => source separation
  - templates of notes => multi-pitch estimation
  - templates of chords => chord recognition




Discriminative non-negative matrix factorization for multiple pitch estimation, ISMIR 2012

# Extension: Voice Conversion

~2014 NMF仍是state-of-the-art

用violin的W得H，用长笛的W乘H

條件：training 想得 $W_{violin}$ 和 $W_{flute}$ 最好有 parallel data（要align好） Deep learning 中常用 Seq2seq 或 transformer

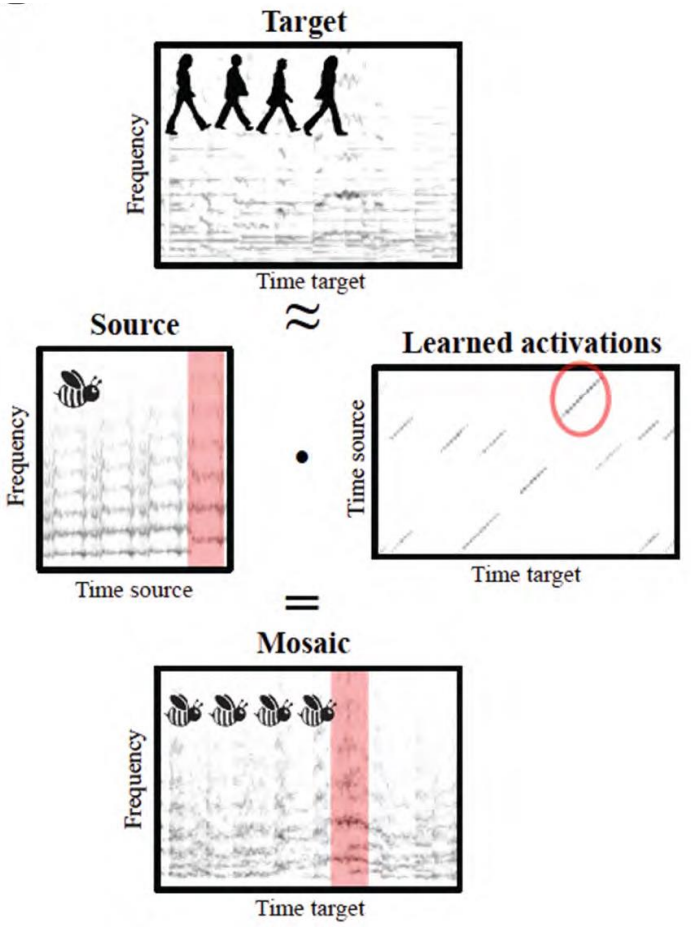
## Quiz

- we have two dictionaries $\mathbf{X}_{\text{vio}} \in \mathbb{R}^{n \times m}$ and $\mathbf{X}_{\text{flu}} \in \mathbb{R}^{n \times m}$ for violin and flute, satisfying
  - each dictionary contains spectral templates of different pitches;
  - the two dictionaries have one-to-one correspondence (i.e. $\mathbf{X}_{\text{vio}}^{(j)}$ and $\mathbf{X}_{\text{flu}}^{(j)}$ correspond to the same pitch, $\forall j$);

  for a violin recording $\mathbf{Y}_*$, we compute $\mathbf{W}_*$ s.t. $\mathbf{Y}_* \simeq \mathbf{X}_{\text{vio}}\mathbf{W}_*$; then, what would happen if we take $\mathbf{X}_{\text{flu}}\mathbf{W}_*$?

# Extension: Audio Mosaicing

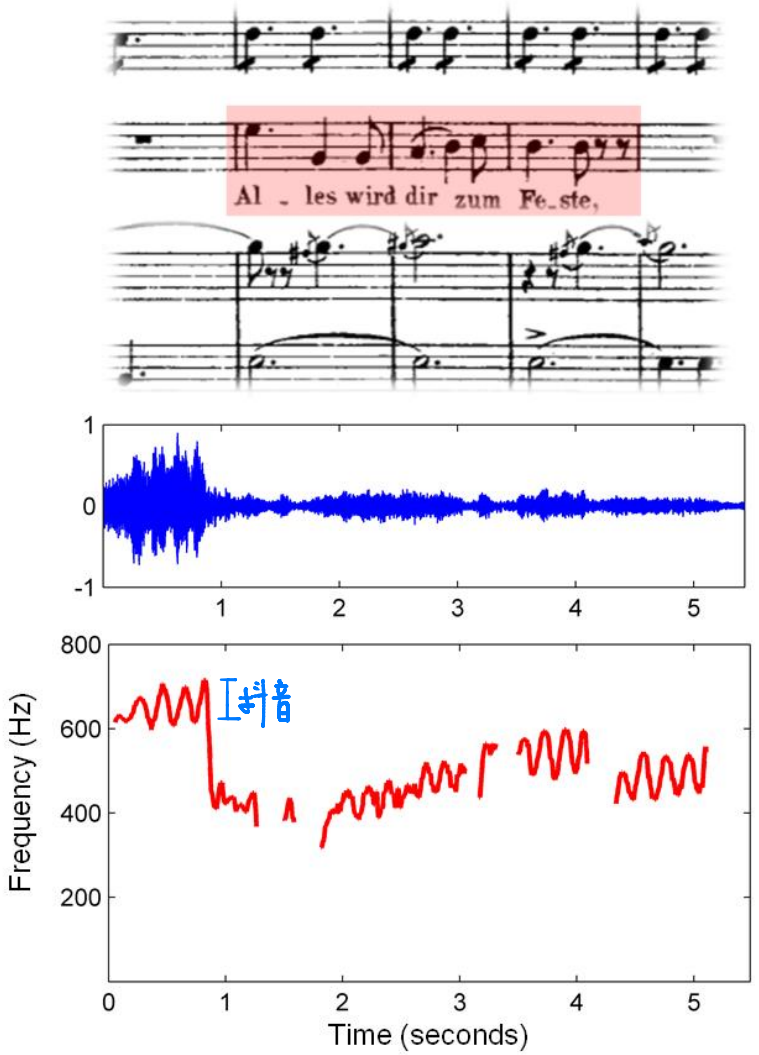
- Given a *target* and a *source* recording, the goal of *audio mosaicing* is to generate a mosaic recording that conveys musical aspects (like melody and rhythm) of the target, using sound components taken from the source
- https://www.audiolabserlangen.de/resources/MIR/2015-ISMIR-LetItBee/ 去瞧瞧



# Melody extraction

這討論人聲

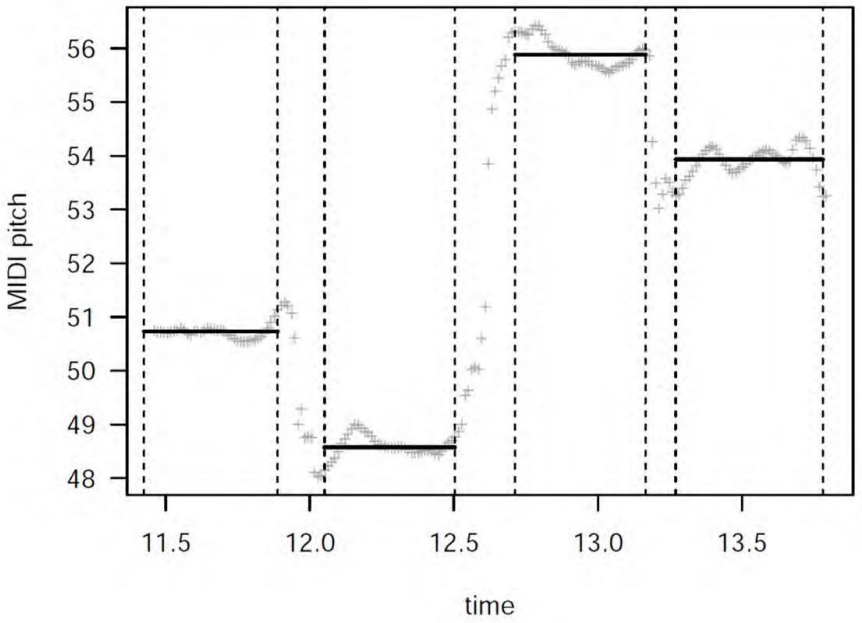
- Knowing the melody contour is helpful for source separation




