Automatic Music Transcription

Li Su

Associate Research Fellow

Music and Culture Technology Lab (MCTL)

Institute of Information Science, Academia Sinica

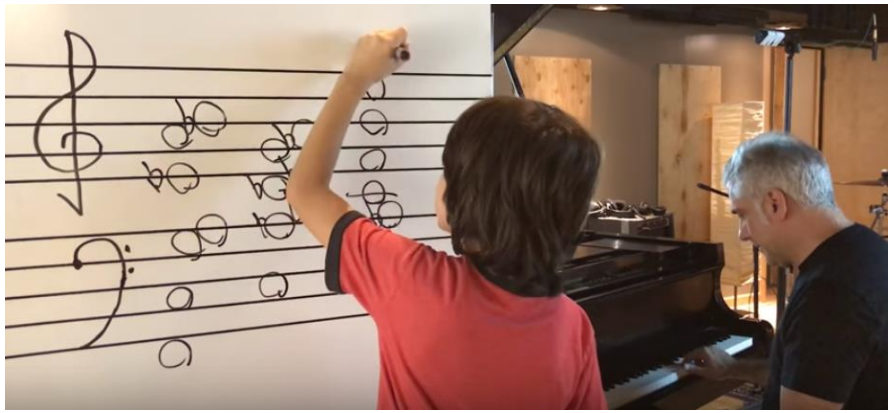
On transcription

- Transcription (採譜 in CH and JP); Automatic Music Transcription (AMT)
  - Literal meaning: audio-to-score
  - What is a score?
- The current definition of automatic music transcription
  - The task to convert the components of interest in acoustic music signals into symbolized music notation such as to
  - 1) record music-meaningful events that musicians executed in a performance;
  - 2) generate a written or printed version of music contents that facilitates human reading and comprehension.
  - Most (if not all) of the AMT research has focused on the first.
- Music is polyphonic: the major complexity of the AMT task

MCTL

AMT for music performance

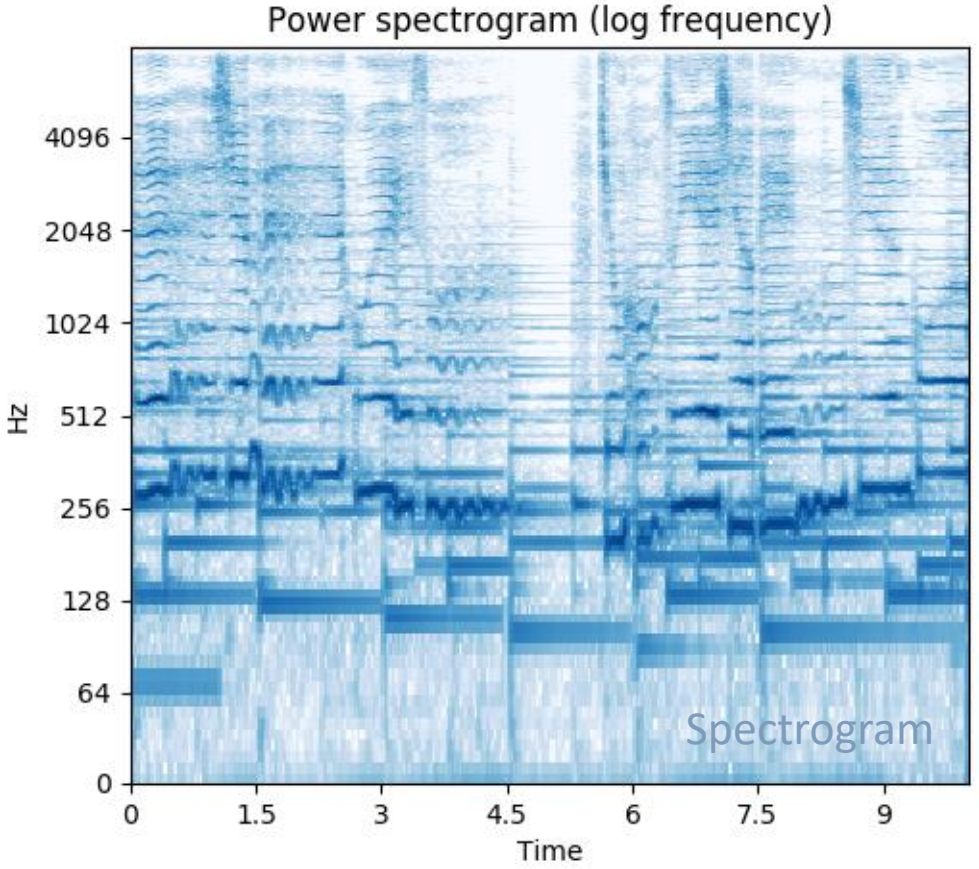
- Similar to a computer vision (CV) problem:
  - Object detection
  - Semantic segmentation
  - Instance segmentation
- Challenges:
  - Multiple objects
  - In CV: occlusion (objects are opaque)
  - In music: objects are transparent; objects overlap with each other both in time (beat position) and in frequency (harmony)
- Audio-to-score AMT: similar to NLP?

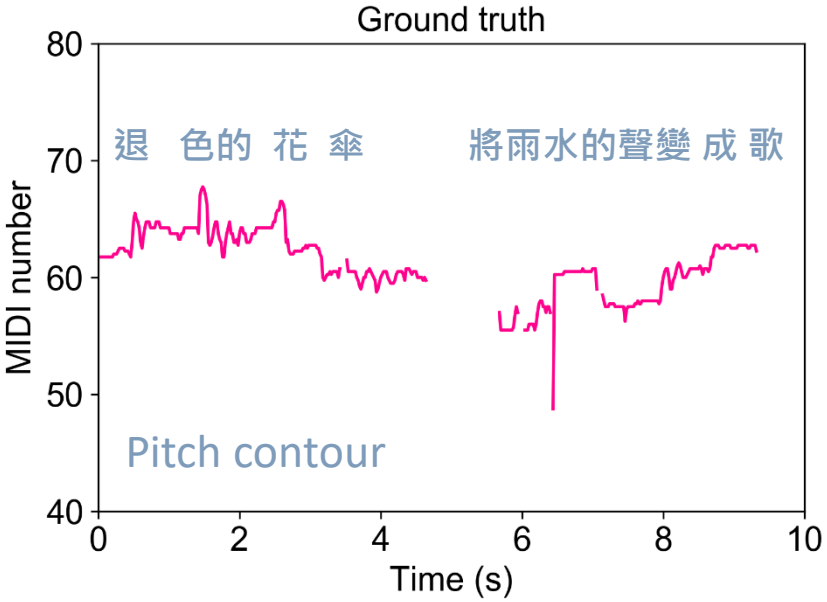


From YouTube: https://youtu.be/hli-9maxDjY

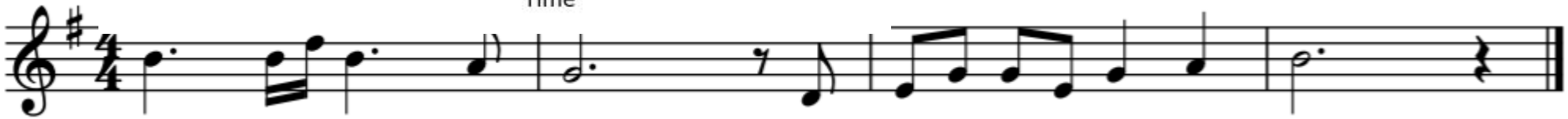
MCTL

Semantic levels in pitch detection: example





Score



Chord G D Em Am D G

MCTL

Types of AMT tasks

			Semantic level		
Info	Track		Frame	Note	Notation
	Track/ voice/ instrument/ stream		F0 contour	pitch, onset, offset, playing technique, etc.	position, note value, pitch spelling, etc.
	Input	Output			
Complexity	Single-track	Single-track	Single-pitch detection	Note tracking (NT)	Score/ notation transcription
	Multi-track	Track-agnostic	Multipitch estimation (MPE)		
		Melody track only	Melody extraction	Melody transcription	
		Track-informed	Multipitch streaming (MPS)	Note streaming (NS), Drum transcription, etc.	Full score transcription
		Track-uninformed			

MCTL

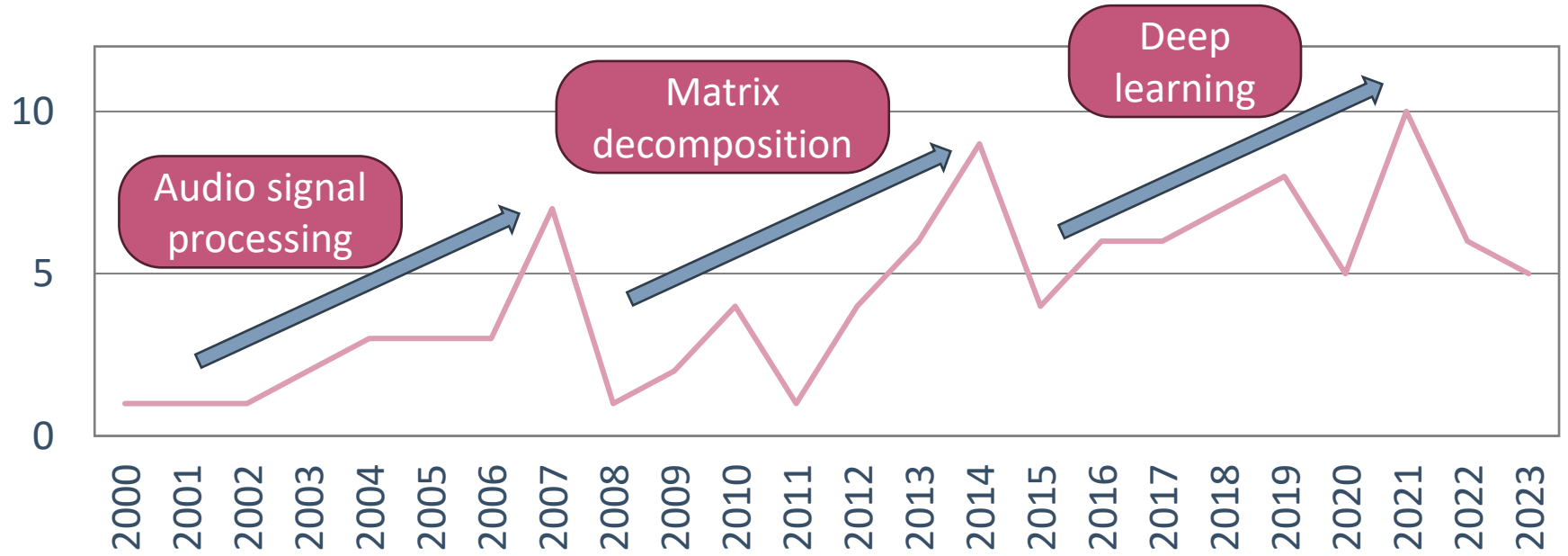
Number of transcription papers in ISMIRs

