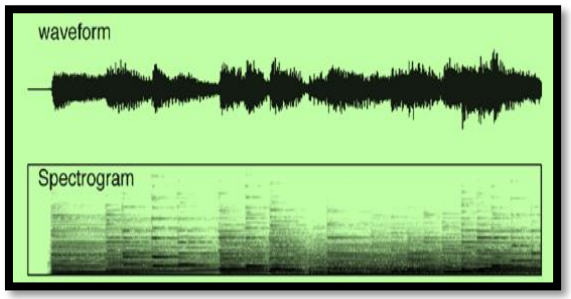
Lecture 8:
Automatic music transcription

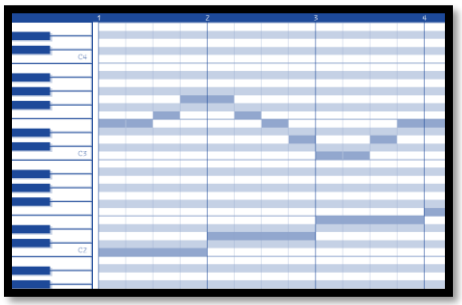
Li Su

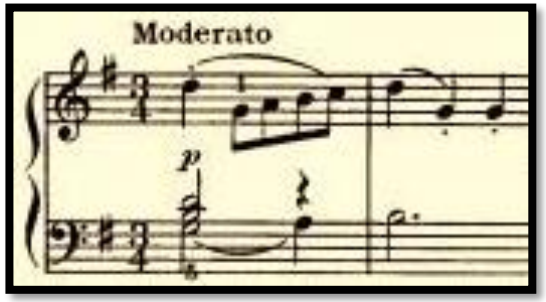
2019/04/22

Automatic music transcription (AMT)

"Polyphonic transcription of digitally encoded music is one of the holy grails of machine listening." (Benetos, 2013)







Physical (audio signal) level:
Musical signal processing
Time-frequency analysis

Symbolic (MIDI) level:
Multi-F0 estimation
Note (onset/offset) tracking
Multi-pitch streaming
Instrument identification

Lexical (score) level:
Voice separation
Key estimation
Meter identification
Score parsing

High-level information: genre/style identification

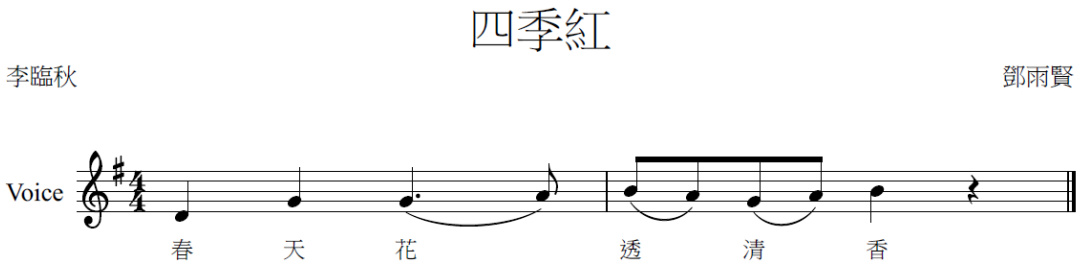
Motivation of AMT

- Not all music piece have scores: we need a written representation of music performance
- Musicological analysis
- A connection between the written music and the performed music
- Structured audio coding 夢想

MusicXML 譜的主流格式

- Open-source, compatible to Finale, Sibelius, MuseScore, etc
- Format:
  - <part> <measure> <attributes> <divisions>
  - <key>
    - <fifths><mode>
  - <time>
    - <beats><beat-type>
  - <clef>
    - <sign><line>

open source



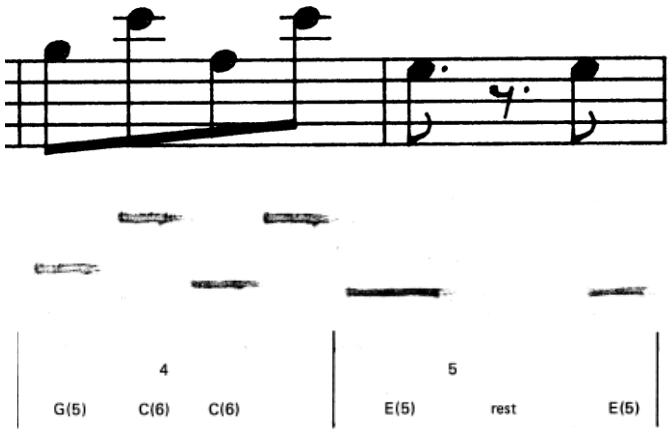

```
56  <part id="P1">
57    <measure number="1" width="533.09">
58      <print>
59        <system-layout>
60          <system-margins>
61            <left-margin>73.88</left-margin>
62            <right-margin>0.00</right-margin>
63            </system-margins>
64          <top-system-distance>180.00</top-system-
65          </system-layout>
66        </print>
67      <attributes>
68        <divisions>2</divisions>
69        <key>
70          <fifths>1</fifths>五度圈上的代號
71          <mode>major</mode>
72          </key>
73        <time>
74          <beats>4</beats>
75          <beat-type>4</beat-type>
76          </time>
77        <clef>
78          <sign>G</sign>
79          <line>2</line>
80          </clef>
81        </attributes>
82      <note default-x="98.40" default-y="-45.00">
83        <pitch>
84          <step>D</step>
85          <octave>4</octave>
86          </pitch>
87        <duration>2</duration>
88        <voice>1</voice>
89        <type>quarter</type>
90        <stem>up</stem>
91        <lyric number="1">
92          <syllabic>single</syllabic>
93          <text>春</text>
94          </lyric>
95        </note>
```

Three levels of polyphonic music transcription 多重音高

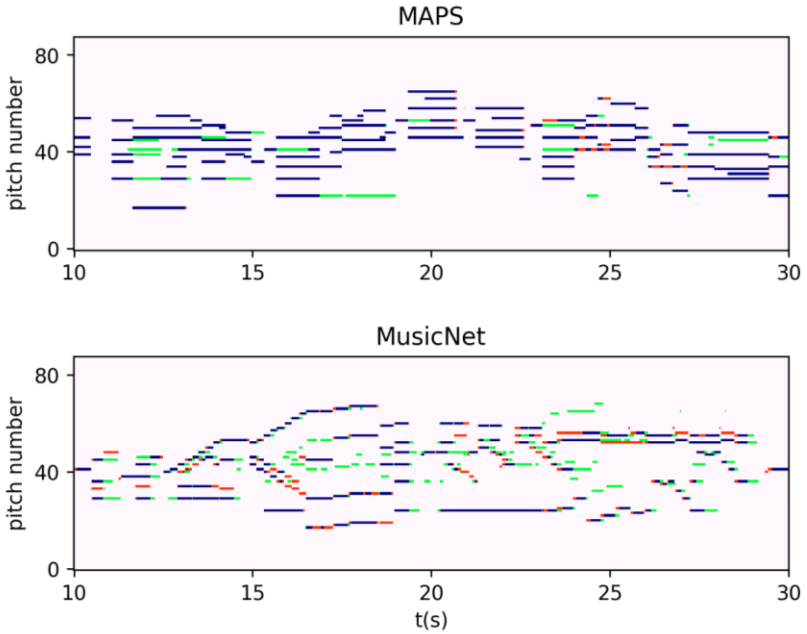
- Multi-pitch estimation (MPE):
  - collectively estimate pitch values of all concurrent sources at each individual time frame, without determining their sources (Multi-F0)
- Note tracking (NT):
  - estimate continuous segments that typically correspond to individual notes or syllables 難：弦樂四重奏中，第一部和第二部小提琴分離
- Timbre tracking, streaming:
  - stream pitch estimates into a single pitch trajectory over an entire conversation or music performance for each of the concurrent sources

AMT: past and now

- Over 40 years of development
  - From simple to complex
  - From monophonic to polyphonic
  - From single instrument to multi-instrument



Piszczalski et al., in MIT Computer Music Journal (1977)



Wu et al, in IEEE ICASSP (2019)