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Melody extraction

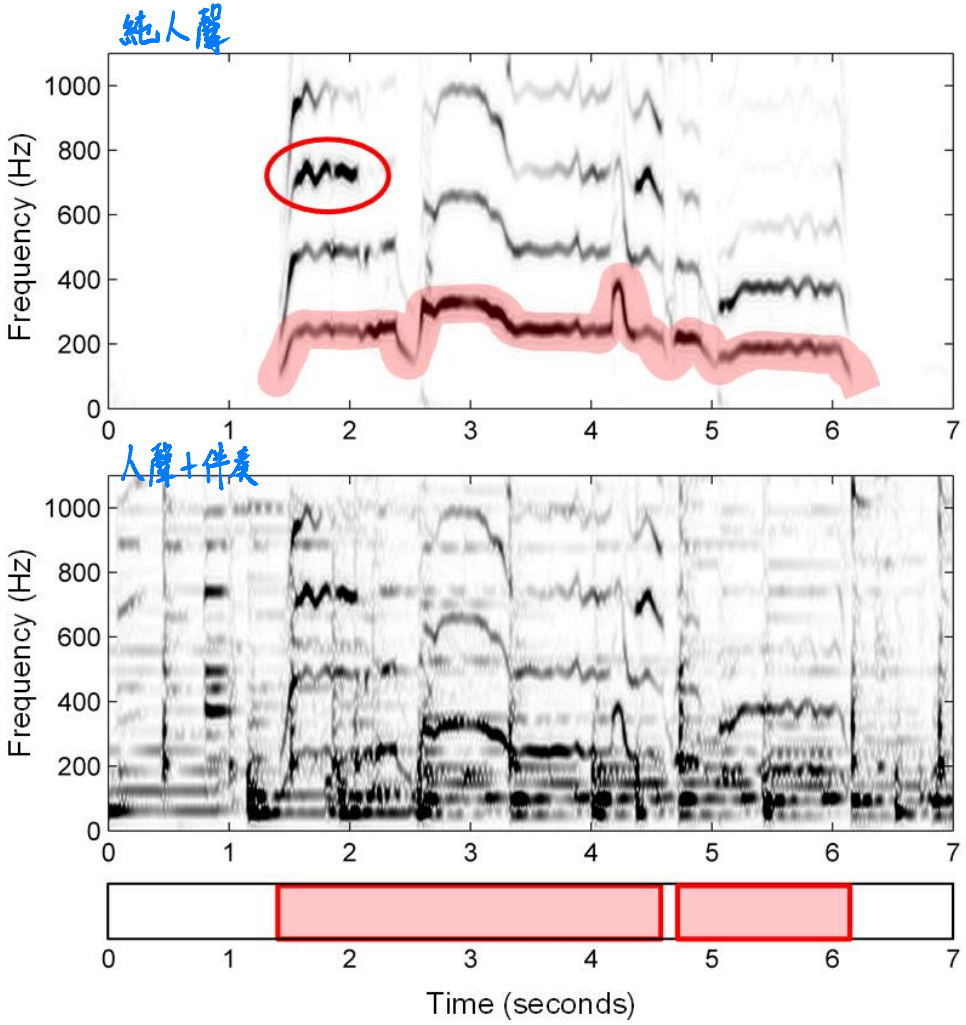
- Definition: linear succession of musical tones expressing a particular musical idea
- Frequency trajectory, instead of a sequence of notes to account for the patterns such as *vibrato*, *tremolo*, or *glissando*
- Pitch contour extraction
  - Pitch
- Note segmentation
  - Pitch, onset, offset
- Voice activity detection



給人聲+伴奏欲得人聲的旋律

# Melody Extraction: Challenges

- Attribute specific t-f patterns to notes of individual instruments
- Resonance and reverberations increase the overlap of different sound sources
- Determine which of the F0-values belong to the predominant melody
- Vocal activity detection



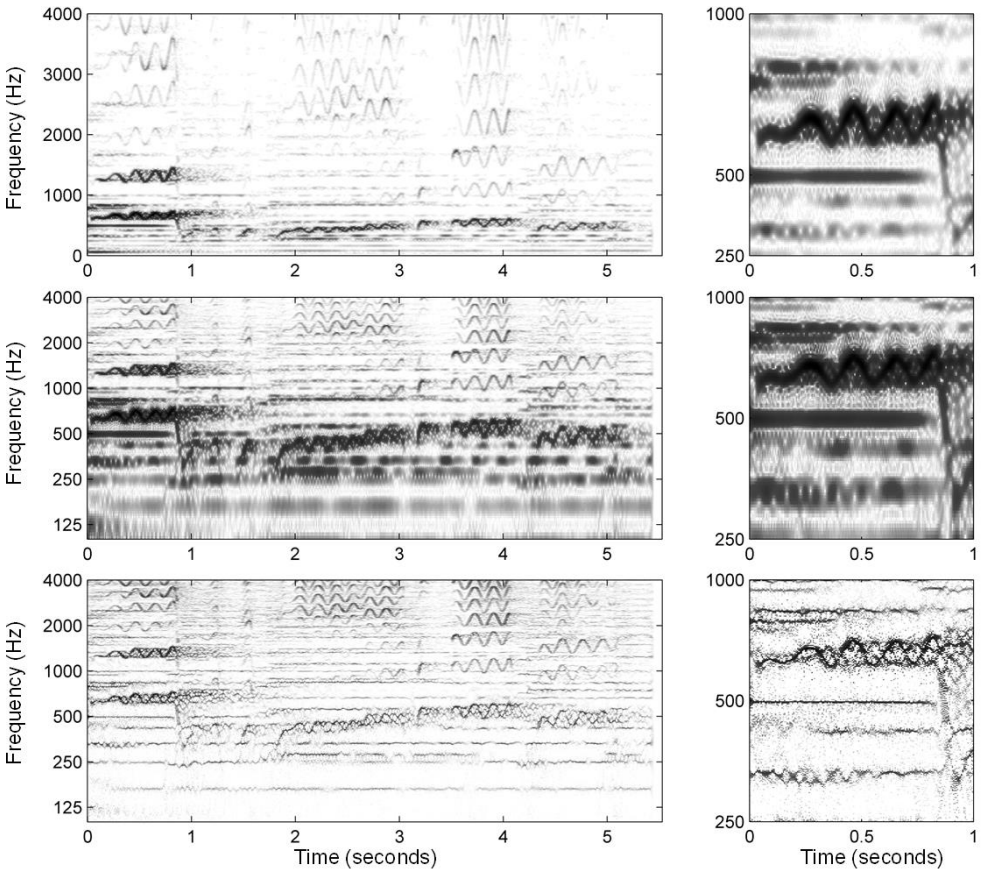

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Log-frequency representation

男聲 $f_0 \approx 100$ Hz 且常常有 missing fundemental，人聽需要低頻的 resolution 細緻

mel-frequency

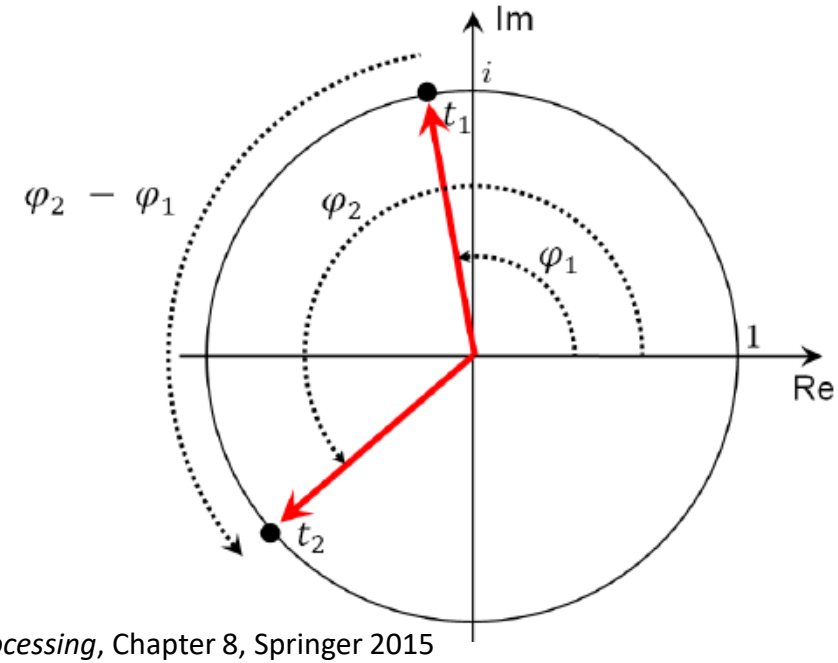
- Log-freq spectrogram
  - log perception of frequency and the musical notation of note's pitch
  - melody can be estimated by amplifying regularly spaced frequency components (*scaled* version vs *translated* version)
- Instantaneous frequency estimation 似 super resolution
  - refine the frequency grid




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Instantaneous Frequency Estimation

- How to achieve “super-resolution” in the time-frequency grid?
- Instantaneous frequency: the derivative of phase 相位角的微分
  - http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/ifgram.m
  - http://tftb.nongnu.org/