- Number of papers having “transcription” in their titles each year

MCTL

Number of transcription papers in ISMIRs

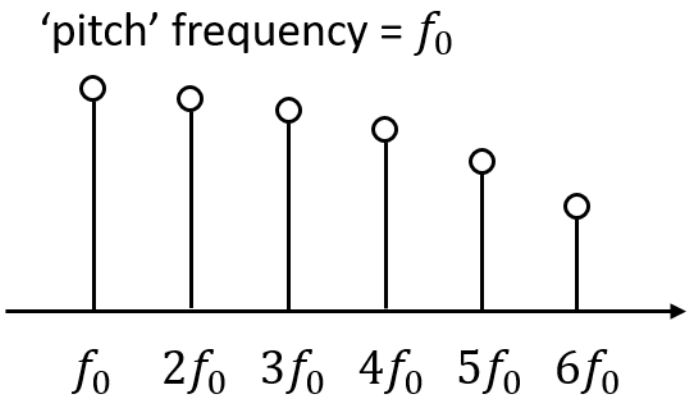
- Number of papers having “transcription” in their titles each year



MCTL

What is pitch?

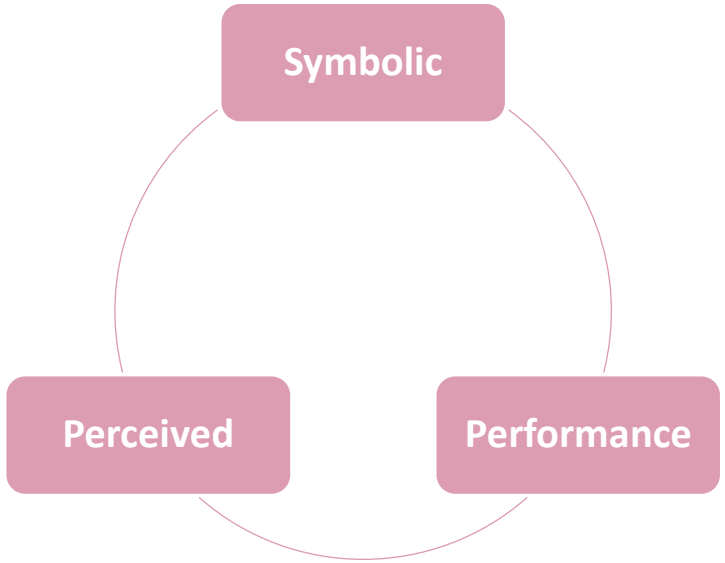
- Pitch: “that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high” (ANSI, 1994)
- Pitch is a *perceptual* quantity, while fundamental frequency (F0) is a *physical* quantity
- In MIR, pitch is often (but not always) considered to be equivalent to F0
- When is pitch not equivalent to F0? missing fundamentals, masking effects, pitch shifts, virtual pitch, dichotic pitch, and the pitches of things that are not there at all …
- What aspects will we consider when pitch is a *musical* quantity?



MCTL

Pitch as a musical object

- Pitch refers to various musical objects:
  - Symbolic pitch, performed pitch, or perceived pitch?
- MIR perspective (1): instantaneous F0/ pitch contour – frame-level pitch
  - Fine resolution in frequency and time: e.g., 20 cents in frequency and 10 ms in time
  - True and nuanced F0 but “unreadable”
- MIR perspective (2): semitone-level pitch – note-level pitch
  - Coarse resolution: in terms of semitone (usually, in Western music); the way reading a note
  - Insufficient for expressive performance (F0-to-note)
  - Three attribute of a note-level pitch: an onset time, an offset time, and a pitch



MCTL

AMT Approach (1)

Audio signal processing for pitch detection

MCTL

Pitch detection methods

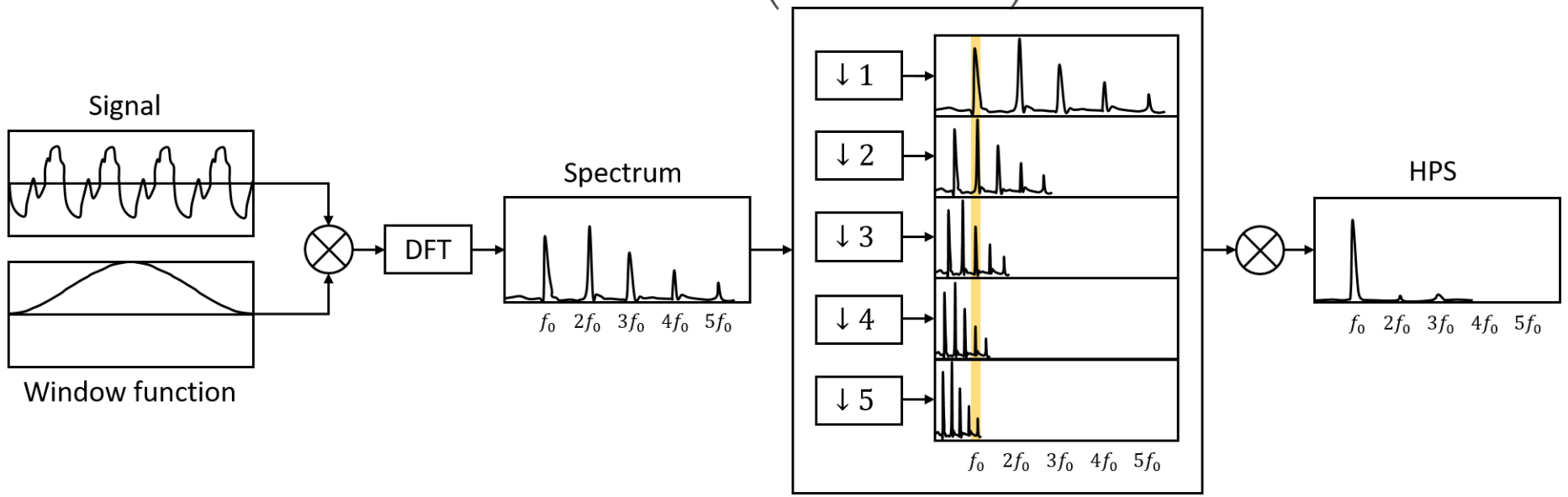
- Signal processing approach
  - Waveform methods: zero-crossing rate
  - Spectral methods: spectrum, harmonic product spectrum
  - Temporal methods: autocorrelation functions, cepstrum and misc.
  - Hybrid methods
- Data-driven approach
  - Using signal processing to extract data representations or just using raw waveform
  - Template matching: k-nearest neighbor and sparse coding
  - Classification: neural networks

MCTL

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 $X[k]$

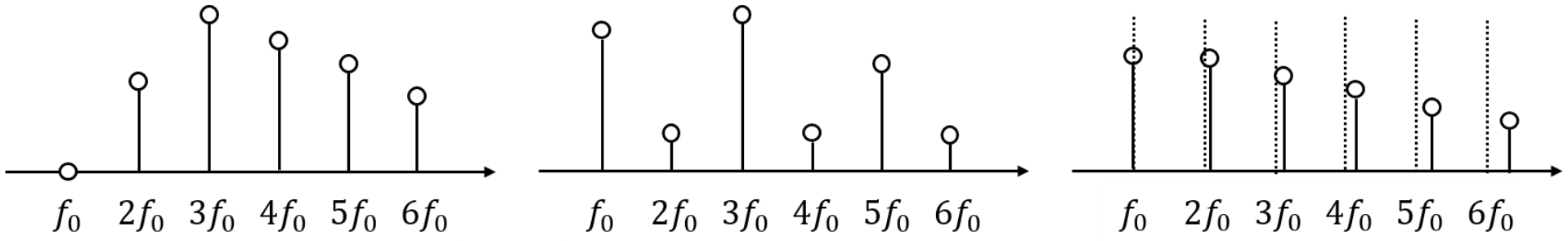
$$\mathrm{HPS}_x[k] = \left(\prod_{m=1}^{M} X[mk] \right)^{\frac{1}{M}}$$



MCTL

Issues of spectral approach in pitch detection

- Why not just taking the index corresponding to the spectral maximum? Any issue?
- 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 $f_n = nf_0\sqrt{1+\beta n^2}$
  - Piano, guitar, and other stuck-string or pluck-string instruments ...