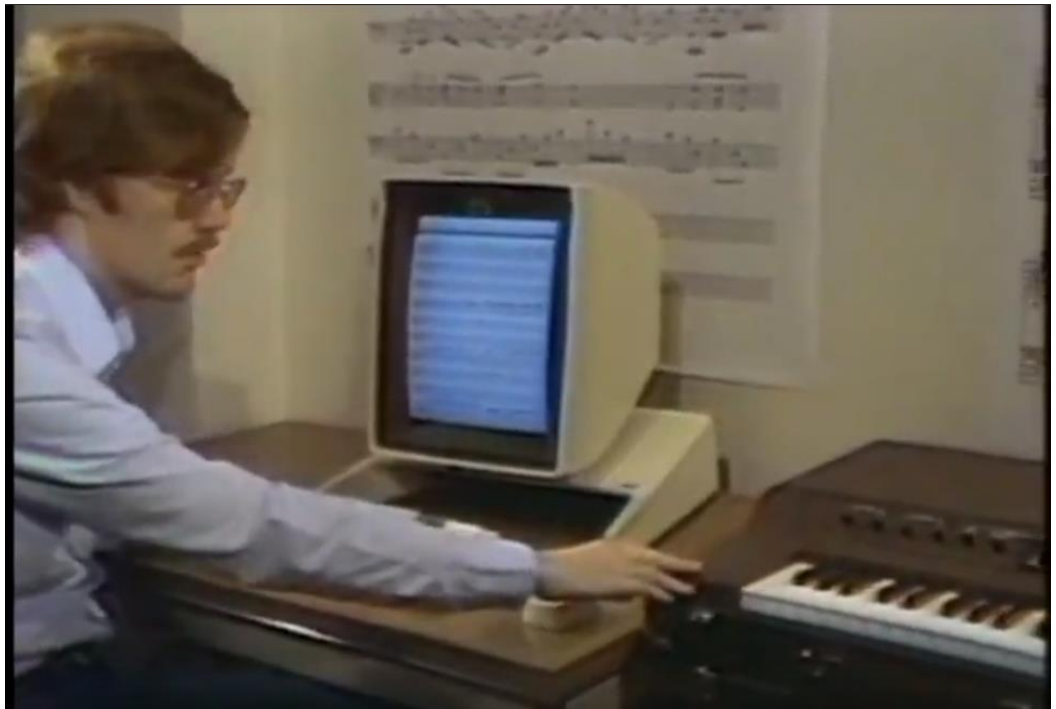
Early AMT systems

- AMT research appears earlier than MIDI
- The first AMT paper
- James Anderson Moorer, On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer, Dissertation, Stanford University (1975)
- Transcribe guitar duet into scores
- Pitch detection and melody grouping

第一篇 AMT

Automatic score-typing systems

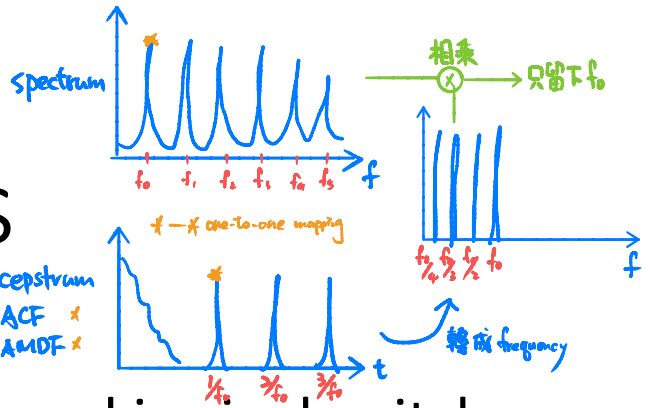
- Mockingbird: a musician's amanuensis (1980)
- https://www.youtube.com/watch?v=0dxaEDKoTys&list=LLzqtAP7sgXppFj2JGQFqnig&index=4&t=368s
- Recording the electric signals



AMT methods

- Core subtask (as an example): multi-F0 estimation
  - Feature-based methods
  - Statistical model-based methods
  - Spectrogram decomposition-based methods
  - Deep learning-based methods

Feature-based methods



Frequency domain　time domain　cepstrum

- Pitch salience functions: from what discussed in single pitch detection (spectrum, AMDF, ACF, etc.)
- Combining frequency and periodicity features: a powerful approach for multi-component signal processing (Peeters, 2006; Su and Yang, 2015; Lin, Su and Wu, 2018)

現在仍很常被用

$$W(t,\omega) = |\mathrm{STFT}(t,\omega)|\mathrm{ReLU}\left(\mathrm{Cepstrum}\left(t,\frac{1}{\omega}\right)\right)$$

Generalized Cepstrum (GC)
$= \mathrm{IFFT}(|\mathrm{FFT}(x)|^{\gamma})$　// $0 < \gamma < 1$, 0.2 or 0.3 常用於 multi-pitch

Combining frequency and periodicity (Peeters, 2006)

- A spectral (*frequency*) feature $U(f)$
  - Fundamental frequency (o) and their multiples (x)
- A temporal (*periodicity*) feature $V(q) = V\left(\frac{1}{f}\right)$
  - Fundamental period (o) and their multiples (x)
  - In other words, fundamental frequency (o) and their sub-harmonics (x)
- Rule of thumb: pitch is the consensus of the fundamental frequency and the fundamental period

$$\hat{f}_0 = \text{argmax}_f U(f) V\left(\frac{1}{f}\right)$$

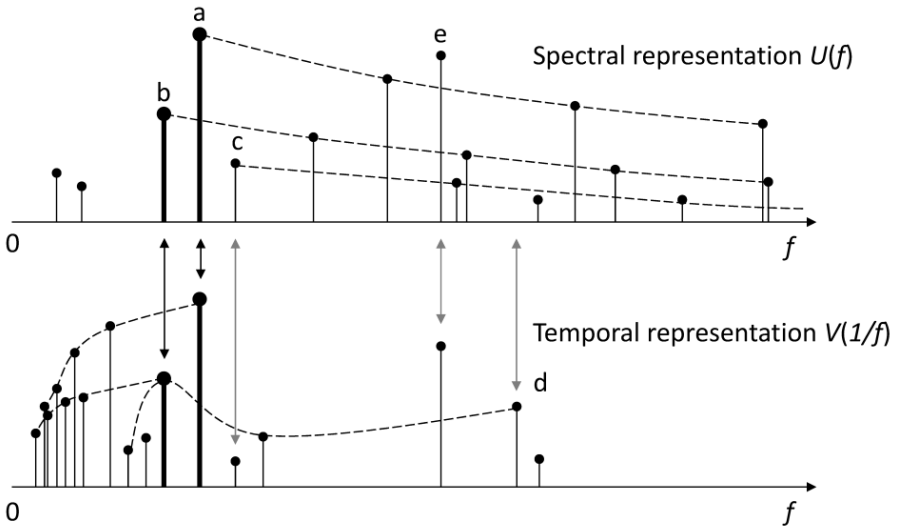
- How about multiple components?

Li Su 覺得實用的 feature representation

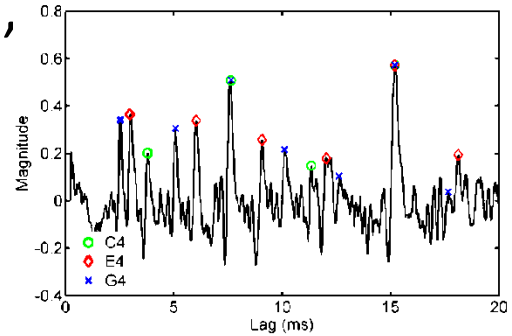
Combining frequency, periodicity, and harmonicity (Su and Yang, 2015)

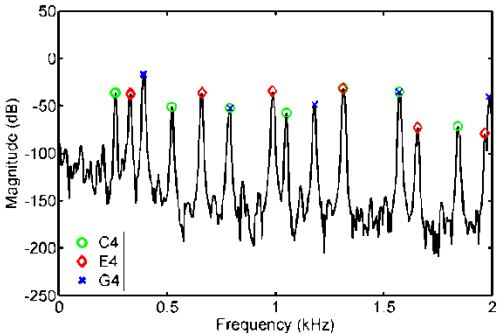
- $U(f)$: magnitude spectrum
- $V(q)$: generalized cepstrum

- The presence of a true pitch at frequency f_0 satisfies the constraints of harmonicity:
  1. A sequence of peaks found around $U(f_0), U(2f_0), U(3f_0), \ldots$
  2. A sequence of peaks found around $V(q_0), V(2q_0), V(3q_0), \ldots$
  3. $f_0 = 1/q_0$



加上 harmonics 必需存在的條件





Statistical model-based methods

- An observed frame $\mathbf{x}$, a set $\boldsymbol{C}$ of all possible F0 combinations
- Maximum a posteriori (MAP) estimation (Emiya et al. 2010):

$$\hat{c}_{MAP} = \underset{c \in \boldsymbol{C}}{\operatorname{argmax}} P(c|\mathbf{x}) = \underset{c \in \boldsymbol{C}}{\operatorname{argmax}} \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})}$$

- Example: PreFEst system (Goto, 2004)
  - Each harmonic is modelled by a Gaussian centered at its position on the log-frequency axis
  - Prior information of note mixture: $P(c)$
- If no prior information: maximum likelihood (ML) estimation

$$\hat{c}_{ML} = \underset{c \in \boldsymbol{C}}{\operatorname{argmax}} P(\mathbf{x}|c)$$

Spectrogram decomposition (2010~2015)

dictionary-based 方法

- 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

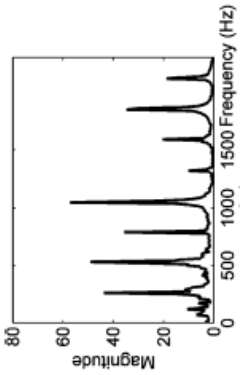
$$\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$$

Basic template matching: single pitch detection

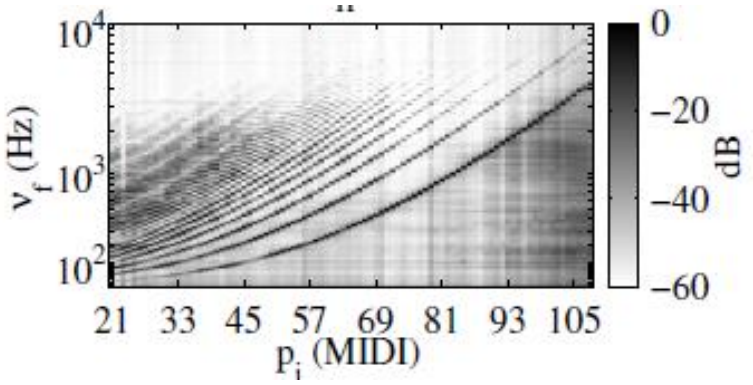
- Input $\mathbf{x}$, dictionary $\mathbf{D} = [d_1, d_2, \cdots, d_{88}]$, each d_k represents one pitch (e.g., d_1 is the spectral pattern of A0, d_{40} is the spectral pattern of C4)
- Find a d_k such that $\mathbf{x} \cdot d_k$ is maximum
- Vector quantization (VQ): "sparsest" approximation

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

single pitch



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?