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

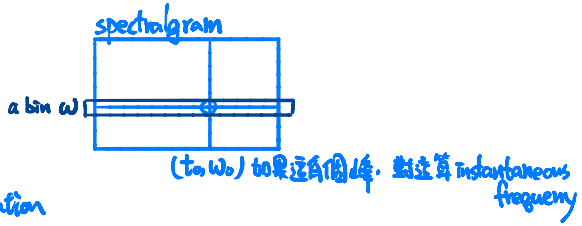
# Instantaneous Frequency Estimation

Φ: phase  ω: freq

- Consider the short-time Fourier transform of $x(t)$ with window function $h(t)$

$$S_x^h(t,\omega) \coloneqq \int x(\tau)h(\tau - t)e^{-j\omega\tau}d\tau = M_x^h(t,\omega)e^{j\Phi_x^h(t,\omega)}$$

$$\log S_x^h = \log M_x^h + j\Phi_x^h$$



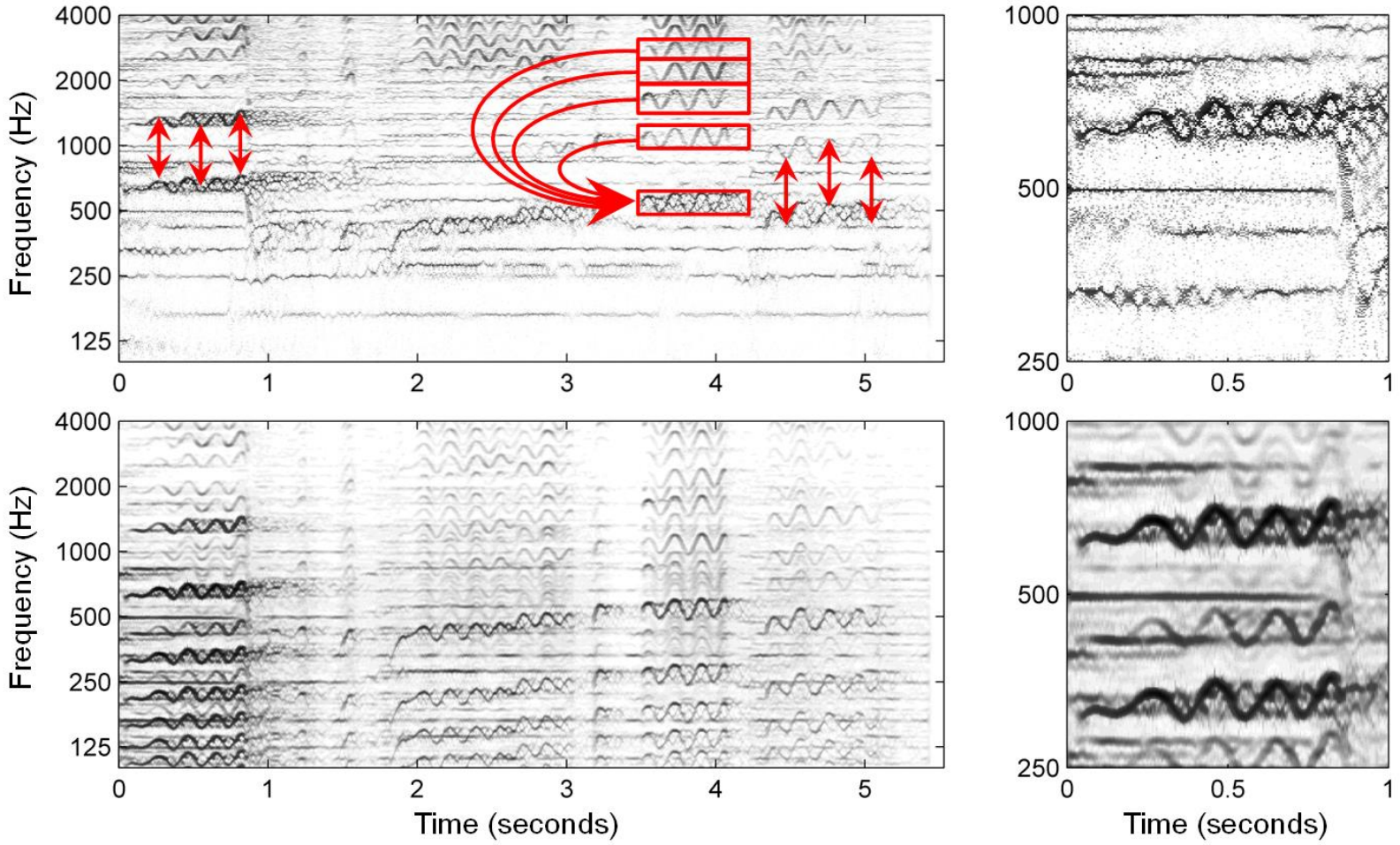

- Instantaneous frequency 對1個時頻點做，可以修正 ω 的 resolution 帶來的誤差

$$\widehat{\omega}(t,\omega) \coloneqq \omega + \frac{\partial\Phi_x^h(t,\omega)}{\partial t}, \quad \frac{\partial\Phi_x^h(t,\omega)}{\partial t} = \underset{\text{虛部}}{Im}\left(\frac{S_x^{Dh}(t,\omega)}{S_x^h(t,\omega)}\right)$$

window function 的微分 → 用 h' 做 STFT
→ 用 h 做 STFT

- Where $Dh \coloneqq \partial h(t)/\partial t$
- For an $\omega$ at a spectral peak (resolution up to a bin), you may use the phase derivative to achieve super-resolution

# Harmonic Summation



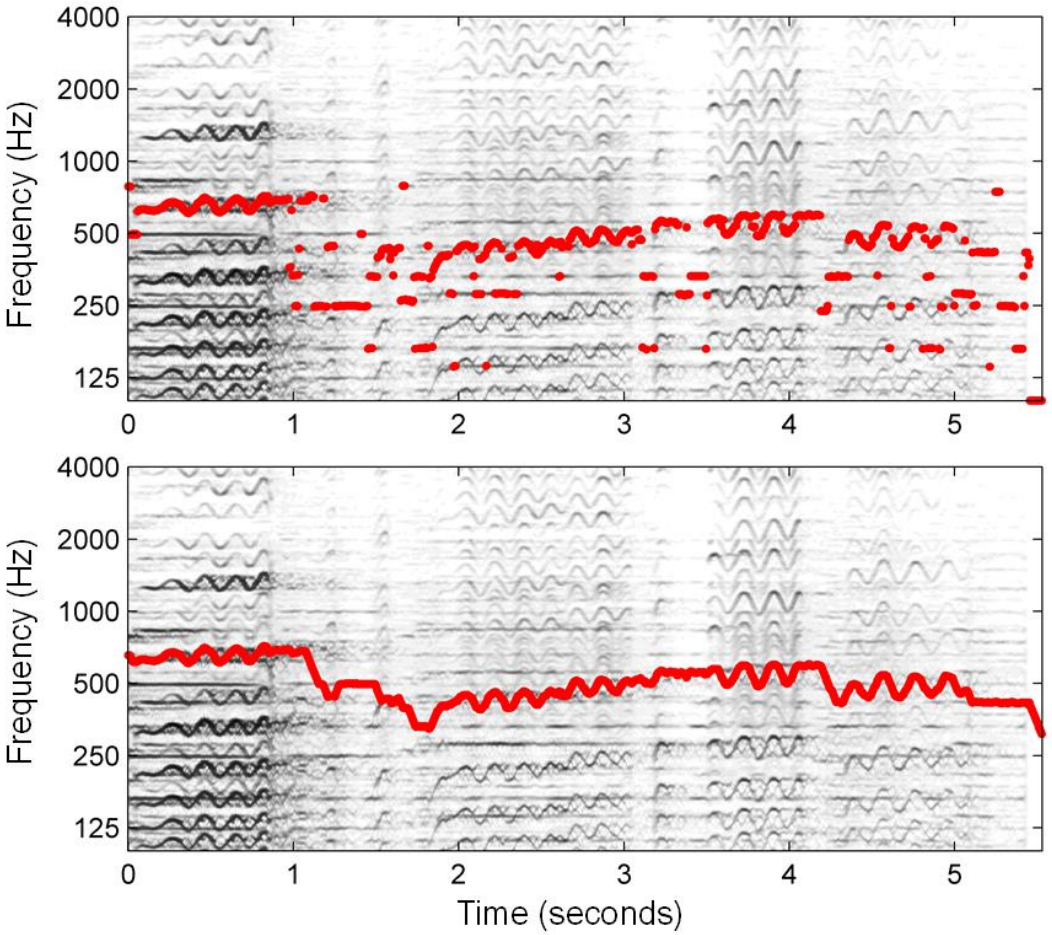

From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Melody Extraction

- Steps

1. Log-freq
2. Harmonic sum
3. Max per frame
4. Smoothing (HMM or median filtering)

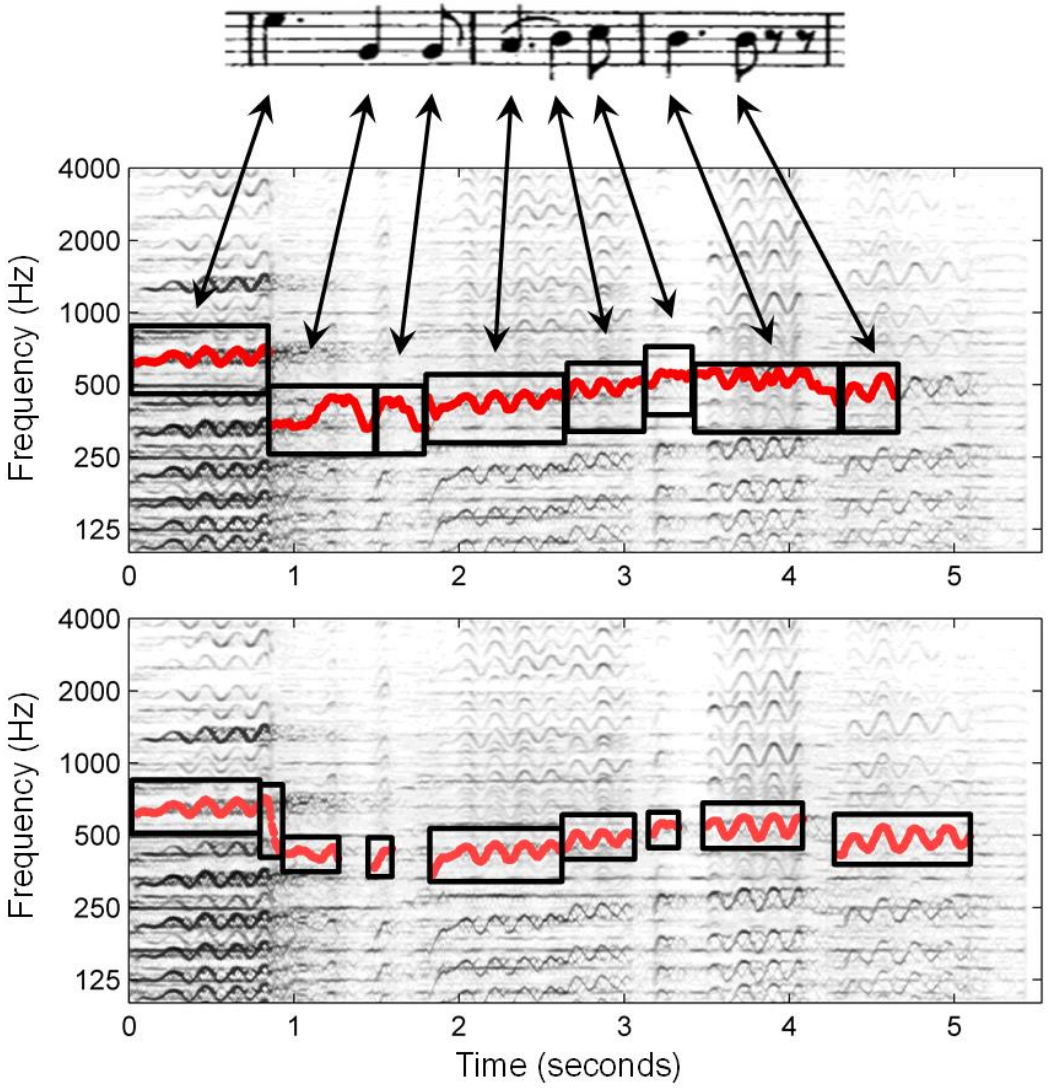
- *Temporal flexibility* (possible jumps) vs *temporal continuity* (smoothness)