MCTL

Basic periodicity detection functions

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

MCTL

Autocorrelation function (ACF)

- Measure the similarity of a signal and itself across time
- Random-process formulation: $R_{xx}(\tau) = \mathbf{E}\left[x(t)x(t+\tau)\right]$
- Continuous-time formulation:

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)d\tau$$

- Discrete-time formulation: N-point (*estimated*) ACF for $x = \left[x[0], x[1], \ldots, x[N-1]\right]$

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

- t: time-domain
- τ: lag-domain (the unit the same as time)

MCTL

Other relevant pitch detection functions

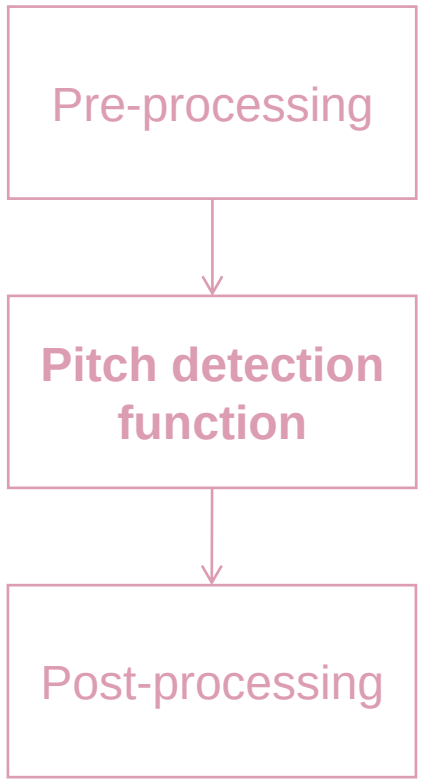
- Average magnitude difference function (AMDF)

$$AMDF[\tau] = \frac{1}{N-1} \sum_{t=0}^{N-1-\tau} |x[t] - x[t+\tau]|$$

- The pitch detection function used in YIN

$$YIN[\tau] = \frac{1}{N-1} \sum_{t=0}^{N-1-\tau} (x[t] - x[t+\tau])^2$$

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



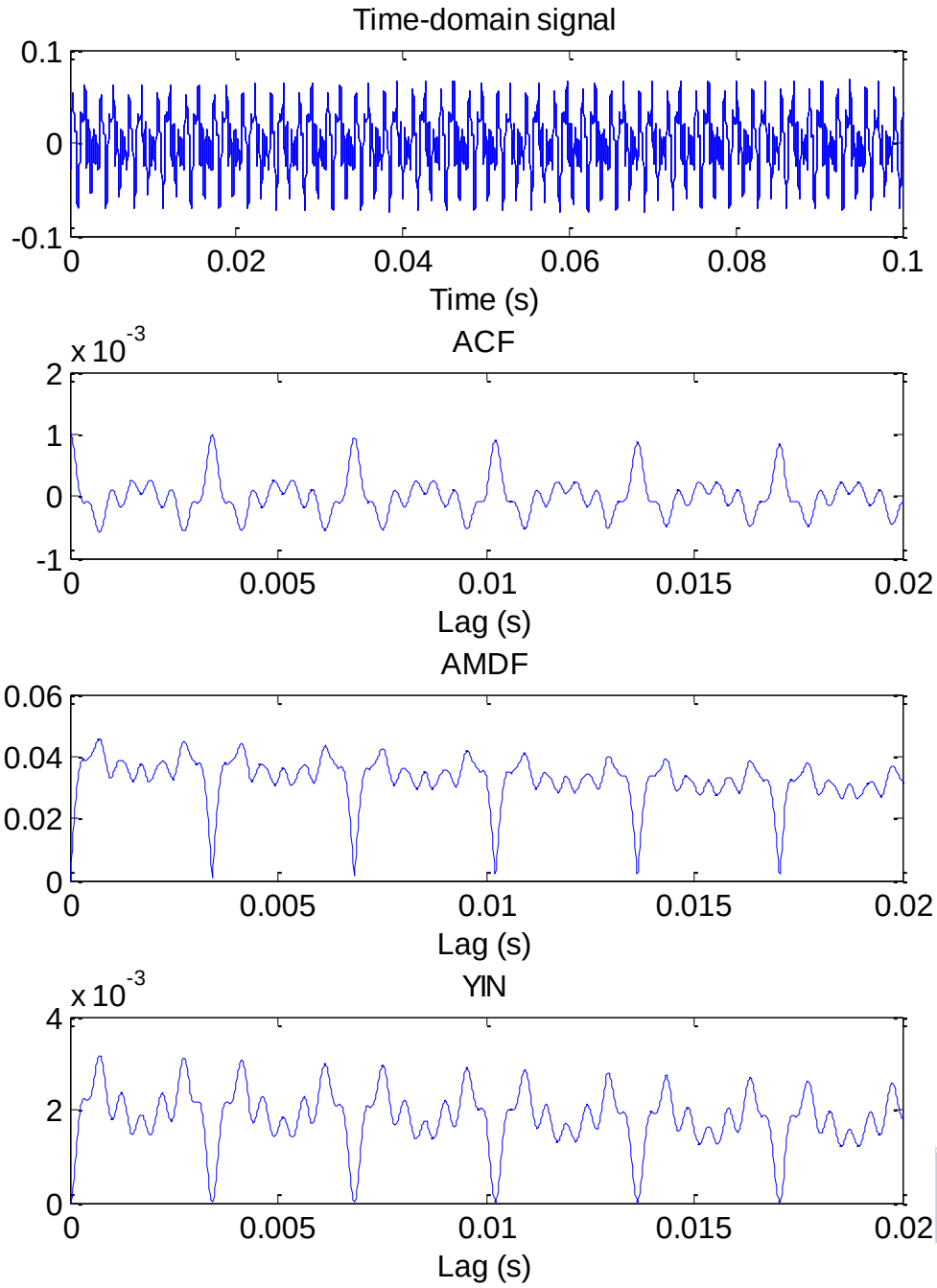
MCTL

Example

- A violin D4: $f_0 = 293$ Hz, $T_0 = 3.41$ msec
- Discarding zero-lag term (for the zero lag of a signal just matches the signal itself)
- The pitch indicator: we have

$$\tau^* = \text{argmax}_\tau \, ACF(\tau)$$
$$\tau^* = \text{argmin}_\tau \, AMDF(\tau)$$

- Then $T_0 \leftarrow \tau^*, f_0 = 1/T_0$



Generalized ACF

- Consider a generalization of ACF:

$$R_{xx}[\tau] = IFFT(|FFT(x[\tau])|^{\gamma}), 0 < \gamma < 2$$

- Why considering a 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.

MCTL

Cepstrum for pitch detection

- From spectrum to cepstrum (倒頻譜)
- Spectrum computed by the 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)

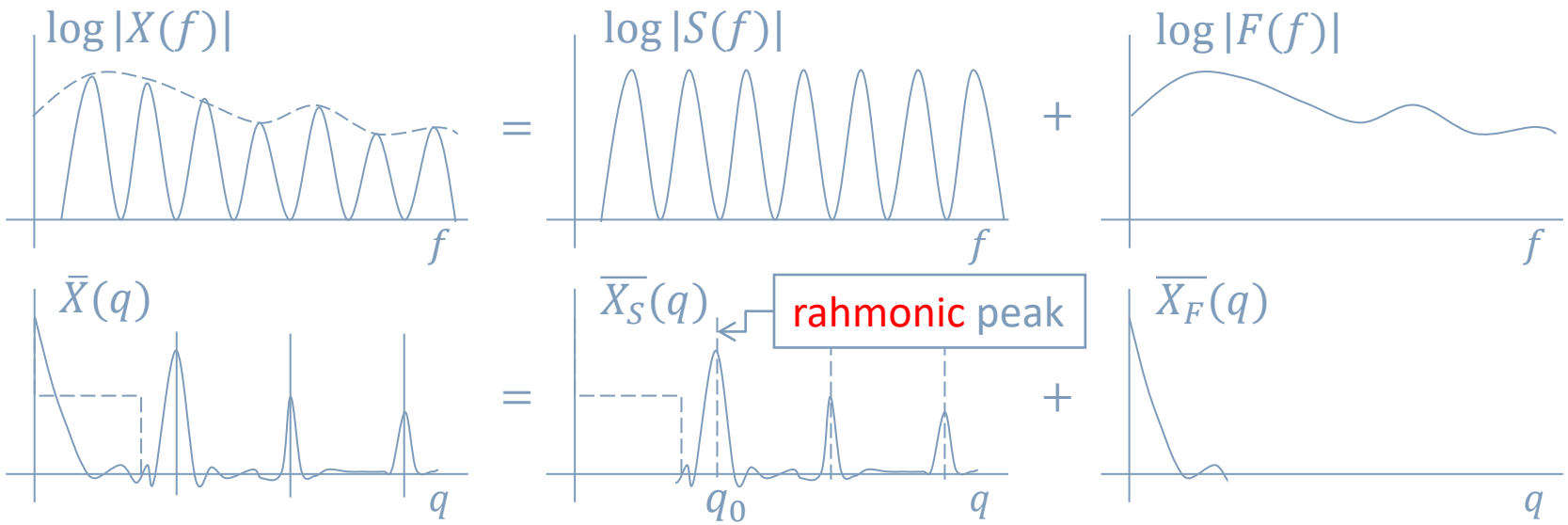
Frequency domain	Cepstrum domain
Frequency	Quefrency
Spectrum	Cepstrum
Harmonic	Rahmonics
Filtering	Liftering
Low-pass filter	Short-pass lifter
High-pass filter	Long-pass lifter

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 seldom used now)

MCTL

Cepstrum as a pitch estimator

- Why considering “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 that the ACF also works in this way (the only difference is the nonlinear scaling term)



AMT Approach (2)

Template matching, matrix decomposition, sparse representation, etc.