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

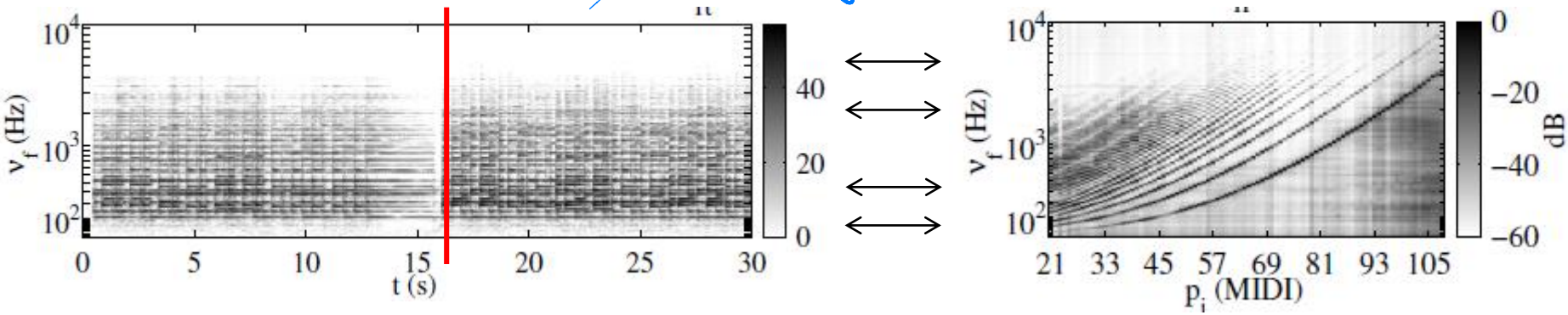
Interpretable

zero norm (非零的個數)

sparse coding / compressed sensing

以 MAPS 為例，同時出現的音大概按下 3~4 音

Beethoven symphony 同時出現約 6 個不同的音



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

Sparse coding

- Perfect reconstruction

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

- Approximation, hard constraint

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

- Approximation, soft constraint

$$\text{minimize} \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2 + \lambda\|\boldsymbol{\alpha}\|_0$$

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: $\min \|\boldsymbol{\alpha}\|_0 \text{ s.t. } \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ …	Signal recovery Feature selection …

Some points in sparse coding

n > 2m
多我几个牌子的钢琴
dictionary 就可以很大

- 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)

$$\text{argmin}_{\boldsymbol{\alpha}} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2 + \lambda\|\boldsymbol{\alpha}\|_1$$

所有 element 的绝对值的和

State of the art

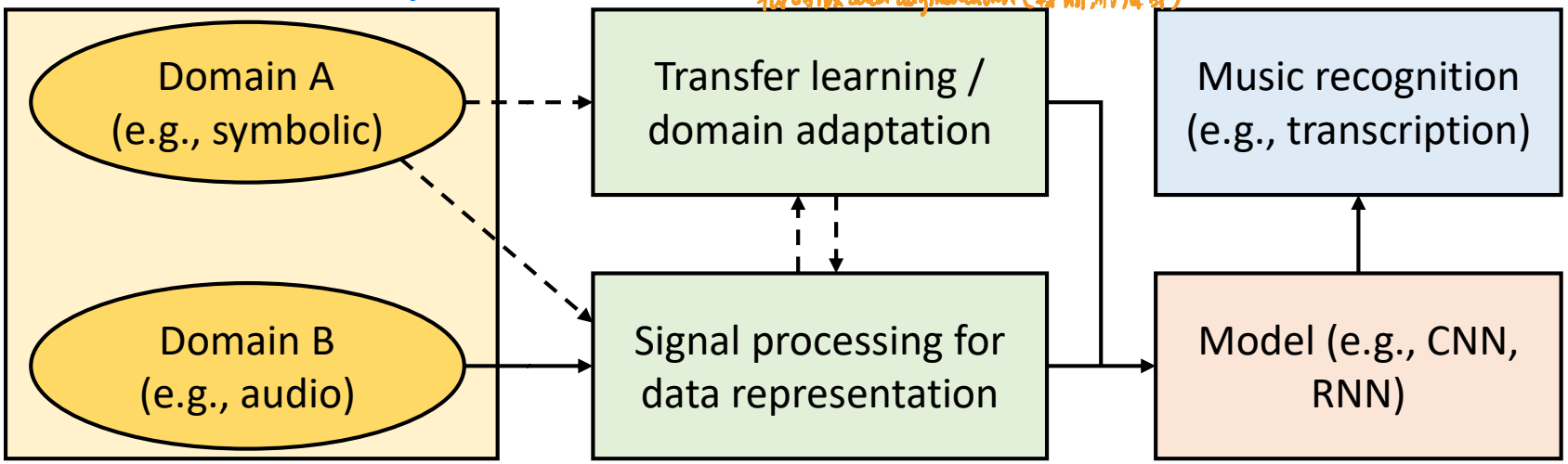
piano transcription 看 Google Magenta

- Example: multi-F0 estimation
- Top 8 in the history of MIREX multi-F0 campaign

Rank	Accuracy	Team	Method
1	0.723	Elowsson and Friberg (2014)	Deep learning
2	0.720	Thickstun et al. (2016)	Deep learning
3	0.692	Yeh and Roebel (2010)	Feature-based
4	0.680	Dressler (2014)	Feature-based
5	0.662	Benetos and Weyde (2013)	Matrix decomposition
6	0.635	Su and Yang (2014)	Feature-based
7	0.569	Duan et al. (2009)	Statistical model
8	0.561	Fuentes et al. (2012)	Matrix decomposition

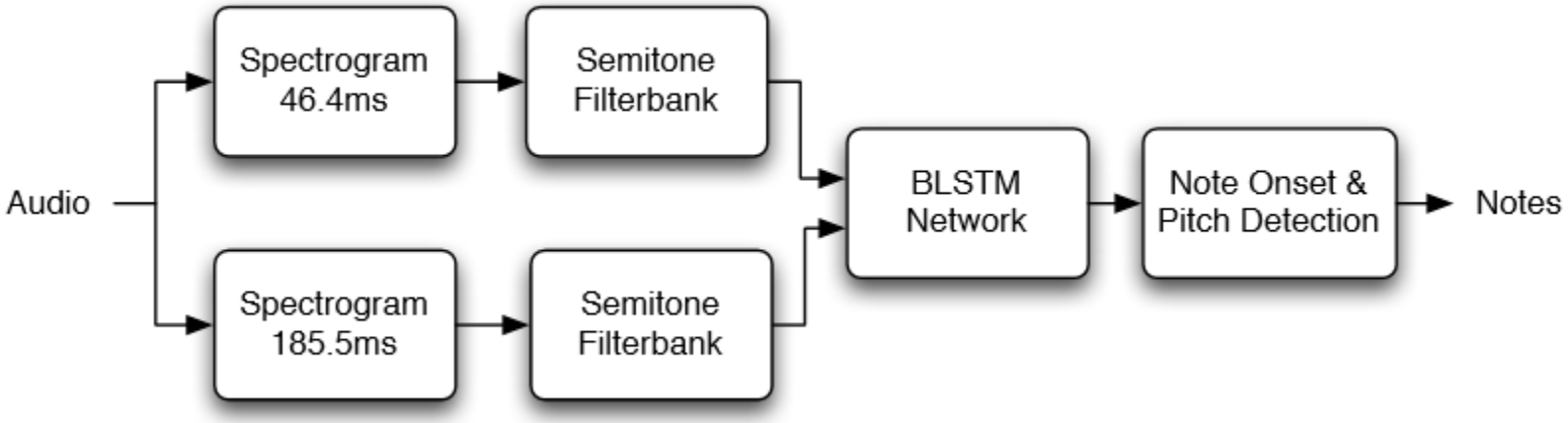
Why deep learning wins

- Flexibility
  - Architecture: information flow control
  - Data representation: multi-view signal characteristics
  - Objective functions: multi-task learning
  - Domain adaptation: from symbolic to audio domain

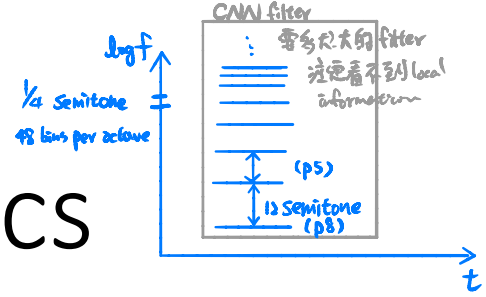


Input feature (1): resolution

- Image data: RGB channels -> multi-channel representation in audio data?
- (Boeck et al., 2012) uses multi-resolution spectrogram and spectral difference to break the performance in onset detection and beat tracking using deep learning
- Multiple ‘views’ of signals: representations depending on window size, window type, etc.