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Melody Extraction

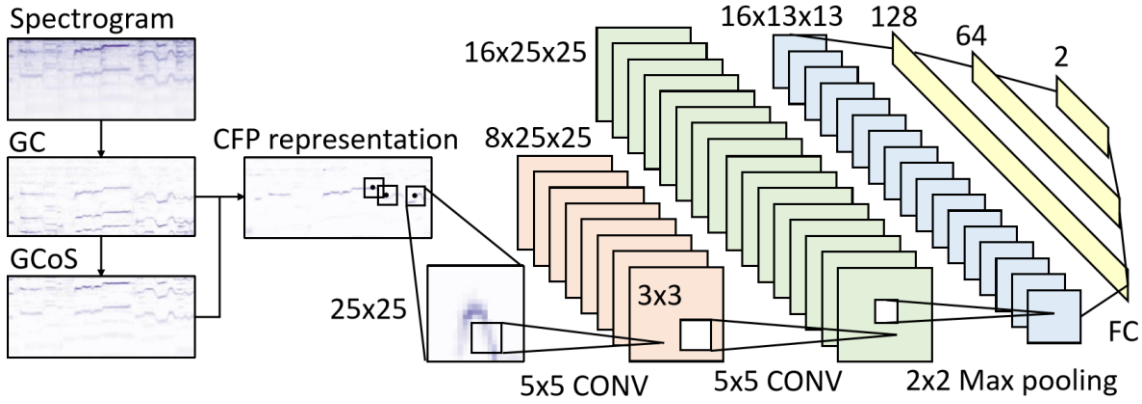
- Score-informed melody extraction
- User-guided melody extraction




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

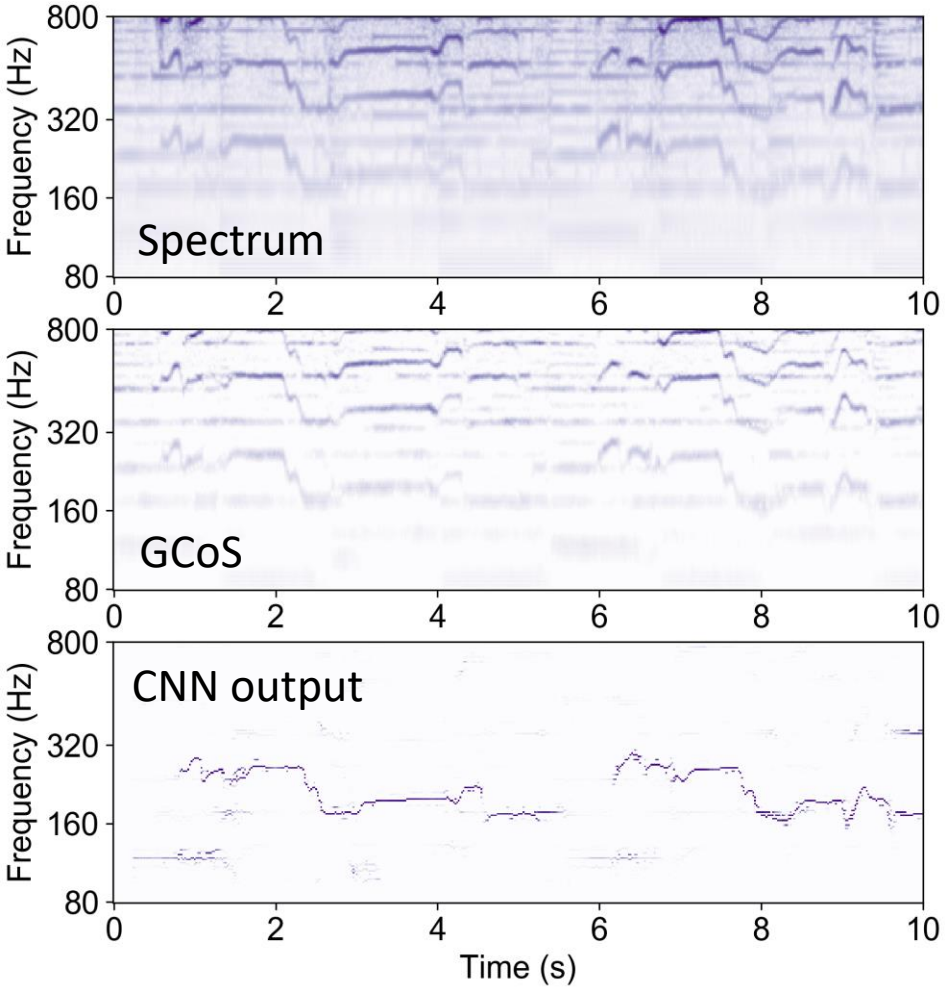
# Vocal melody extraction – patch-based CNN

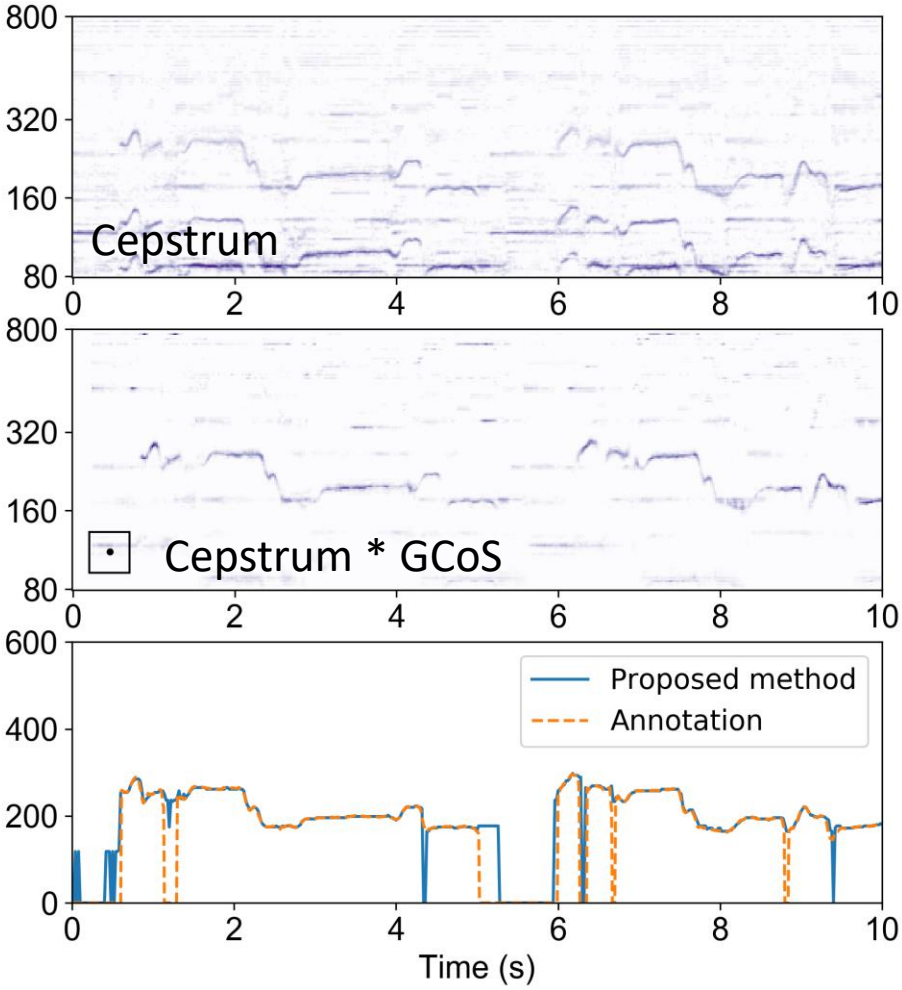
- Pitch contour of human voice
  - Special characteristics of vibrato, tremolo, sliding
  - Evan different from most string instruments
- Discriminate human voice only from pitch contour
  - Conventional method: vocal activity detection (VAD) + predominant pitch detection
  - Proposed method: cepstrum + deep learning