MCTL

Template matching for pitch detection: basic

- Also known as spectrogram decomposition: finding the spectra of the individual pitches which sum up to the input (usually multipitch) spectrum
- A "dictionary" $\mathbf{D} \in R^{m \times n}$ be a set of spectra of all single pitches
- $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n]$, column $\mathbf{d}_k \in R^m$ is called an "atom" or a "template"
- Input spectrum (input feature vector): $\mathbf{x} \in R^m$
- Encoding process: solve $\boldsymbol{\alpha} \coloneqq [\alpha_1, \alpha_2, \cdots, \alpha_n] \in R^n$ for the linear equation

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \mathbf{d}_i$$

- Or minimize $\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|$: a regression problem

MCTL

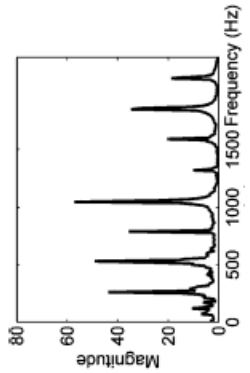
Template matching: single pitch detection

- Input spectrum $\mathbf{x}$, dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_{88}]$, each $\mathbf{d}_k$ is the spectrum of the kth pitch (e.g., $\mathbf{d}_1$ is the spectral pattern of A0, $\mathbf{d}_{40}$ is the spectral pattern of C4)
- Mono-pitch detection: find a $\mathbf{d}_k$ that minimizes $dist(\mathbf{x}, \mathbf{d}_k)$, $dist(\cdot,\cdot)$ being a distance function (e.g., Euclidean distance, cosine distance)
- Vector quantization (VQ): 1-nearest neighbor (1-NN) approximation

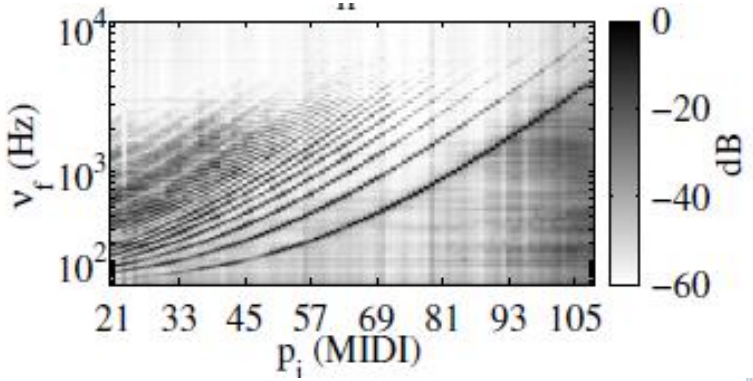
$$\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha} \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_0 = 1$$

The l_p-norm of an n-dim vector $\mathbf{x} \coloneqq [x_i]_{i=1}^n$:

$$\|x\|_p \coloneqq \left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$$



C4



Templates from A0 to C8

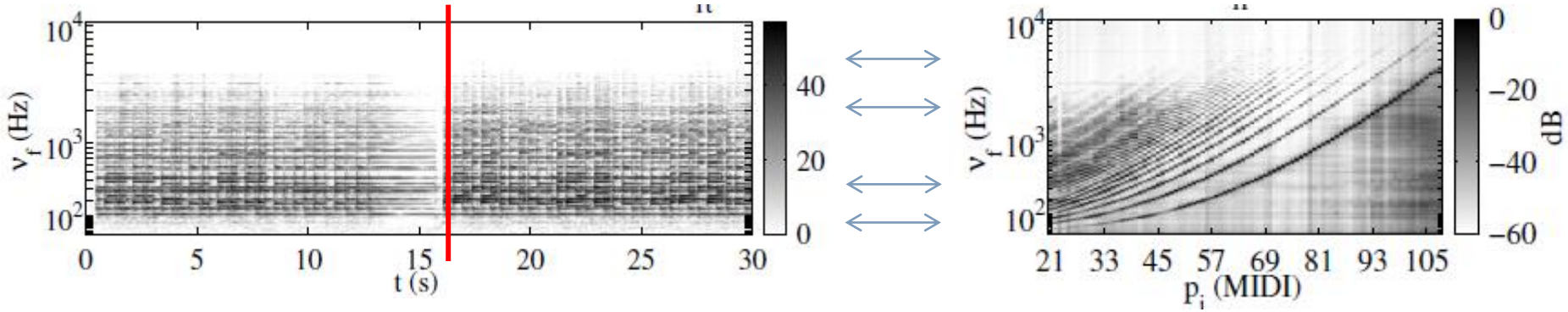
MCTL

Template matching for multipitch estimation

- [Method 1] k-nearest neighbor (kNN): find the atoms having the kth smallest $dist(\mathbf{x}, \mathbf{d}_k)$, or kth largest $\mathbf{x} \cdot \mathbf{d}_k$, etc.
- [Method 2] The sparse coding problem

$$\min \ \|\boldsymbol{\alpha}\|_0 \ \text{ such that } \ \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 < \epsilon$$

- $\|\cdot\|_0$: l_0-norm, the number of non-zero element in x; $\|\cdot\|_2$: l_2-norm, the Euclidean norm



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

MCTL

Algorithms for sparse coding

- The problem that [minimizing $\|\boldsymbol{\alpha}\|_0$ such that $\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 < \epsilon$] is an NP problem (you need to check all the combination of dictionary atoms)
- [Algorithm 1] Method of frames (MoF): solving [minimizing $\|\boldsymbol{\alpha}\|_2$ subject to $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$]
  - A closed-form solution: $\boldsymbol{\alpha} = \mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{x}$
  - But this is actually a non-sparse solution
- [Algorithm 2] Matching pursuit (MP) (Mallat and Zhang, 1993)
  - $\mathbf{x}^{(0)} = 0$, $\mathbf{r}^{(0)} = \mathbf{x}$
  - A greedy algorithm: at the kth step, finding $\mathbf{d}_k$ that maximizes $\mathbf{x} \cdot \mathbf{d}_k$ and determine the weight α_k in order to minimize $\|\mathbf{x} - \alpha_k \mathbf{d}_k\|_2$
  - $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha_k \mathbf{d}_k$, $\mathbf{r}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$
  - End when $\mathbf{r}^{(k)} < \epsilon$

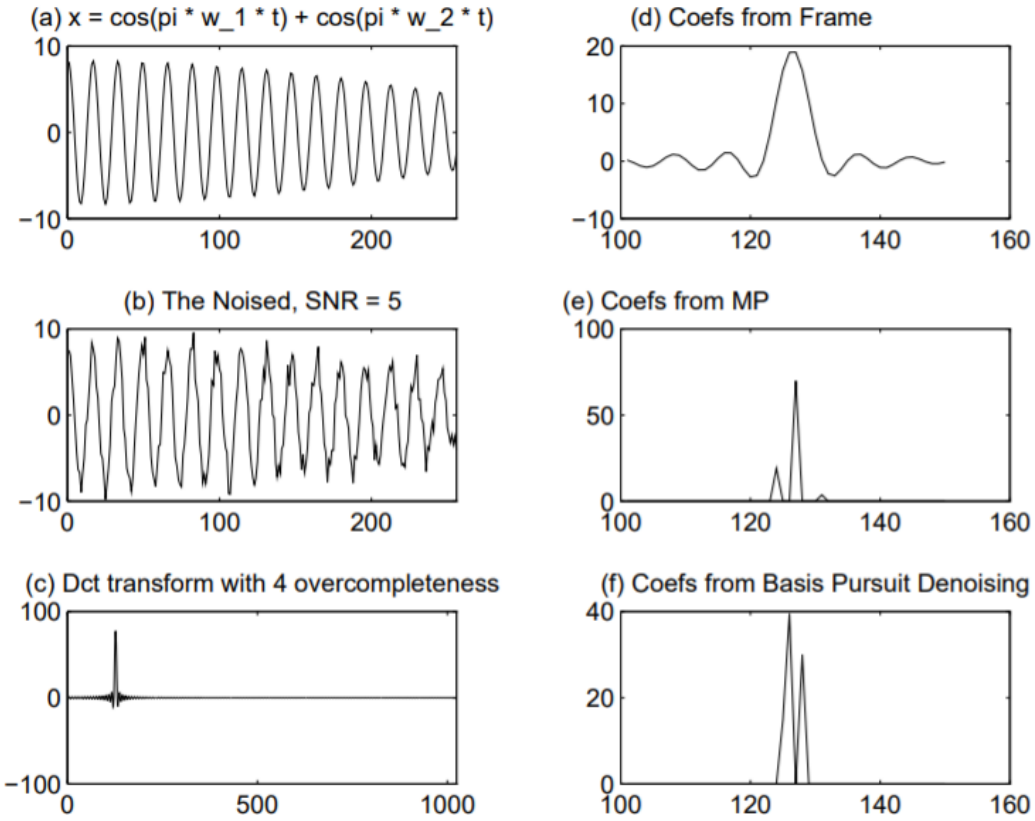
MCTL

Algorithms for sparse coding

- [Algorithm 3] Basis pursuit (BP) (Chen and Donoho, 1992)

$$\min_{\alpha} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

- Global optimization, can be solved with linear programming
- Overcomplete dictionary
  - For $n > 2m$, sparse solution is guaranteed
- l_1-norm regularization
  - l_0-norm: non-convex, no guarantee of global optimal solution
  - l_1-norm: a compromise between convexity and sparsity
  - New algorithms: interior point, homotopy…



Shaobing Chen and David Donoho, "Basis Pursuit"