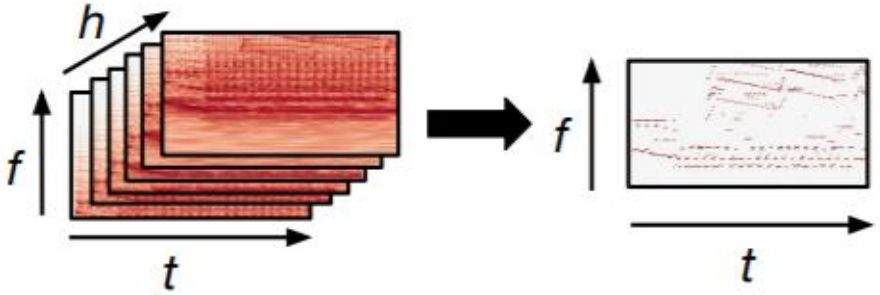
Input feature (2): harmonics

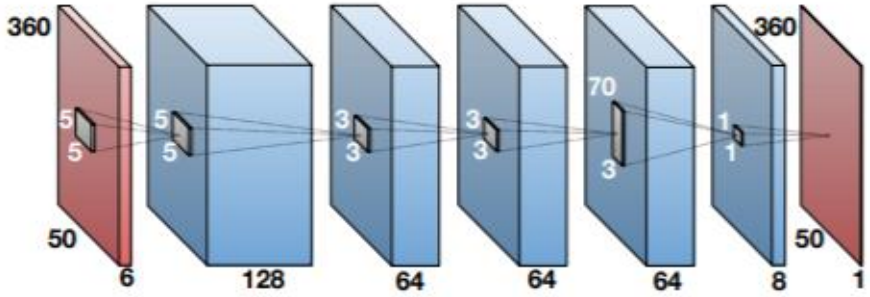


第2個channel為原CQT向下shift 8 pitches 同理亦可用於cepstrum 使filter不用大

- (Bittner et al., 2017) proposed the harmonic constant-Q transform (HCQT): CQTs with different minimal frequencies (f_{min}) positioned at a harmonic series
- (Wu et al., 2019) extends it to cepstral features

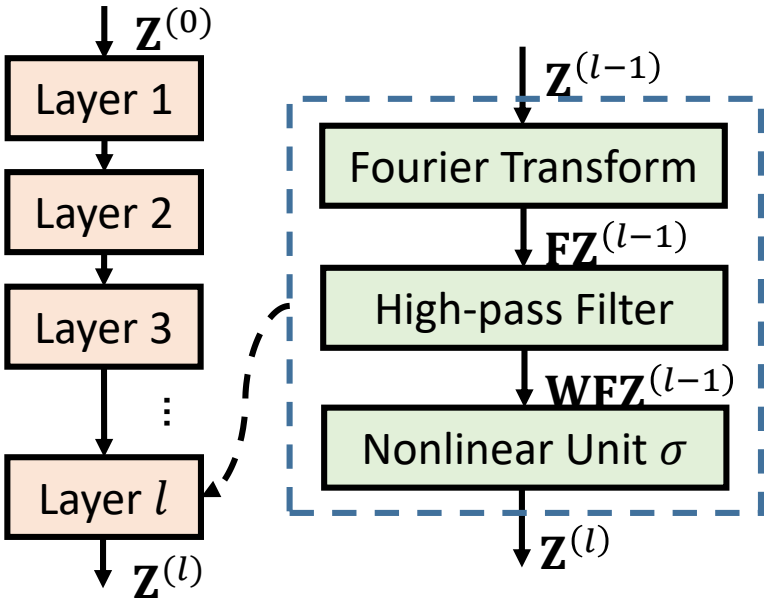
$$\mathbf{Z}_f^{(m)}[k,n] := \mathbf{Z}_f\left[k+\eta\left(m\right)\cdot\delta, n\right]$$
$$\mathbf{Z}_q^{(m)}[k,n] := \mathbf{Z}_q\left[k-\eta\left(m\right)\cdot\delta, n\right]$$
$$\eta(m) := \mathrm{round}(12\log_2 m)$$

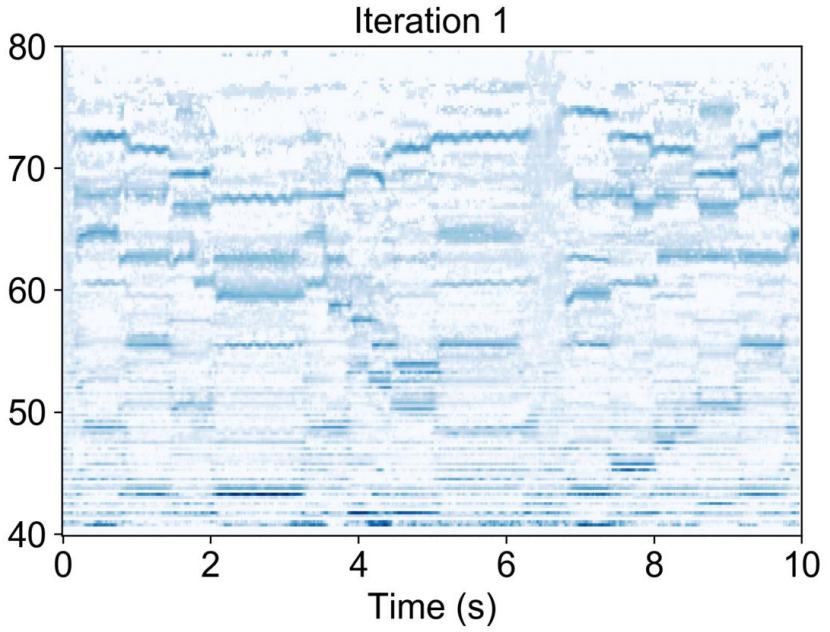




Input feature (3): cepstrum

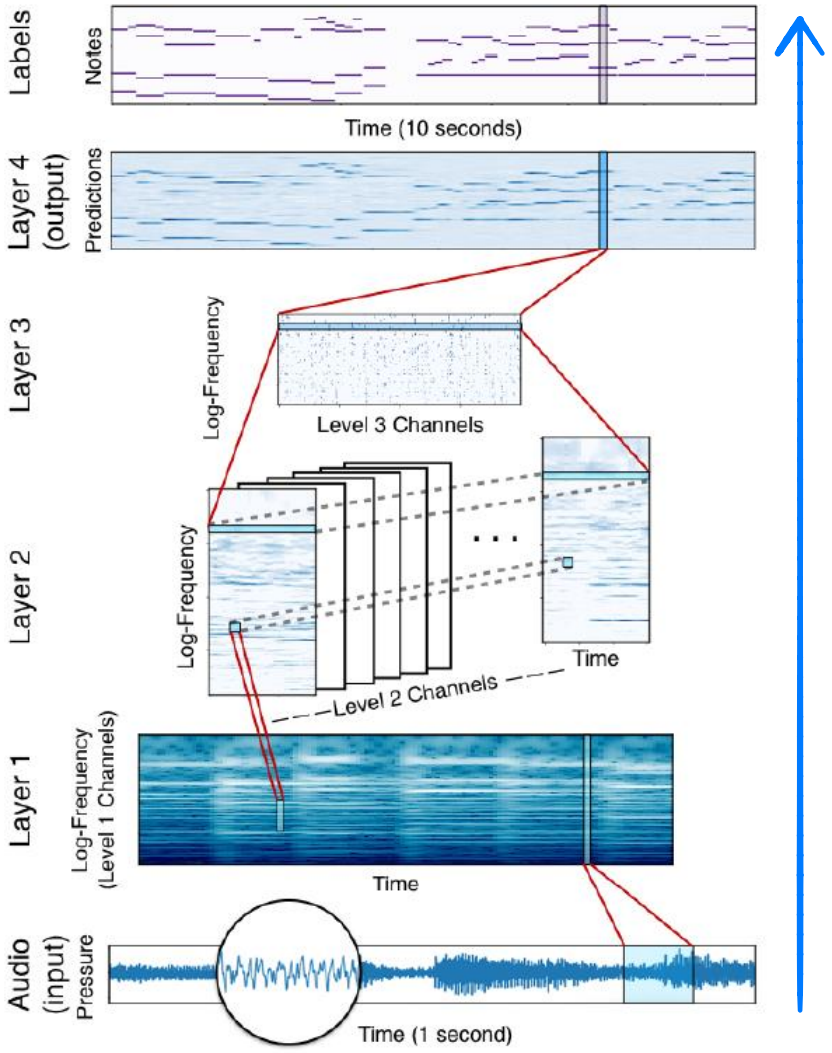
- Frequency- and periodicity-based features for deep learning
- (Su and Yang, 2015) CFP: harmonic suppression!
- (Yu and Su, 2018) Multi-layered cepstrum (MLC): refining pitch saliency by taking Fourier transform again and again