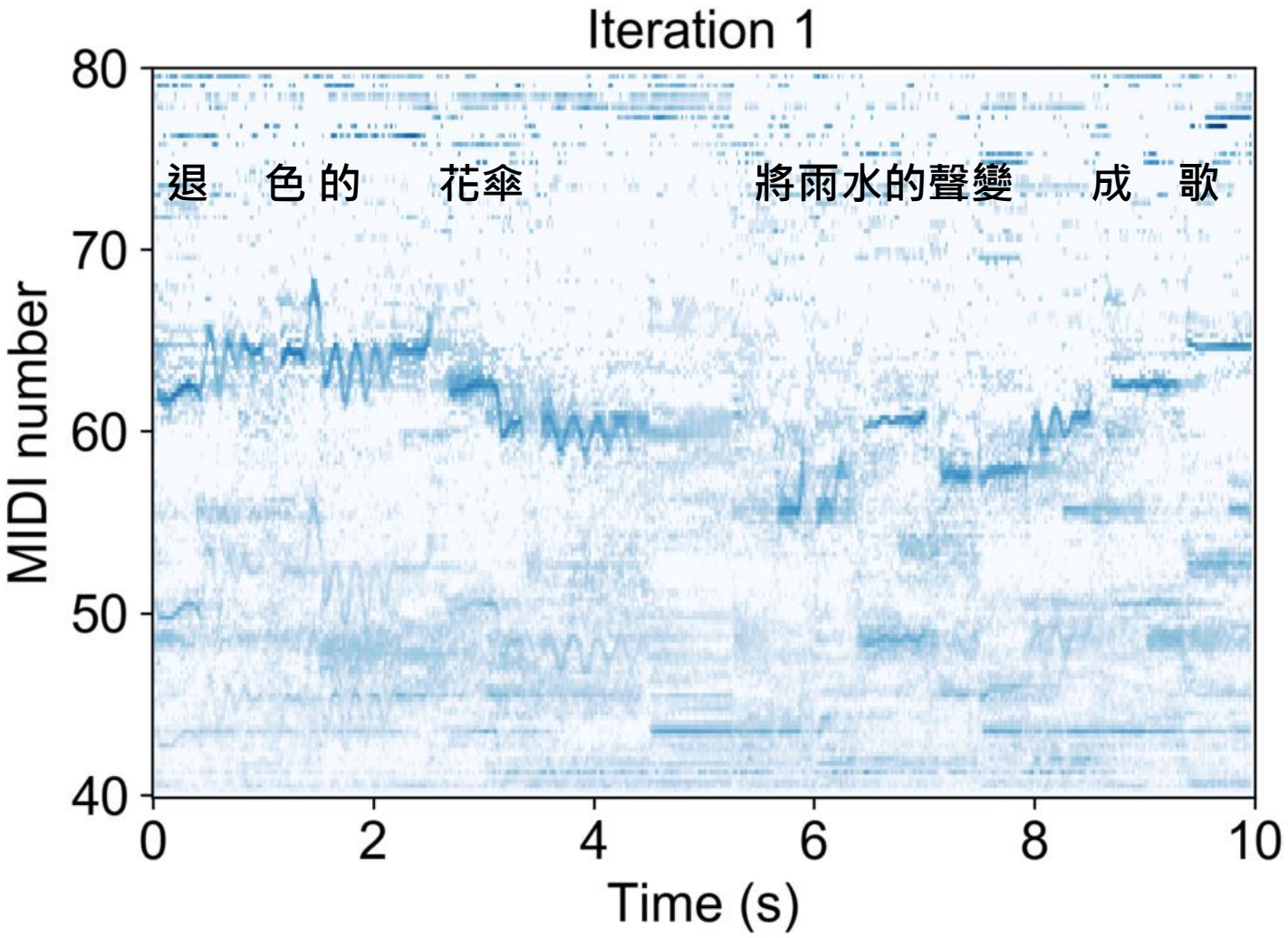

Li Su, "Vocal melody extraction using patch-based CNN," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

# Example

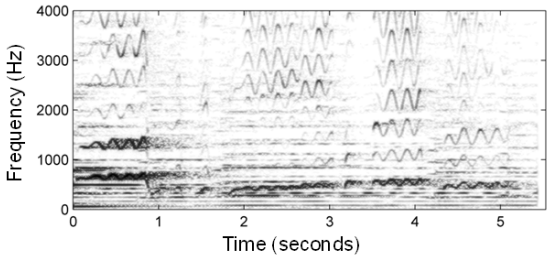

Spectrum
GCoS
CNN output
Frequency (Hz)
800
320
160
80
0
2
4
6
8
10
Time (s)

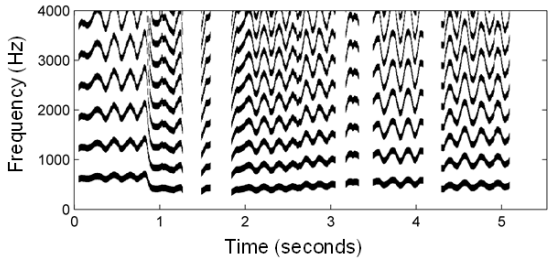

Cepstrum
Cepstrum * GCoS
Proposed method
Annotation
800
320
160
80
600
400
200
0
0
2
4
6
8
10
Time (s)

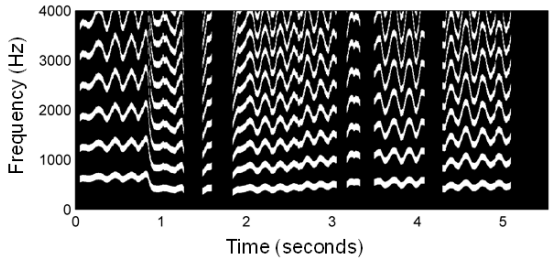
# Example

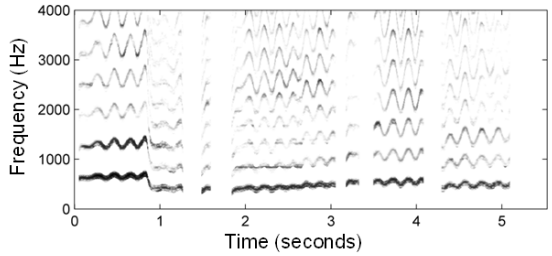

Iteration 1
退 色 的 花傘 將雨水的聲變 成 歌
MIDI number
80
70
60
50
40
0
2
4
6
8
10
Time (s)