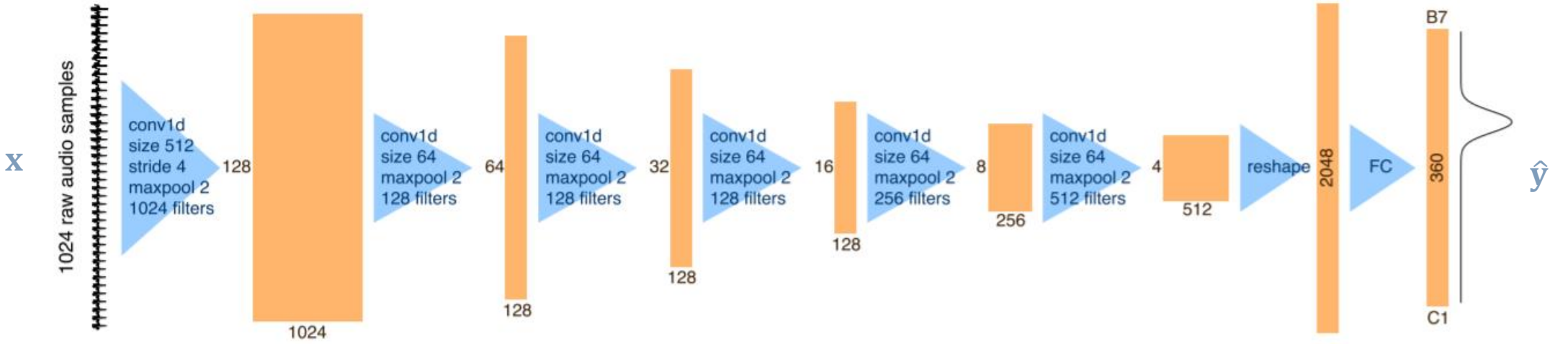
MCTL

AMT Approach (3)

Deep learning

Single-pitch detection with deep learning

- An example: the CREPE library (https://github.com/marl/crepe)
- Raw waveform (1,024 samples) input ($\mathbf{x}$), convolutional neural network (CNN) $\hat{\mathbf{y}} = f(\mathbf{x})$
- One-hot prediction: 360-dimensional output vector after a sigmoid layer, output resolution in 20 cents (equivalent to 6 octaves, from C1 to B6)
- Training: minimizing the binary cross entropy $\mathrm{BCE}(\text{predicted } \hat{\mathbf{y}}, \text{ground truth } \mathbf{y})$

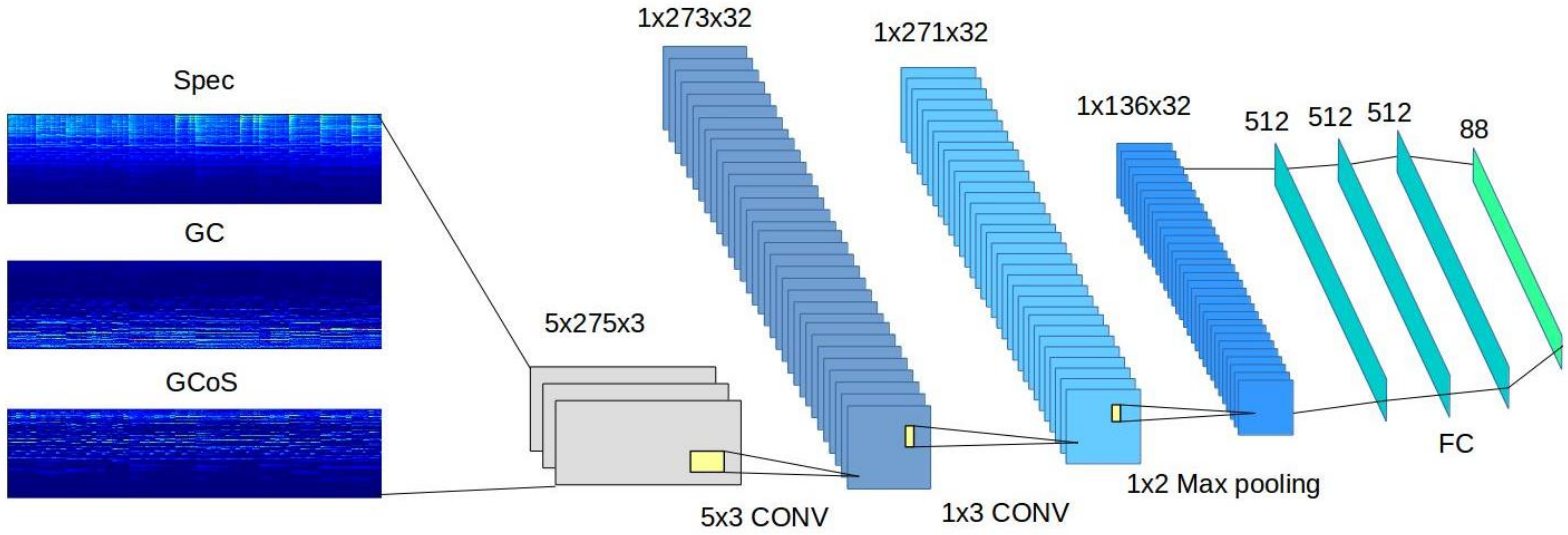


Kim, Jong Wook, et al. "CREPE: A convolutional representation for pitch estimation." IEEE ICASSP 2018

MCTL

Multipitch estimation with deep learning (1)

- Multi-hot prediction [Wu, Chen and Su, in ICASSP 2018]
- 2-D CNN: predict pitch from the input features over a context window
  - Example: predict the pitch at the ith frame from the $(i-\delta)$th to $(i+\delta)$th features
- CNN allows multi-channel inputs; each channel can be any spectral-, cepstral-, and harmonic-based features (in this case, spectrum, generalized cepstrum, and generalized cepstrum of spectrum)

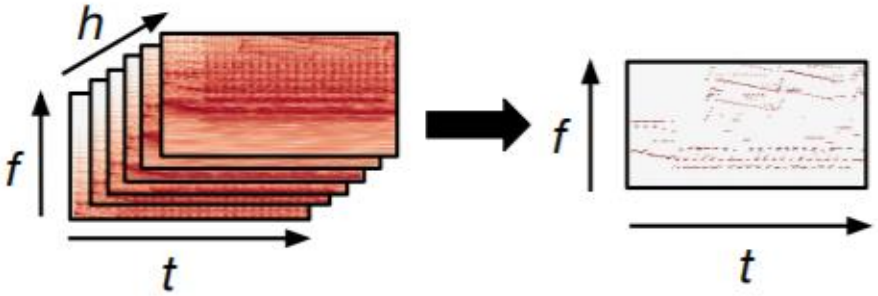


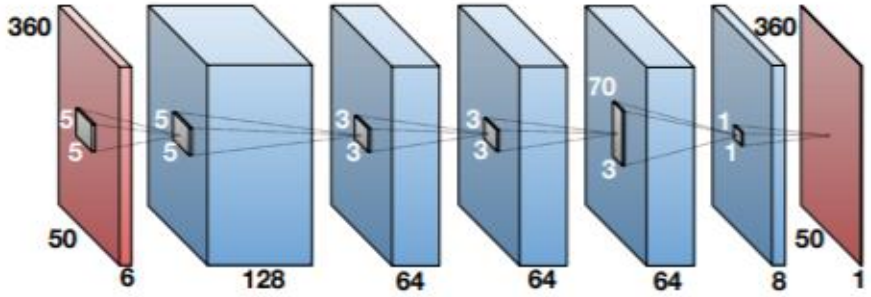
Multipitch estimation with deep learning (2)

- Deep salience map (Bittner et al., 2017): CNN-based image-to-image translation
- Input: harmonic constant-Q transform (HCQT), multi-channel CQTs with different minimal frequencies (f_{min}) positioned at a harmonic series
- Shift-invariant properties: harmonics in log scale are shift invariant for different pitches
- The mth channel of the HCQT:

$$S^{(m)}[k,n] \coloneqq S[k+\eta(m)\cdot\delta,n], \eta(m) \coloneqq \mathrm{round}(12\log_2 m)$$