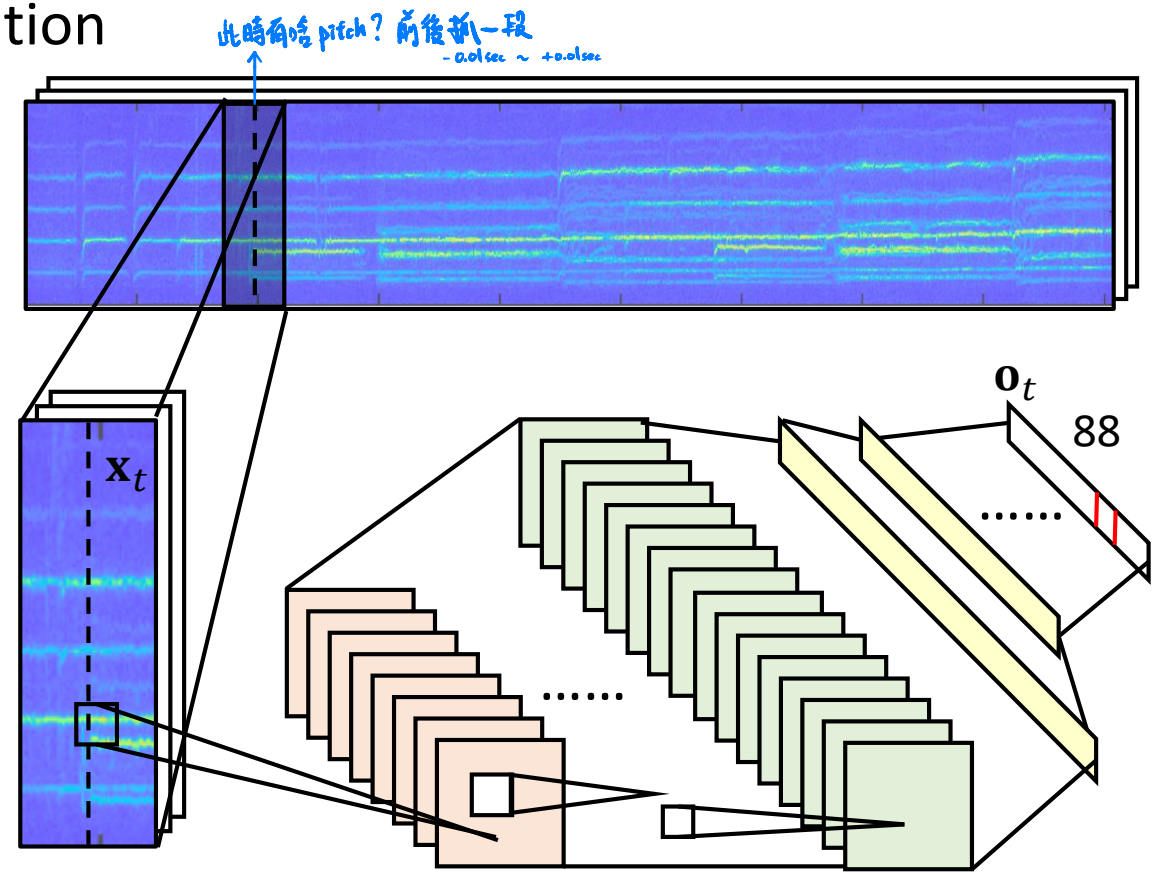
Input feature (4): raw data

- (Thickstun et al., 2018) translation-invariant network: directly learn useful kernels on a long signal (16,384 samples) for filterbank



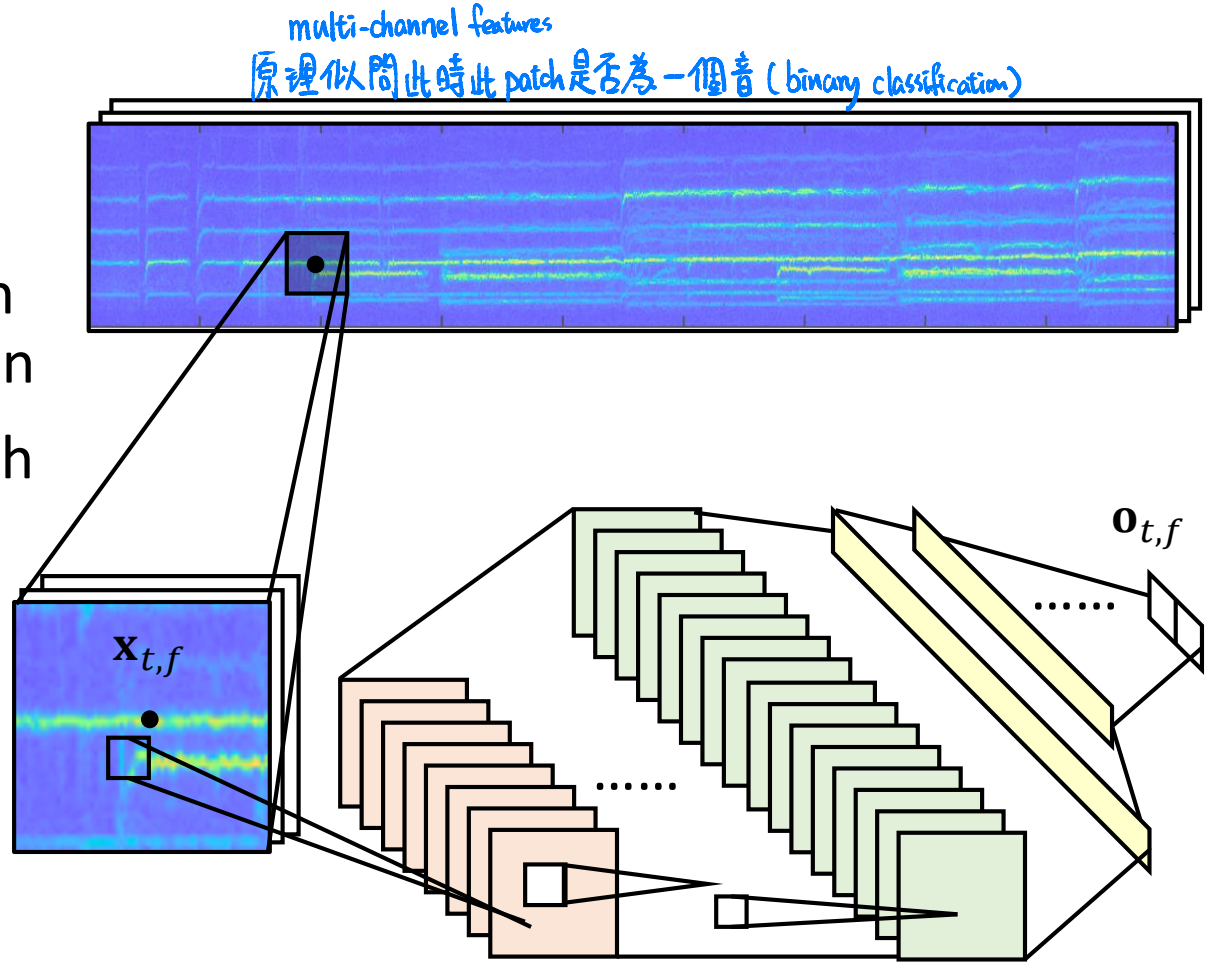
Network architecture (1): frame

- The most widely used architecture
- Multi-label classification
- Network: N
- $\mathbf{o}_t = N(\mathbf{x}_t)$