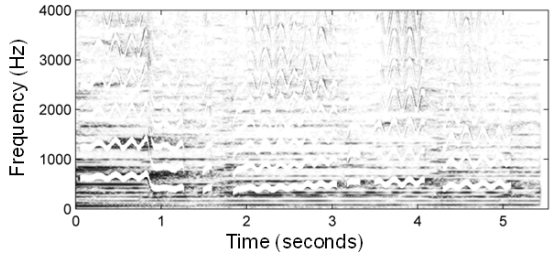
# Melody-informed separation



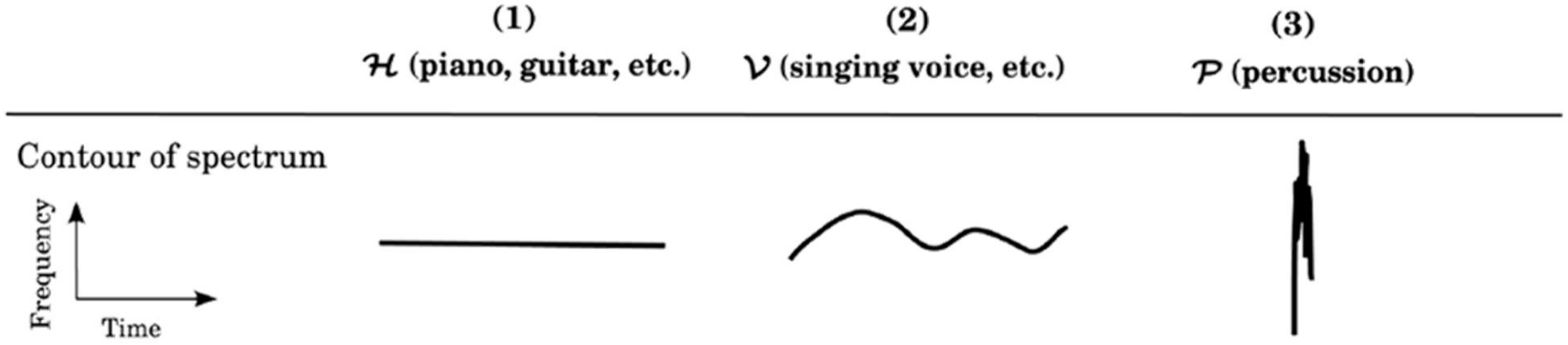



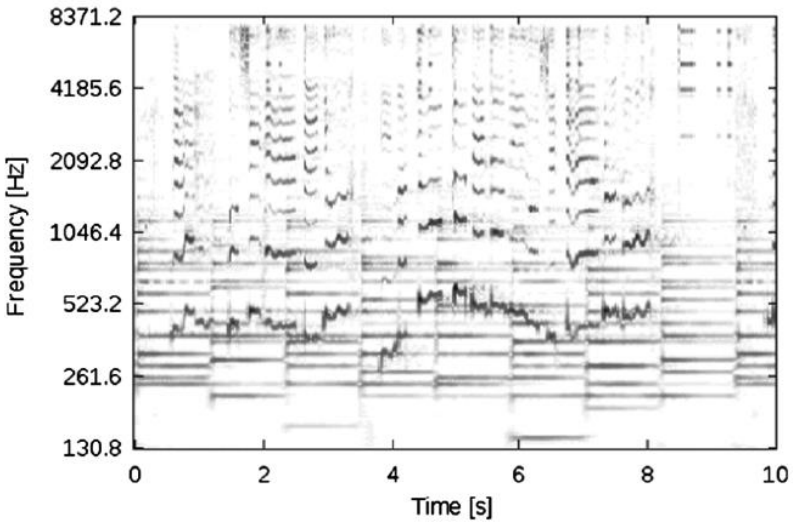



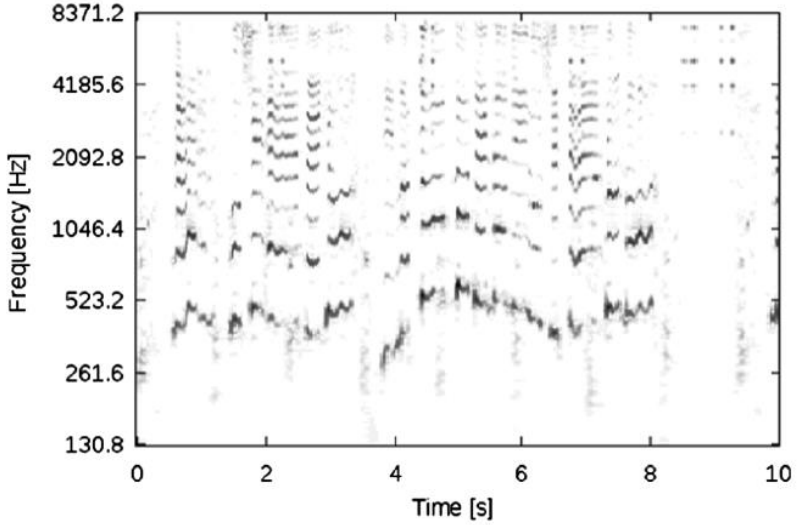




From: M. Mueller, *Fundamentals of Music Processing*, Chapter 8, Springer 2015

# Singing Voice Separation

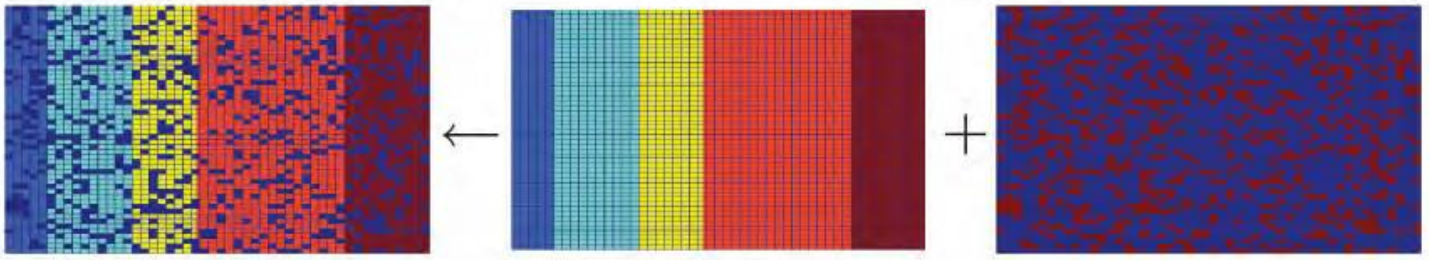


# The Low-Rank Assumption

- A matrix can be decomposed into the sum of a low-rank and a sparse matrix
- ***Rank***: number of linearly independent rows or columns



$$M = L_0 + S_0$$

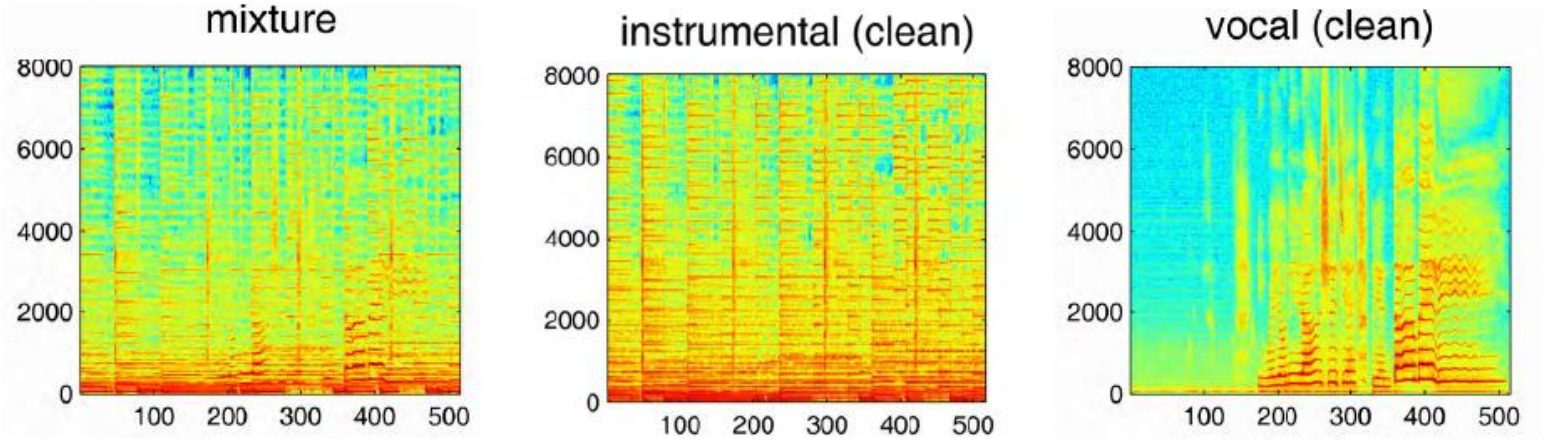
- $M$: data matrix (observed)
- $L_0$: low-rank (unobserved)
- $S_0$: sparse (unobserved)

# The Low-Rank Assumption

- Foreground/background separation
- Singing voice is the foreground and is sparse
- Accompaniment is the background and is low-rank

$$M = L_0 + S_0$$

- $M$: data matrix (observed)
- $L_0$: low-rank (unobserved)
- $S_0$: sparse (unobserved)



# The Low-Rank Assumption

- Foreground/background separation

# Classical Principal Component Analysis (PCA)

$$M = L_0 + N_0$$

- $L_0$: low-rank (unobserved)
- $N_0$: (small) perturbation

Dimensionality reduction (Schmidt 1907, Hotelling 1933)

$$\begin{array}{ll} \text{minimize} & \|M - L\| \\ \text{subject to} & \text{rank}(L) \leq k \end{array}$$