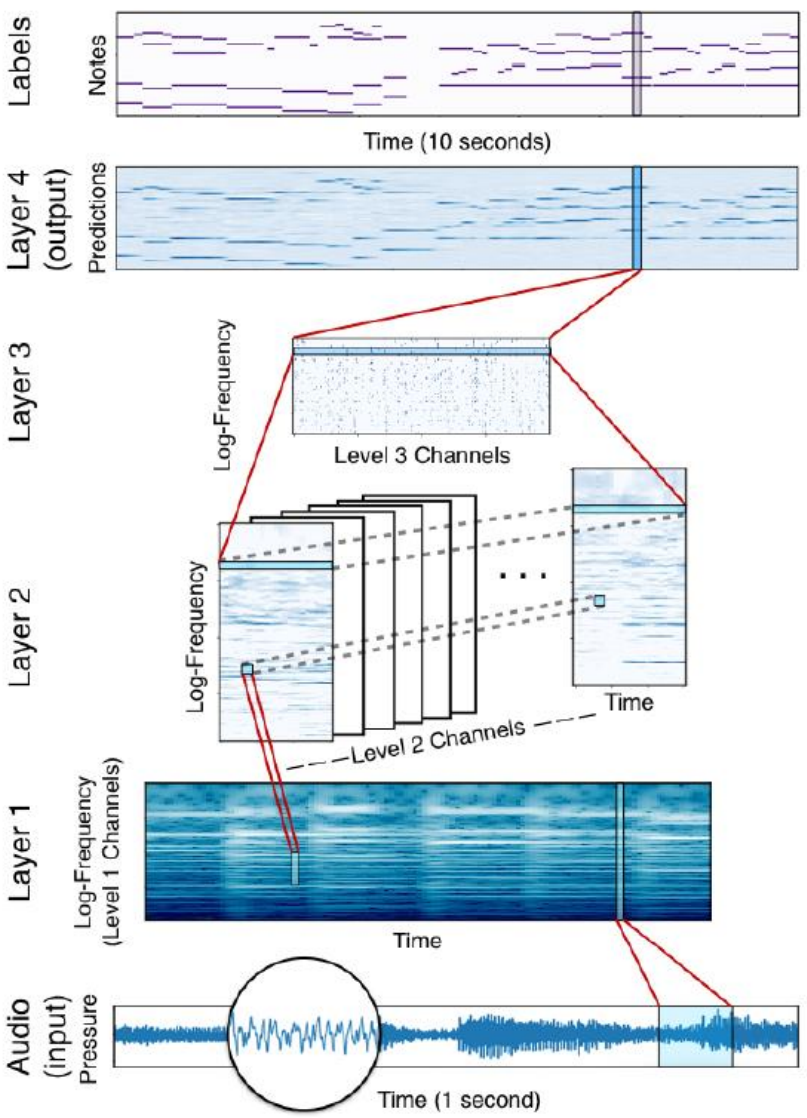
- Recap: harmonic product spectrum



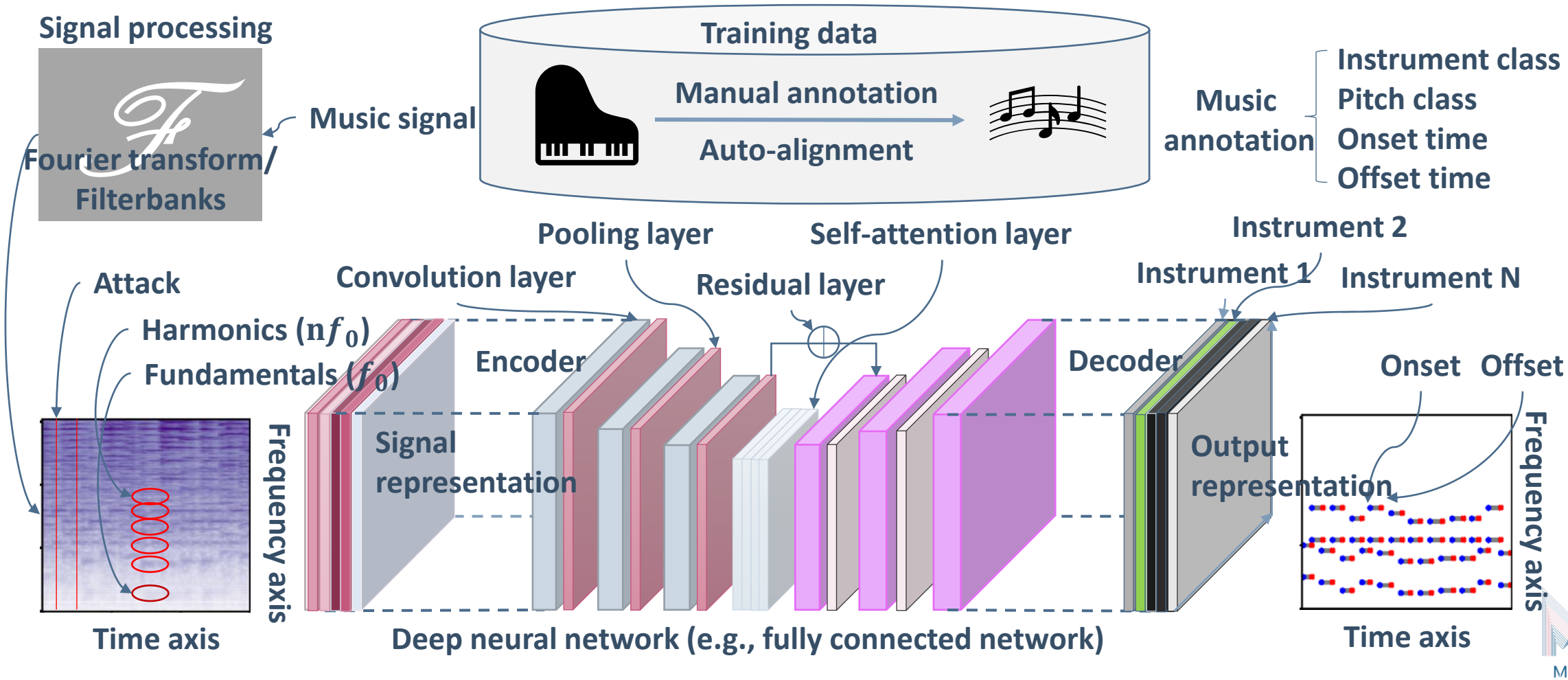


Multipitch estimation (3)

- (Thickstun et al., 2018) translation-invariant network: directly learn useful kernels (for filterbanks) on a long raw audio signal (16,384 samples)
- Output piano rolls



Multi-instrument AMT: Omnizart



Demo of multi-instrument transcription

- Beethoven, Violin Sonata No. 10 in G major, 3. Scherzo: Allegro – Trio


- Beethoven, String Quartet No. 13 in B-flat major, 2. Presto


- Mozart, Serenade for Winds in E-flat major, K. 375, 4. Menuetto II


Yu-Te Wu, Berlin Chen, and Li Su, "Multi-instrument Automatic Music Transcription with Self-Attention-Based Instance Segmentation," *IEEE/ACM Trans. Audio, Signal Language Proc. (TASLP)*, volume 28, pages 2796 - 2809, October 2020.

MCTL

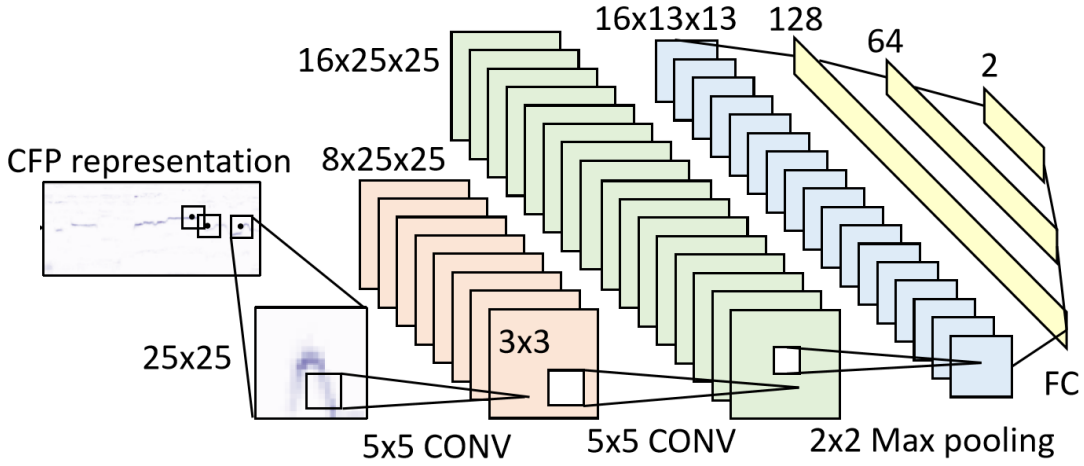
Melody extraction/ transcription & Drum transcription

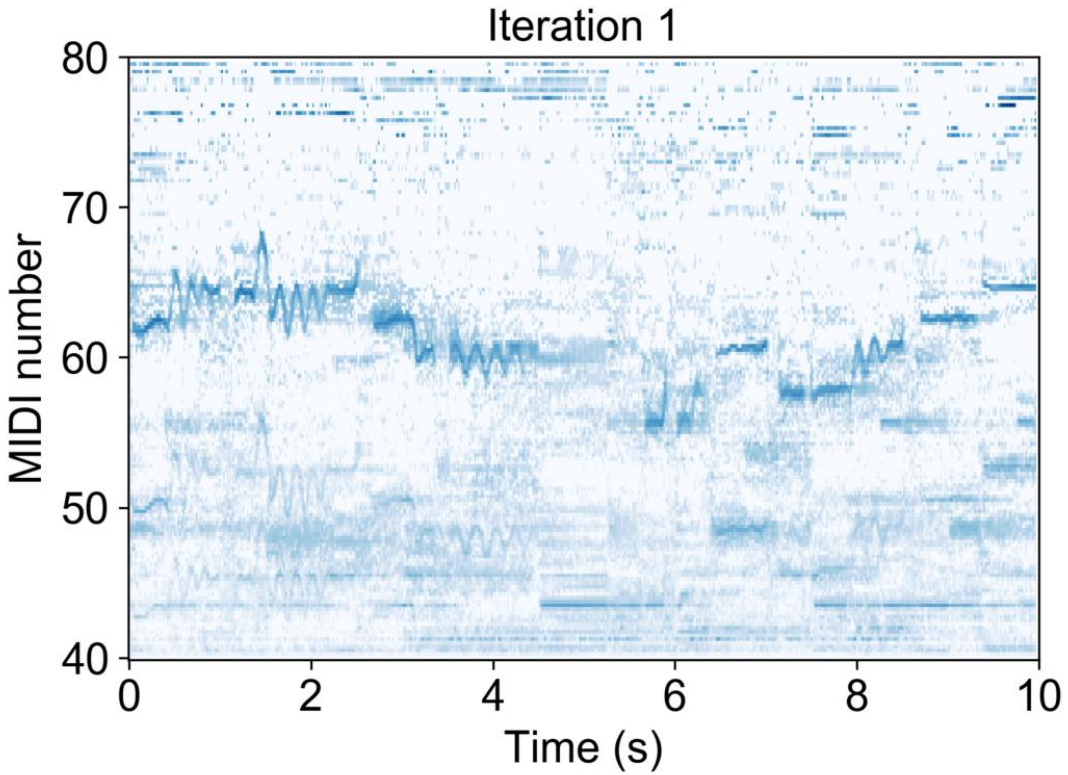
Transcribe specific track(s) in polyphonic music

MCTL

Vocal melody extraction

- Pitch contour classification: the pitch contour of human voice is quite different from other instruments
  - Vibrato / tremolo
  - Slides
- An illustration with the multi-layered cepstrum
- A patch-CNN for vocal melody extraction