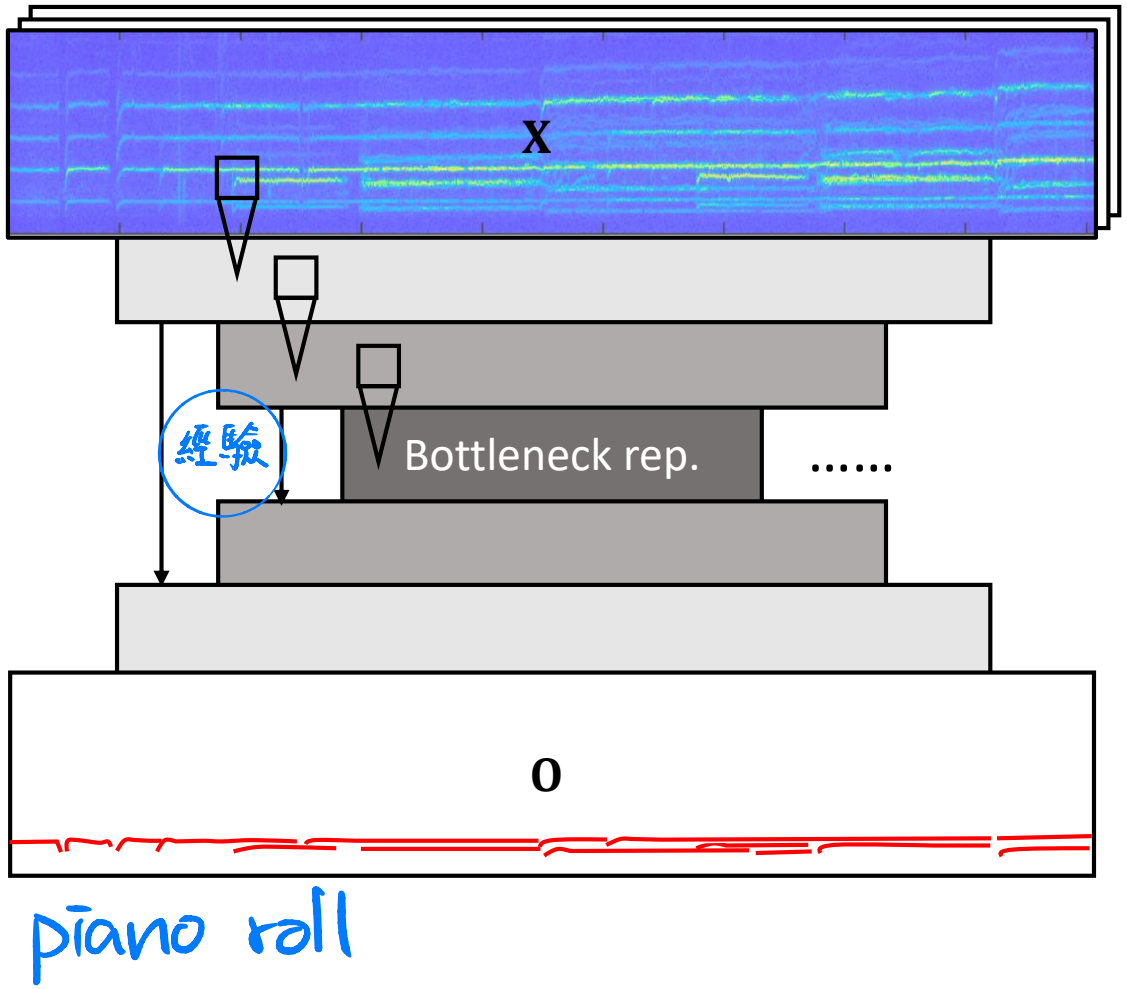
Network architecture (2): patch

- (Su, 2018)
- Local, binary detection: focus on local representation
- Activation and pitch detected jointly
- $\mathbf{o}_{t,f} = N(\mathbf{x}_{t,f})$



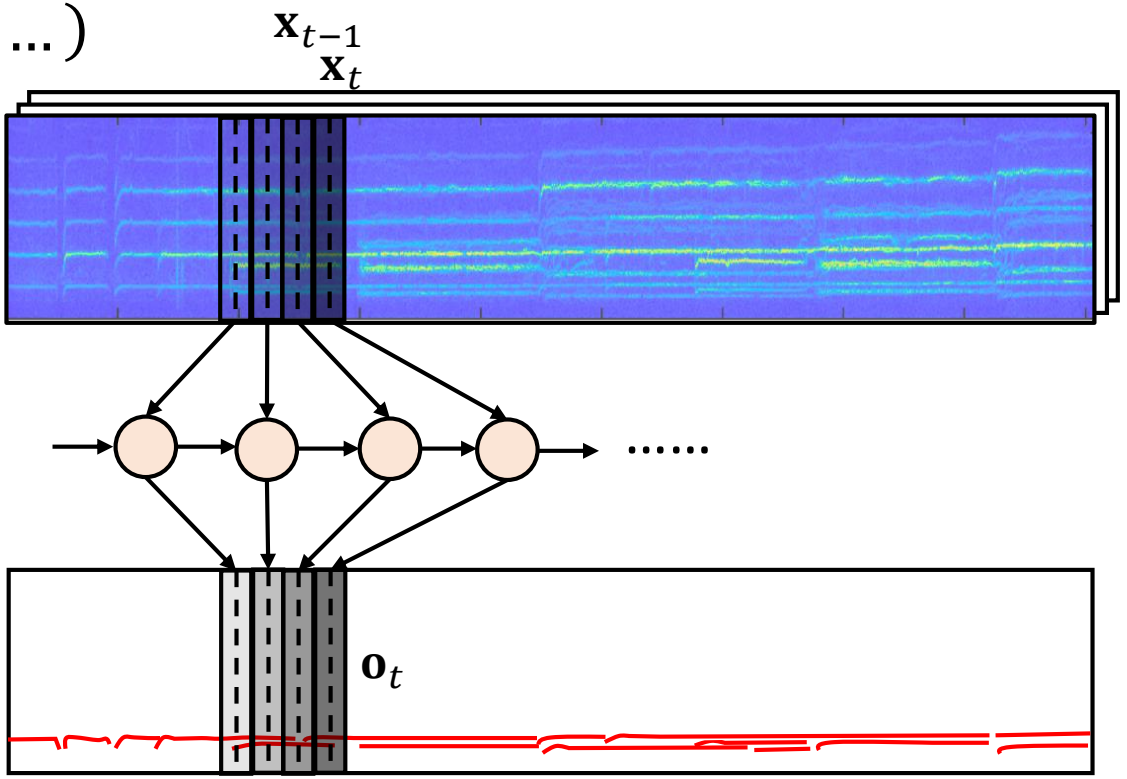
Network architecture (3): image

- (Bittner et al., 2017; Lu and Su, 2018; Wu et al., 2019)
- U-net (Olaf et al., 2015): semantic segmentation in computer vision
- Segmenting pitch contours with image-to-image processing
- $\mathbf{O} = N(\mathbf{X})$



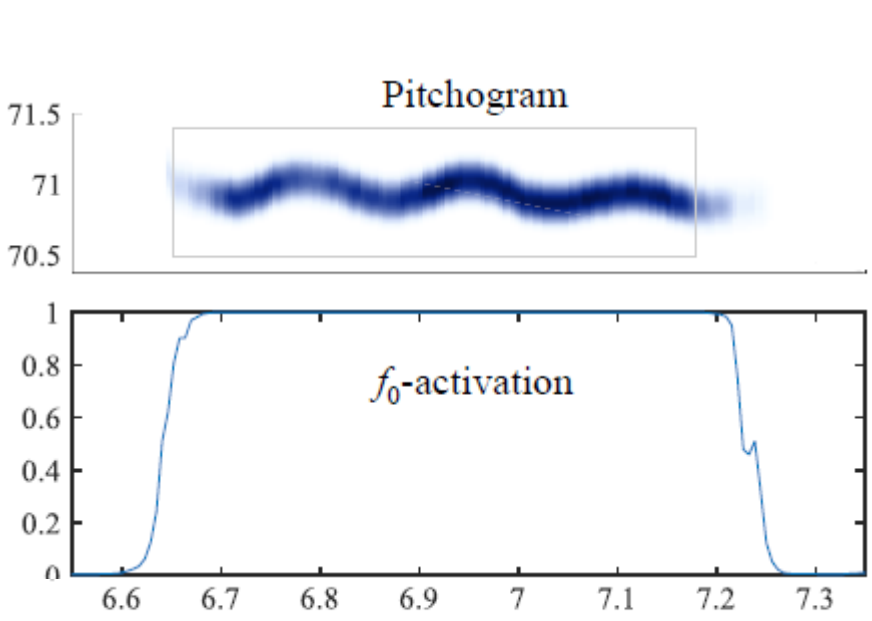
Network architecture (4): sequence

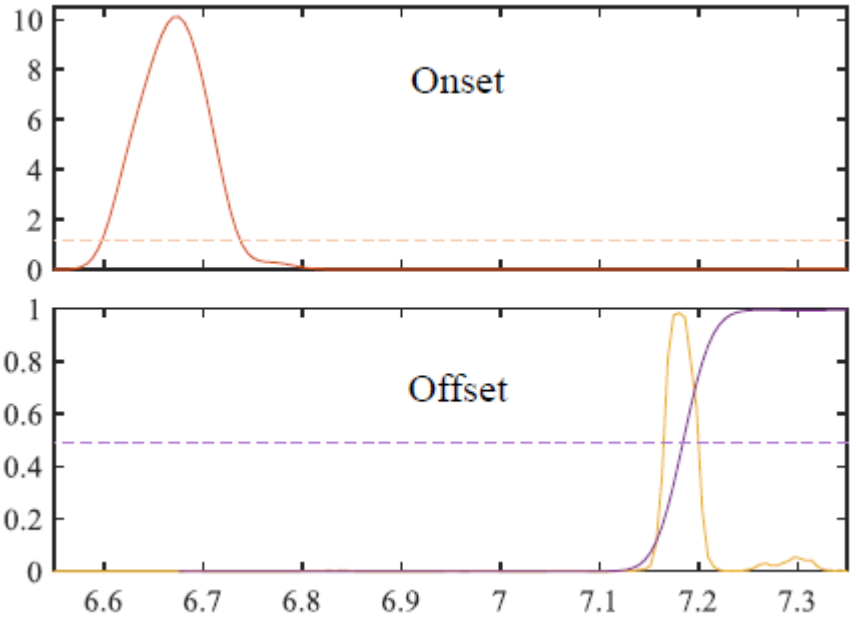
- Recurrent neural networks (Boeck et al., 2014; Ycart and Benetos, 2017), music transformer (Hawthorne et al., 2018) , …
- $\mathbf{o}_t = N(\mathbf{x}_t, \mathbf{x}_{t-1}, \dots)$