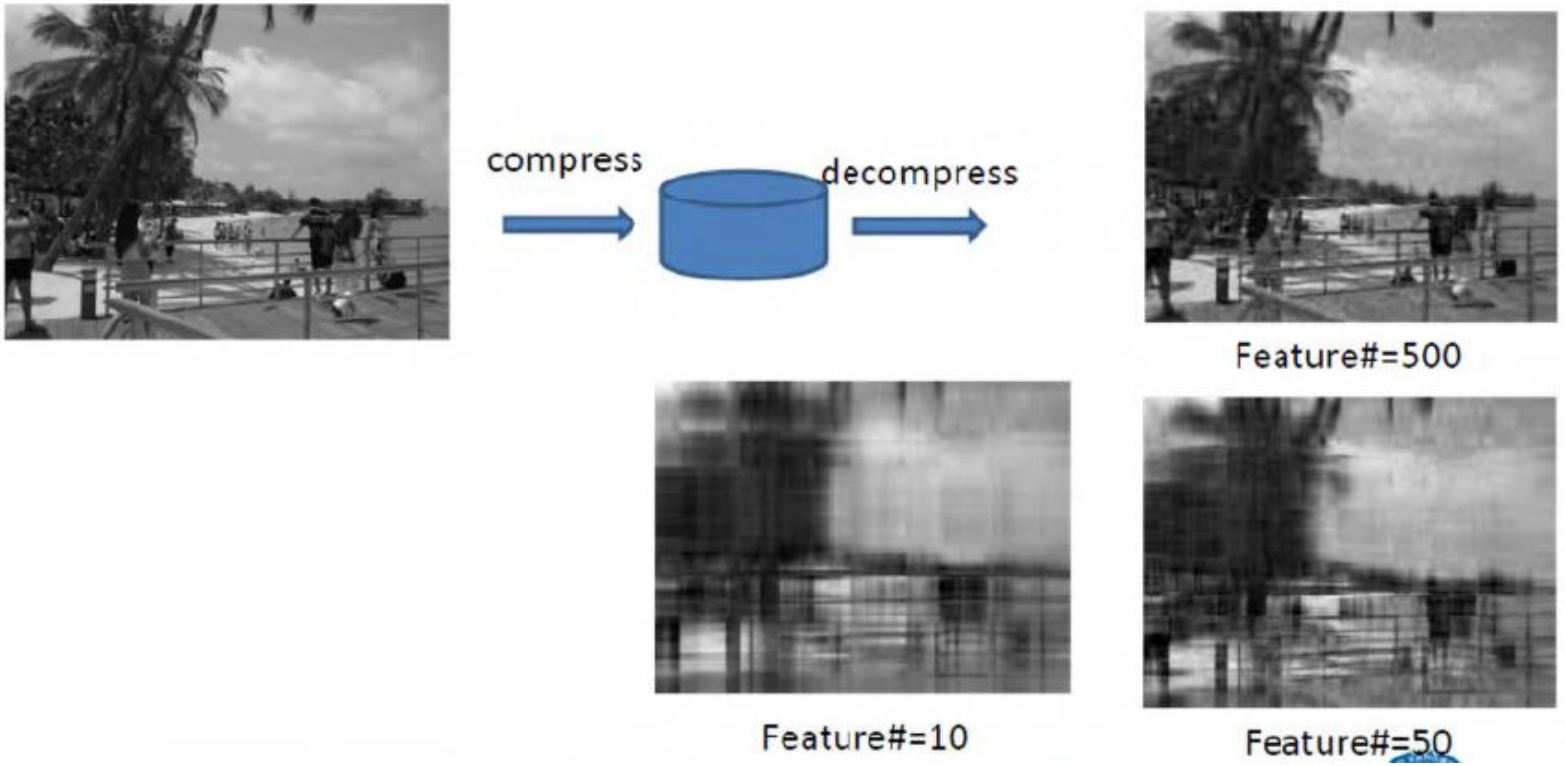
Solution given by truncated SVD

$$M = U\Sigma V^* = \sum_i \sigma_i u_i v_i^* \quad \Rightarrow \quad L = \sum_{i \leq k} \sigma_i u_i v_i^*$$

Fundamental statistical tool: enormous impact

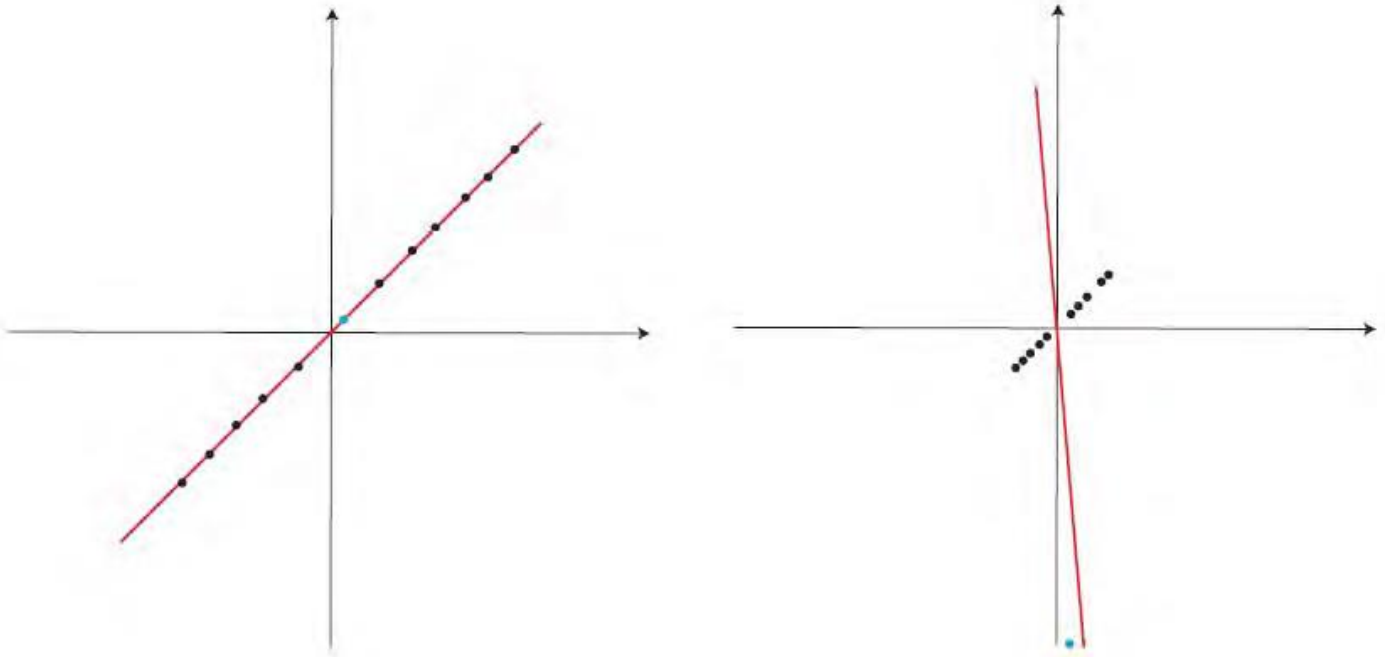
# Classical PCA

- Dimension reduction: using only the first few principal components so that the dimensionality of the transformed data is reduced



# PCA and corruptions/outliers

PCA: very sensitive to outliers



Breaks down with one (badly) corrupted data point

# Robust Principal Component Analysis (RPCA)

- Decomposes $M$ into a low-rank matrix $L$ plus a sparse matrix $S$

$$\min \|L\|_* + \lambda \|S\|_1$$
$$\text{s.t. } L + S = M$$

- $l_1$-norm: $\|S\|_1 = \sum_{ij} |S_{ij}|$ (sum of absolute values; a convex relaxation of $l_0$ norm)
- Nuclear norm: $\|L\|_* = \sum_i \sigma_i(L)$ (sum of singular values; a convex relaxation of the rank of the matrix)

# Algorithm: Augmented Lagrangian Approach

$$\begin{aligned} &\text{minimize} && \|L\|_* + \lambda\|S\|_1 + \frac{1}{2\tau}\|M - L - S\|_F^2 \\ &\text{subject to} && L + S = M \end{aligned}$$

Lagrangian

$$\mathcal{L}(L, S; Y) = \|L\|_* + \lambda\|S\|_1 + \frac{1}{\tau}\langle Y, M - L - S\rangle + \frac{1}{2\tau}\|M - L - S\|_F^2$$

Easy to minimize over $L$ and $S$ separately

$$\arg\min_L \mathcal{L}(L, S, Y) = \mathcal{D}_\tau(M - S + Y)$$
$$\arg\min_S \mathcal{L}(L, S, Y) = \mathcal{S}_{\lambda\tau}(M - L + Y)$$

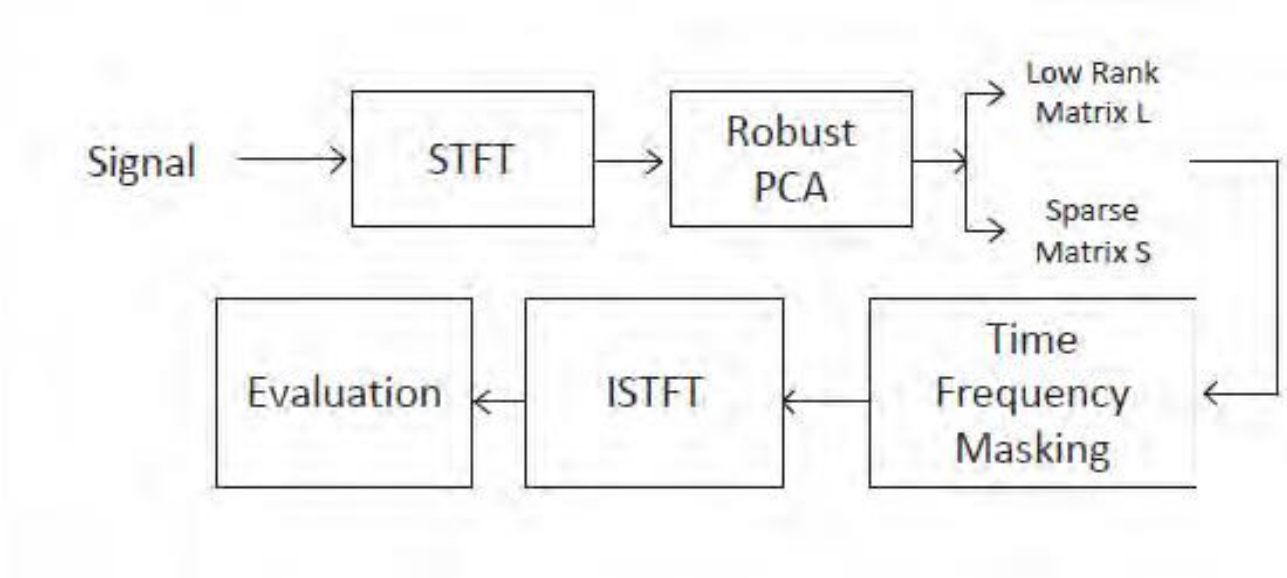
Scalar shrinkage: $\mathcal{S}_\tau[x] = \text{sgn}(x)\max(|x| - \tau, 0)$