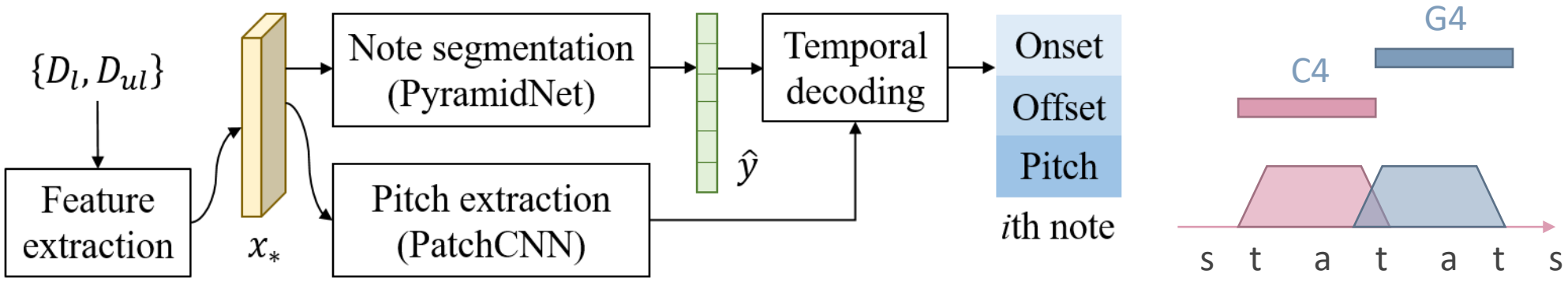




Li Su, "Vocal melody extraction using patch-based CNN," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

VOCANO: the system

- Vocal note transcription
- Recognizing onset, offset, pitch of human singing voice
  - But the classification problem is to classify silence (s), activation (a) and transition (t)
- Temporal decoding: transition state as 1) onset, 2) offset, 3) offset followed by onset
- Pitch contour extracted by PatchCNN



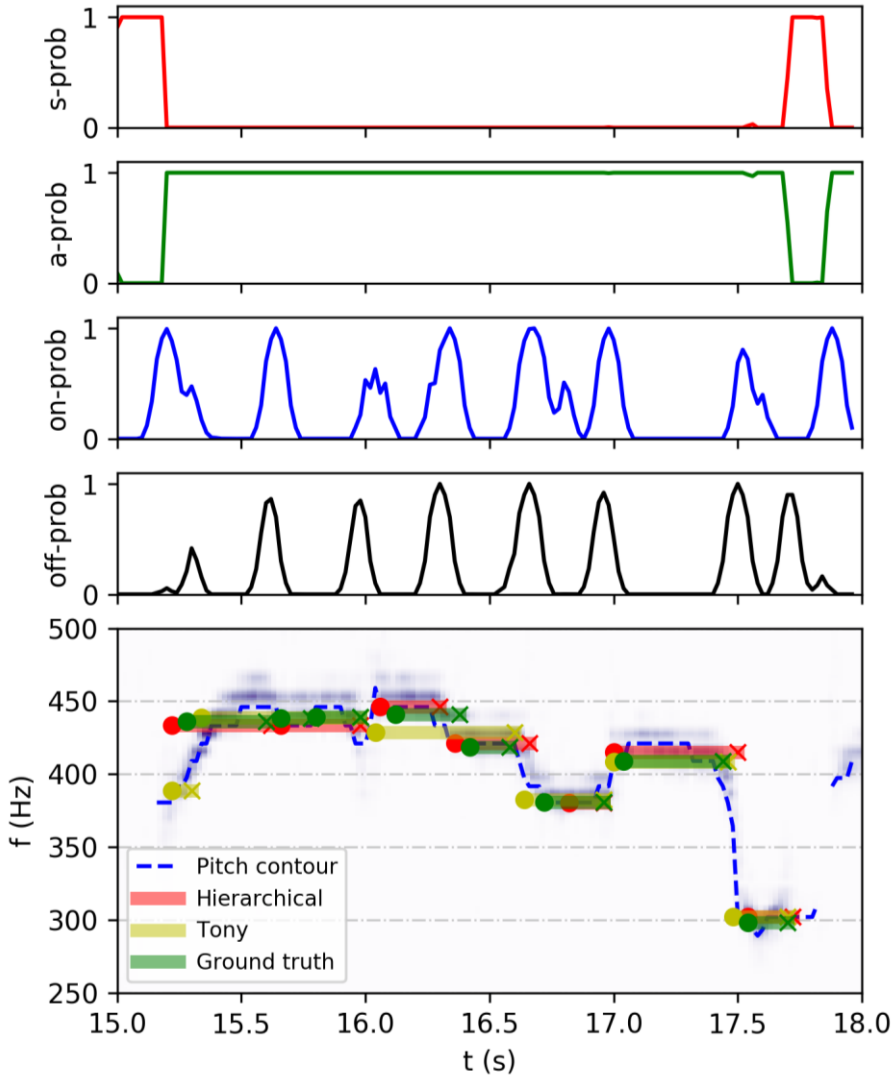
Jui-Yang Hsu, Li Su, "VOCANO: A note transcription framework for singing voice in polyphonic music," *International Society of Music Information Retrieval Conference (ISMIR)*, November 2021.

MCTL

Listening examples

- Transcription result “better than ground truth!”

	male8.wav	child5.wav	child8.wav
Audio			
Ground truth(MIDI)			
Ours			

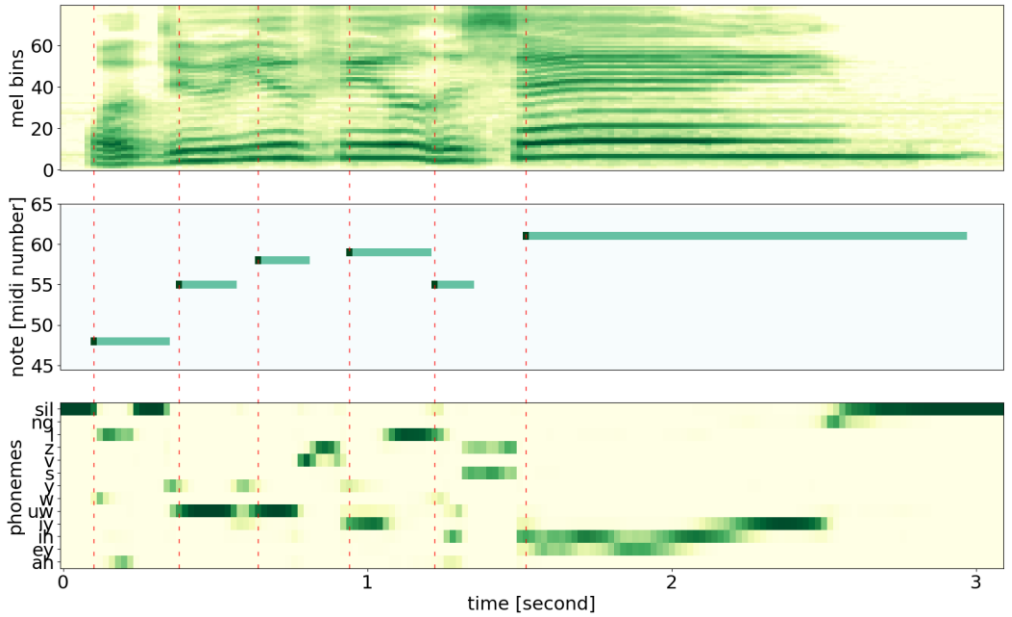


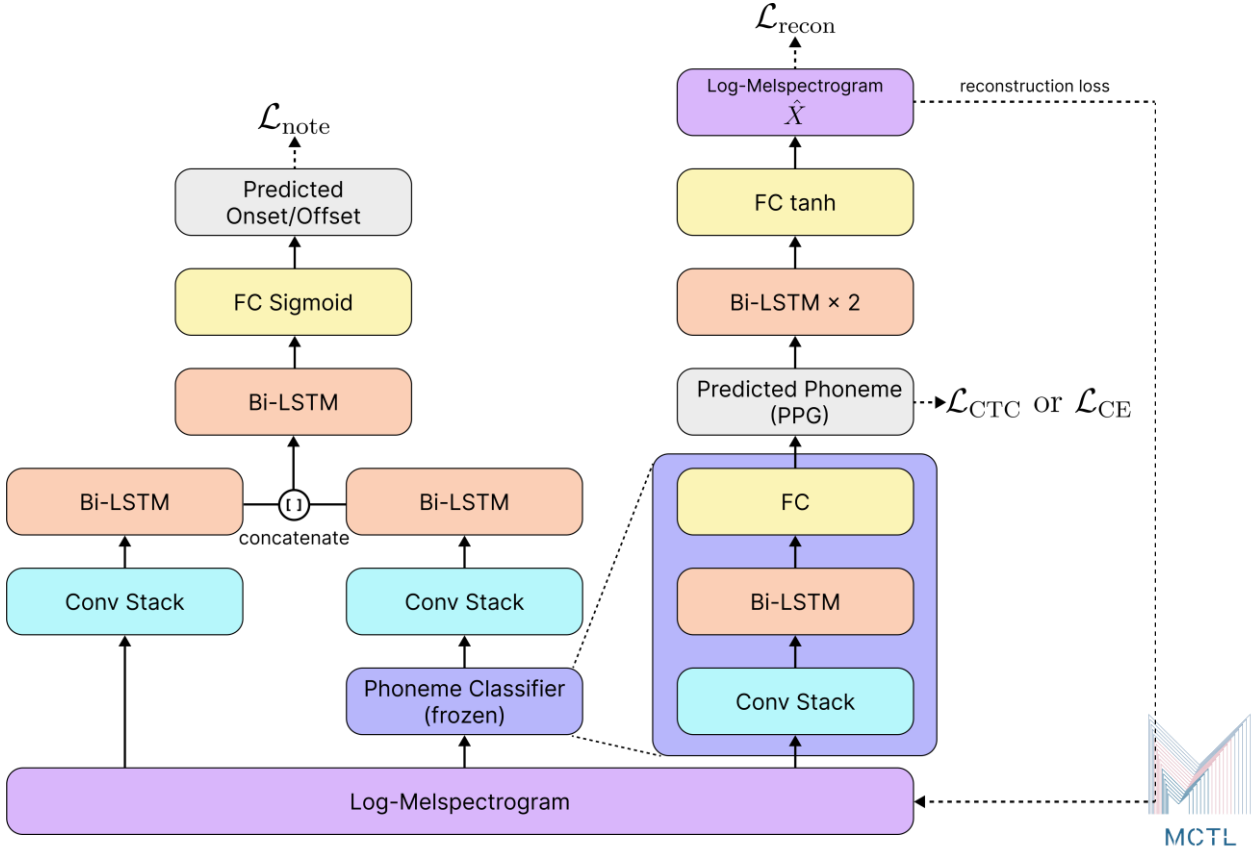
Zih-Sing Fu and Li Su, "Hierarchical classification networks for singing voice segmentation and transcription," International Society of Music Information Retrieval Conference (ISMIR), November 2019.