Objective function (1): multi-state

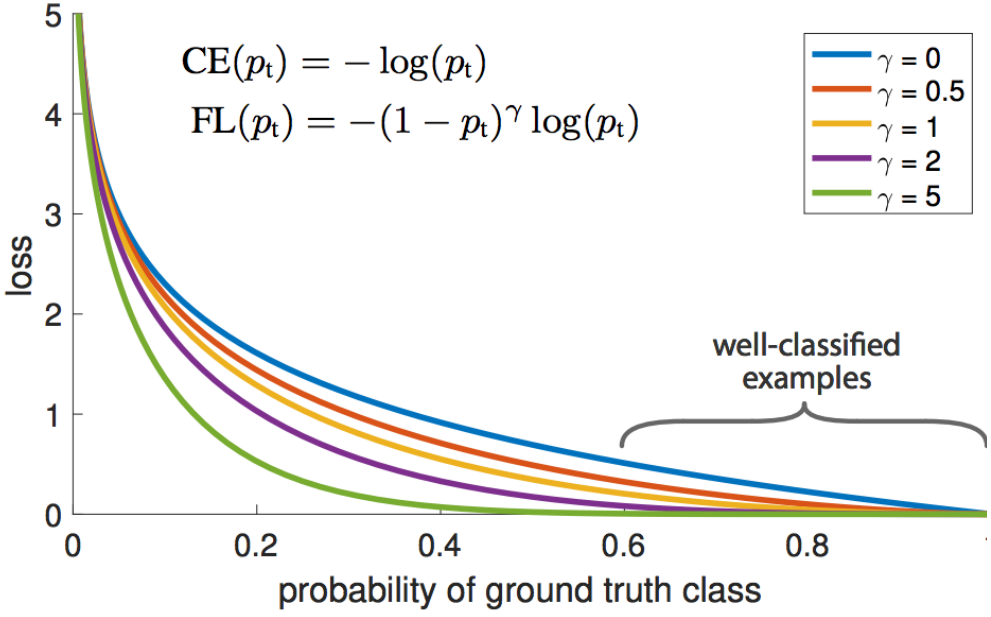
- (Elowsson, 2018) uses two networks to detect note offsets: 1) offset curve and 2) offset detection activation
- (Fu and Su, 2019) uses hierarchical classification approach
- One event, multiple states (and therefore multi-objectives)

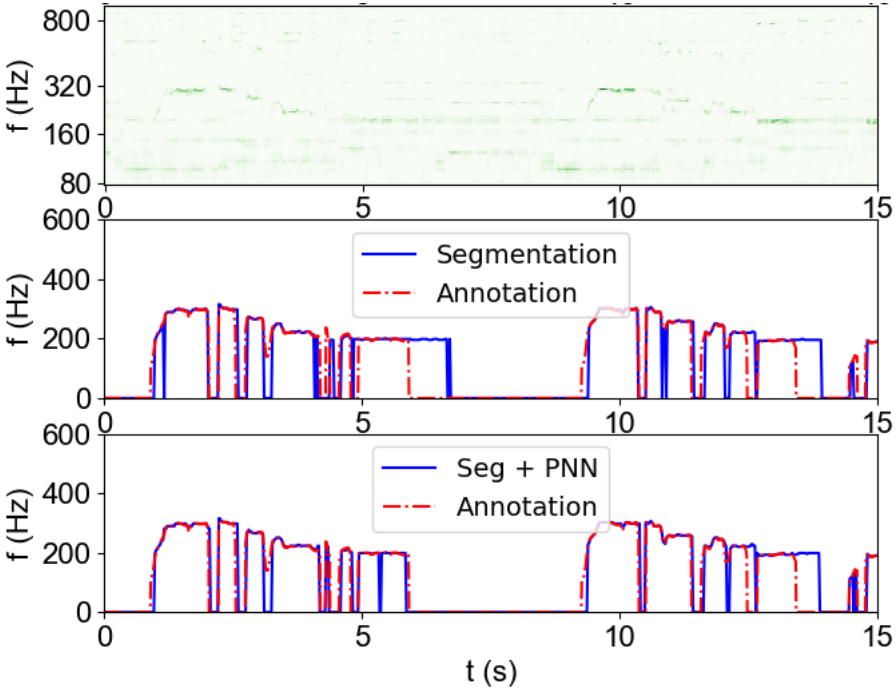




Objective function (2): focal loss

- Some objects (e.g., melody contours) only occupy a small portion of the time-frequency image -> class imbalance!
- (Lin et al., 2017) focal loss for dense object recognition
- (Lu and Su, 2018) focal loss for melody extraction

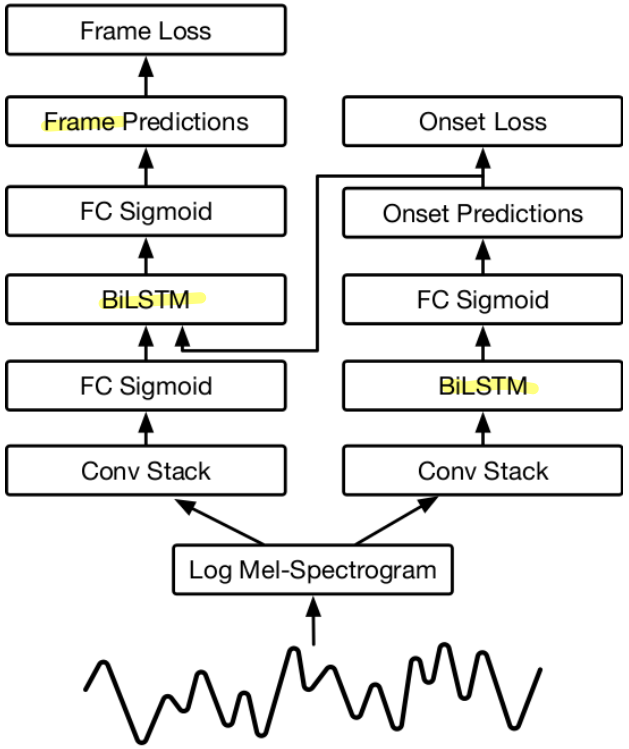


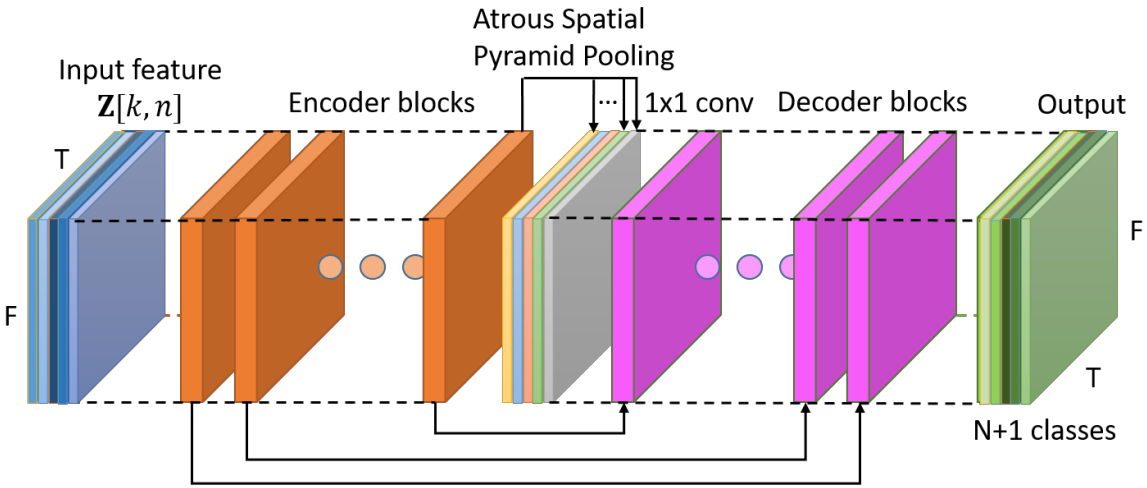


Objective function (3): multi-task

By Google Magenta 建議先看

- (Hawthorne et al., 2018) onsets and frames
- (Bitter et al., 2018; Wu et al., 2019) use multi-task loss for multi-instrument music transcription