- Componentwise thresholding $\mathcal{S}_\tau(X)$
- Singular value thresholding $\mathcal{D}_\tau(X)$

$$\mathcal{D}_\tau(X) = U\mathcal{S}_\tau(\Sigma)V^* \qquad X = U\Sigma V^*$$

# RPCA for Singing Voice Separation

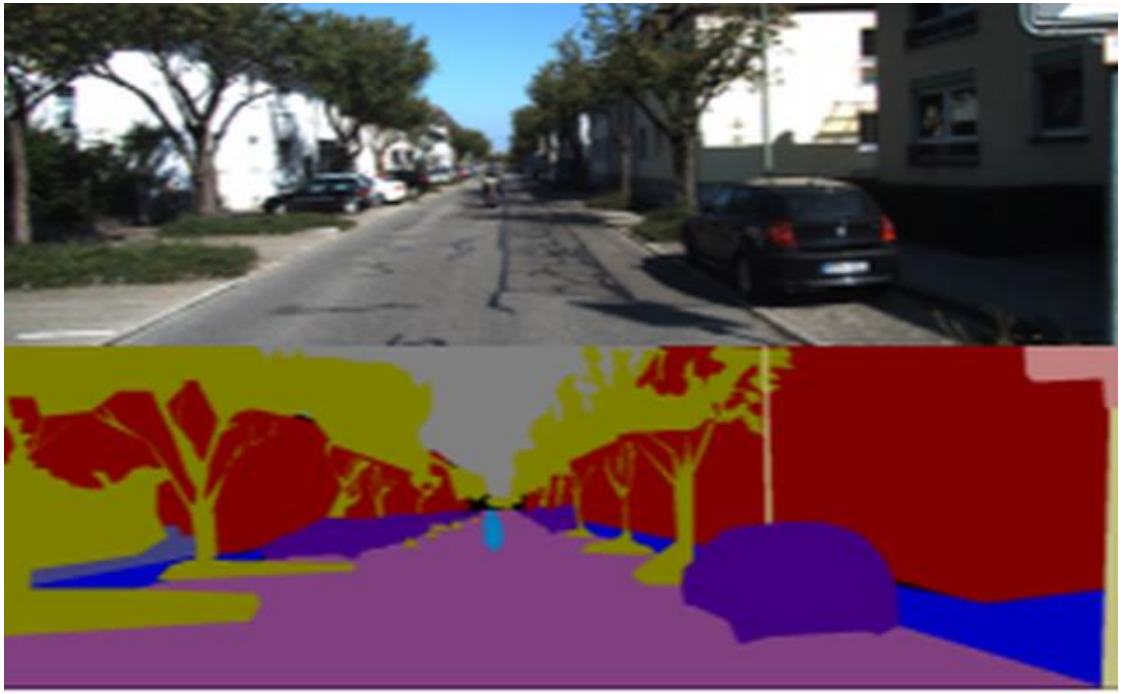
- Huang et al.'s RPCA approach contains the first complete sparse voice (foreground) plus low-rank music (repeating background) model




Singing-voice separation from monaural recordings using robust principal component analysis, ICASSP 2012

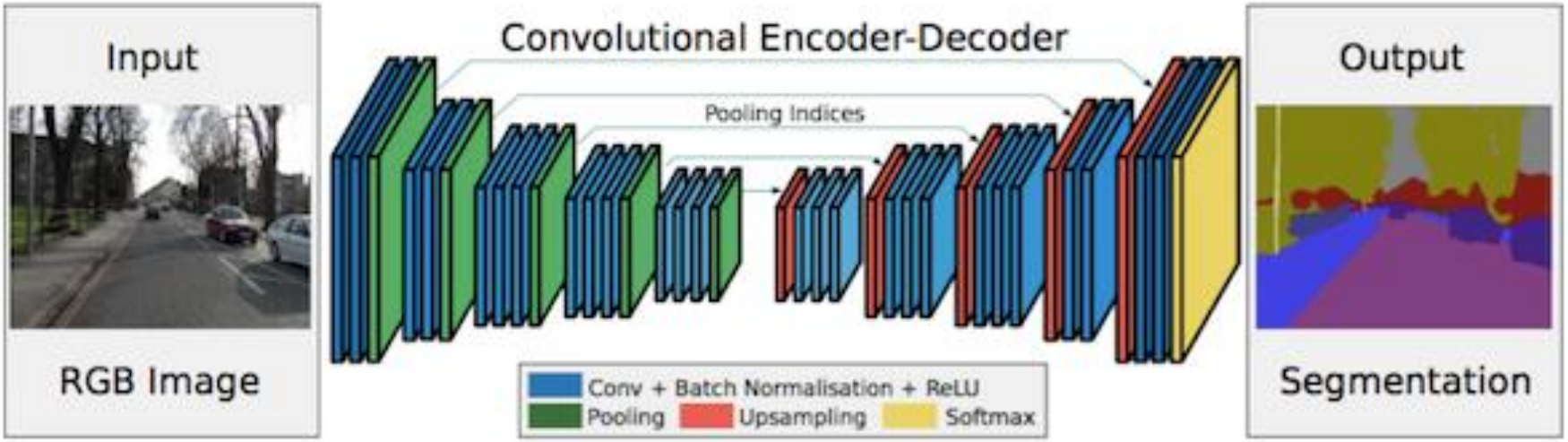
# Source separation as semantic segmentation

- Deep-learning approaches: a great progress in recent years
- Should be helpful for segmenting time-frequency mask for source separation



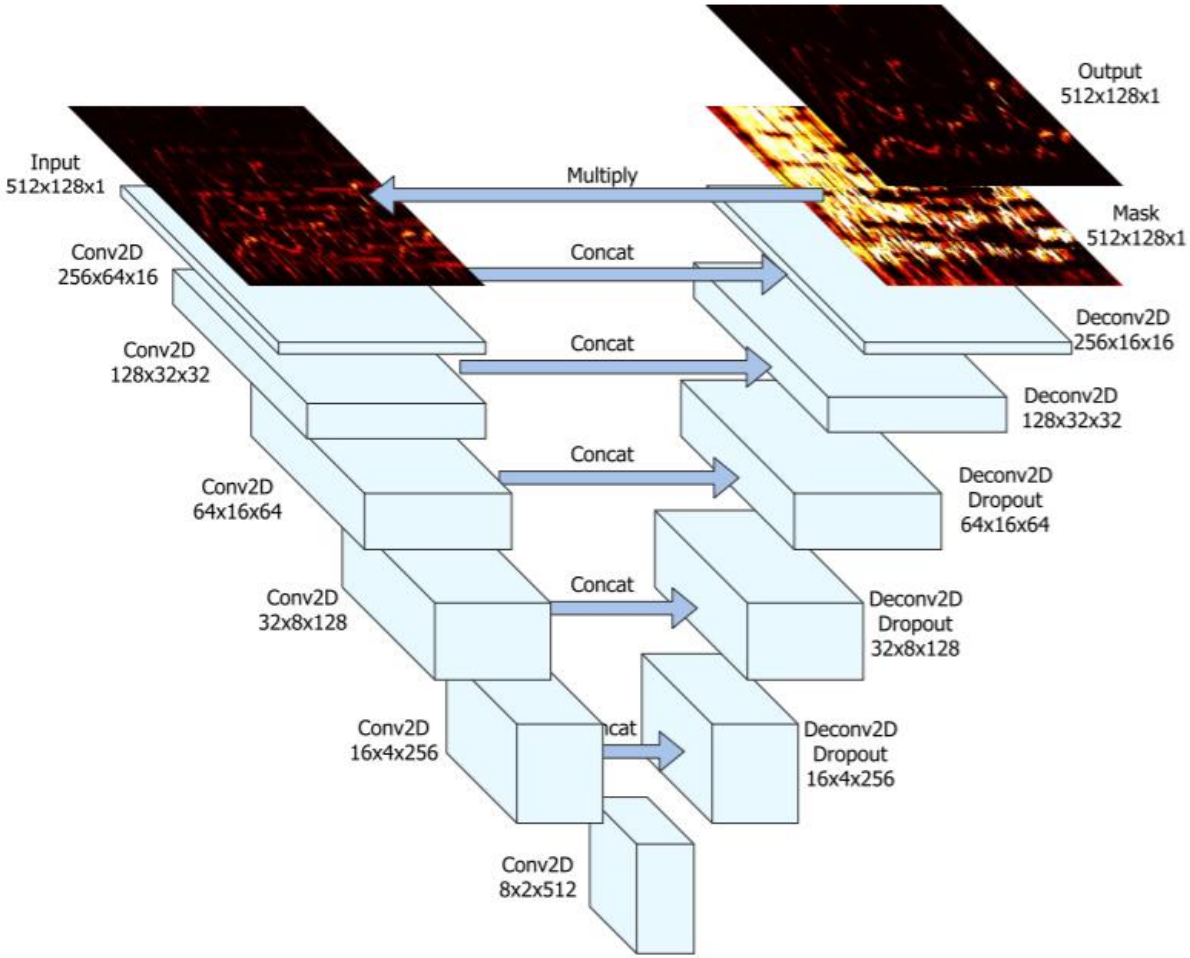
# Examples

- Fully convolutional networks (FCN)
  - U-net, SegNet, …
- R-CNN, mask R-CNN, …



# U-net for source separation

- Learn the time-frequency mask
- Optimize the L1-norm loss




Jansson, Andreas, et al. "Singing voice separation with deep U-Net convolutional networks." ISMIR, 2017.

# Challenges

- Audio sources are “transparent” objects
- Spectrogram is still an suboptimal representation
  - Image: what it looks like is exactly the object
  - Audio: harmonics, time-frequency uncertainties, …