MCTL

Phoneme-informed singing transcription

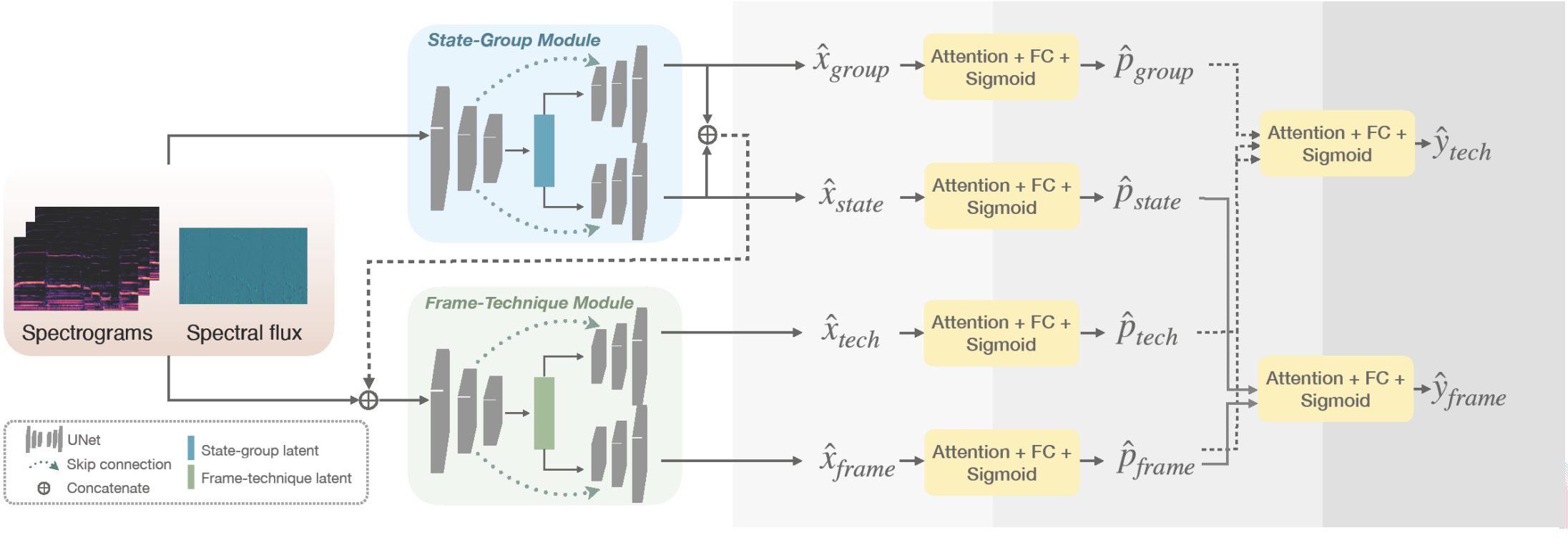
- Sangeon Yong, Li Su, Juhan Nam, "A Phoneme-informed Neural Network Model for Note-level Singing Transcription," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2023.





Guitar solo transcription

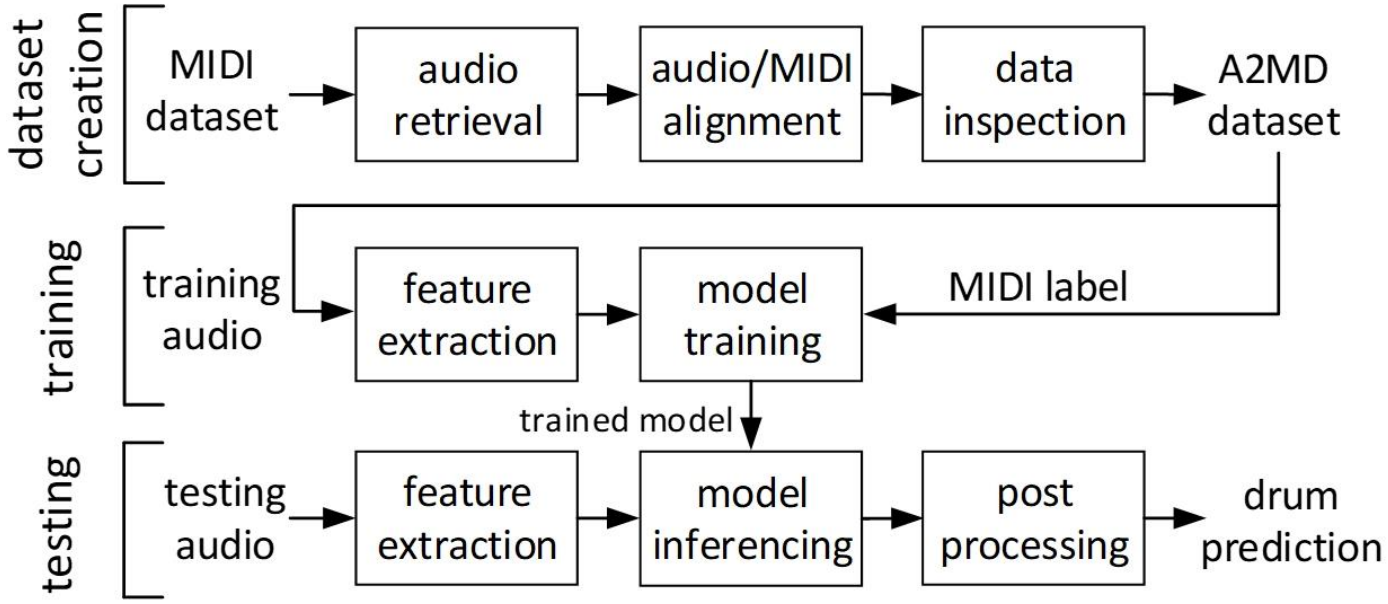
- TungSheng Huang, Ping-Chung Yu, Li Su, "Note and playing technique transcription of electric guitar solos in real-world music performance," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2023.



MCTL

Drum transcription

- Challenge: drums are multi-track and there are very few aligned MIDI data (e.g., label) available for training
- But there are un-aligned MIDI data available
- Solution: audio-to-MIDI alignment by dynamic time warping (DTW)



MCTL

Drum transcription demo

- Michael Jackson - Billie Jean
- Original song

- Transcription result

I-Chieh Wei, Chih-Wei Wu, Li Su, "Improving automatic drum transcription using large-scale audio-to-midi aligned data," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2021.

MCTL

Other issues

Data, annotation, evaluation

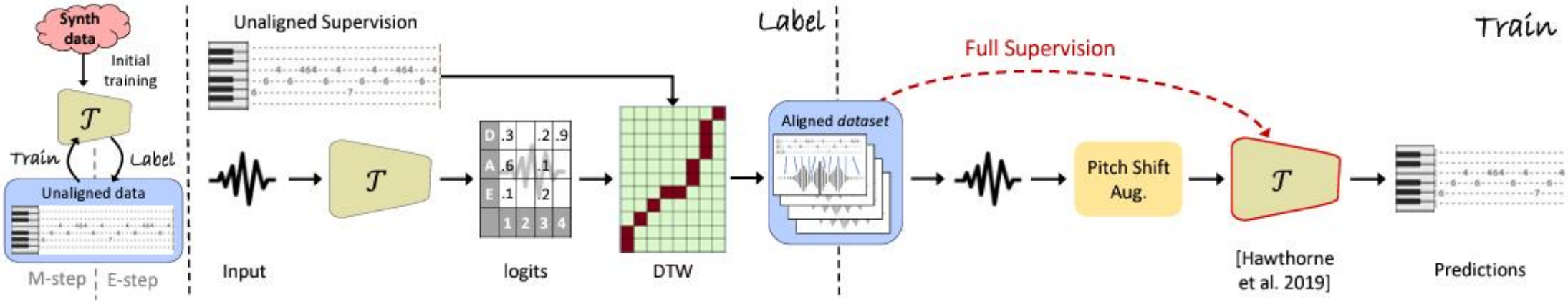
Data annotation

- Full-supervised learning for AMT:
  - Need audio-aligned labels, i.e., audio-aligned MIDI/ music notation files
- Challenge: how to obtain these labels?
  - Manual labels
  - Synthesized labels: using MIDI synthesizers
  - Automatic-aligned labels: using audio-to-MIDI alignment algorithms (e.g., DTW)
  - Machine-assisted labels: auto-piano, guitar string pickup, etc.
- Summary
  - The data annotation you can use is usually less than you imagine
  - The music data you can train is usually limited to the data with MIDI (but such kinds of data do not require the AMT technique)

MCTL

Improving alignment for AMT

- Ben Maman, Amit H. Bermano, Unaligned Supervision For Automatic Music Transcription in The Wild, ICML 2022
- “MIDI-to-MIDI alignment”



Evaluation

- Frame-level evaluation
  - The portion of correct frames (e.g., pitch deviation within 50 cents)
  - Average deviation of pitch contour
- Note-level/ track-level evaluation
  - Onset (e.g., time deviation within 50 ms)
  - Onset-pitch
  - Onset-offset-pitch (offset time deviation with 0.2*note duration)
  - Onset-offset-pitch-track
- Score-level evaluation
  - A subjective question!

MCTL