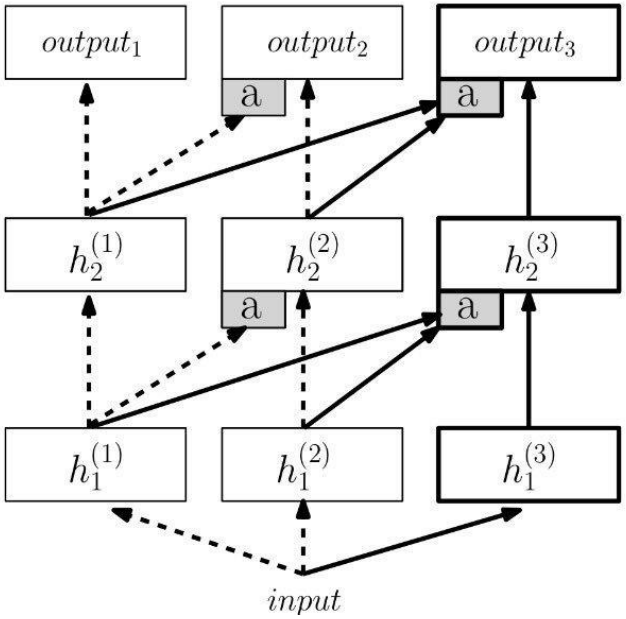


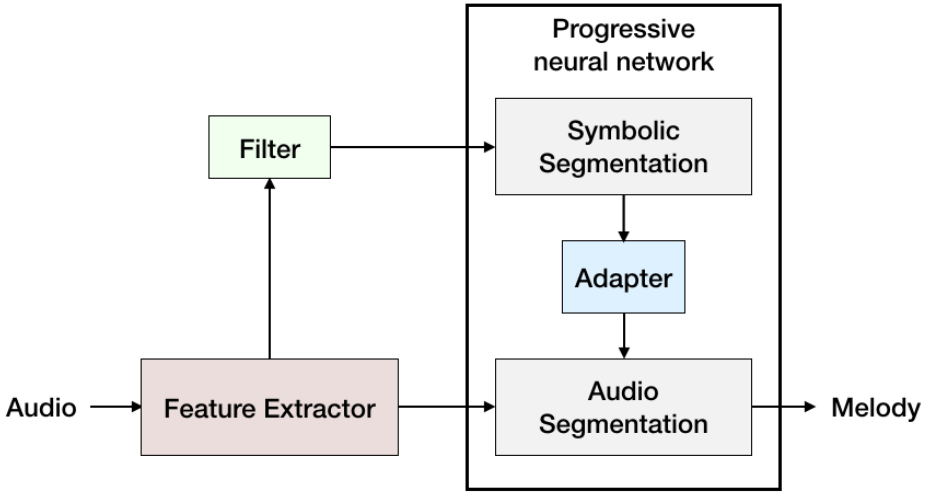
Domain adaptation

目前做的人還不多

- Transfer learning from symbolic to audio domain: 1) capture high-level semantics and 2) enhance data augmentation
- (Lu and Su, 2018) uses *progressive neural networks* (Andrei et al., 2016) for melody extraction

找主旋律

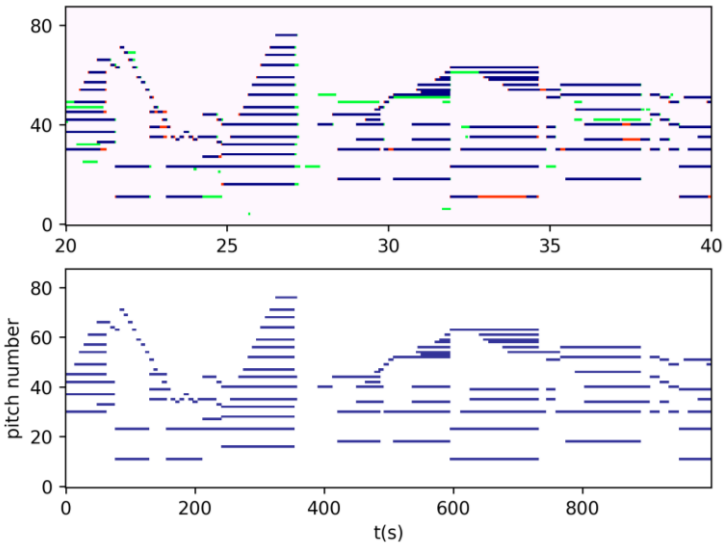




Benchmarks

MAPS 用自動彈奏鋼琴做的

- MIREX multi-F0 campaign
  - MIREX dataset
  - Su dataset 我 piano expert 在電子琴上 跟表名醫部
- MAPS dataset
  - Configuration I
  - Configuration II



Performance of Configuration II,MAPS dataset

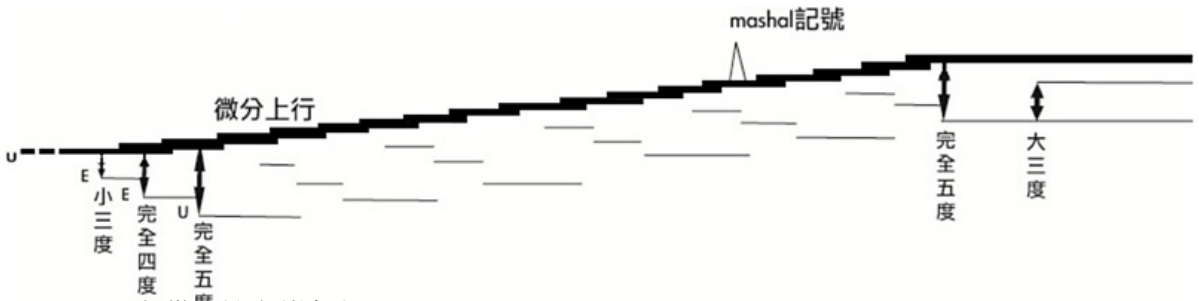
	P	R	F1
Hawthorne et al. in ISMIR 2018 (Google Meganta)	88.53	70.89	78.30
Kelz et al. in ISMIR LBD 2018	90.73	67.85	77.16
Hawthorne et al. in ICLR 2019 (Google Meganta)	92.86	78.46	84.91
Yu-Te Wu, Berlin Chen and Li Su in ICASSP 2019	87.48	86.29	**86.73**

有提出一個dataset

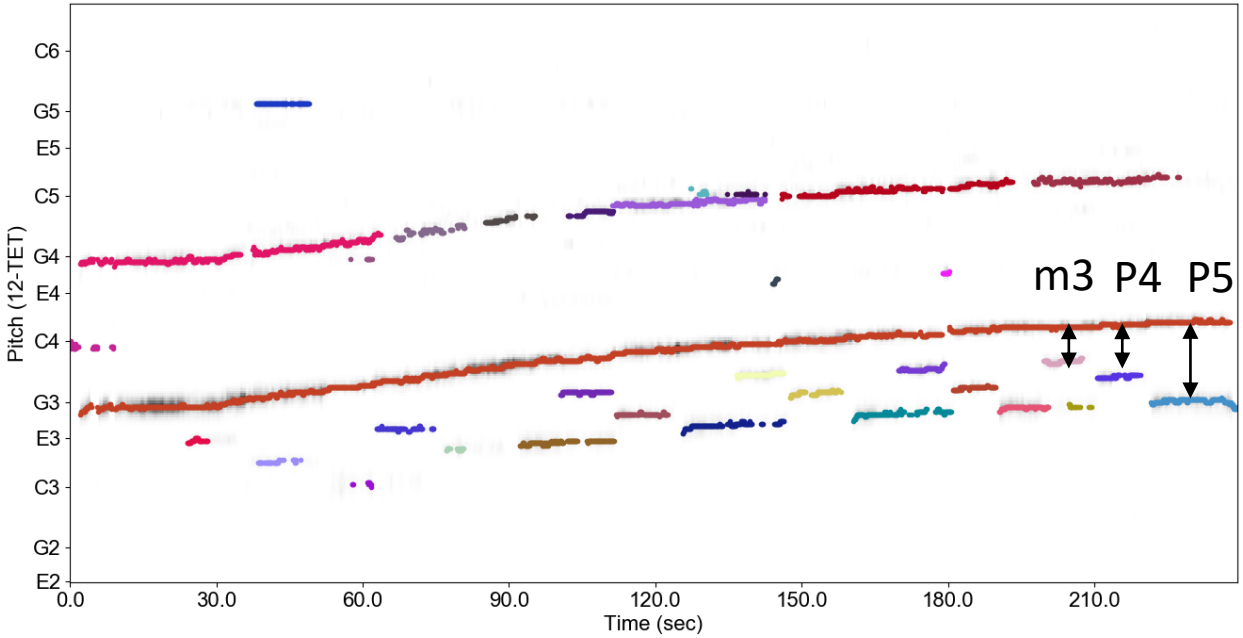
Transcribing Pasibutbut

布農族八部合音

- Polyphonic music in Bunun tribes
- “Praying for a Rich Harvest”
- Performance analysis of microtone, polyphonic singing

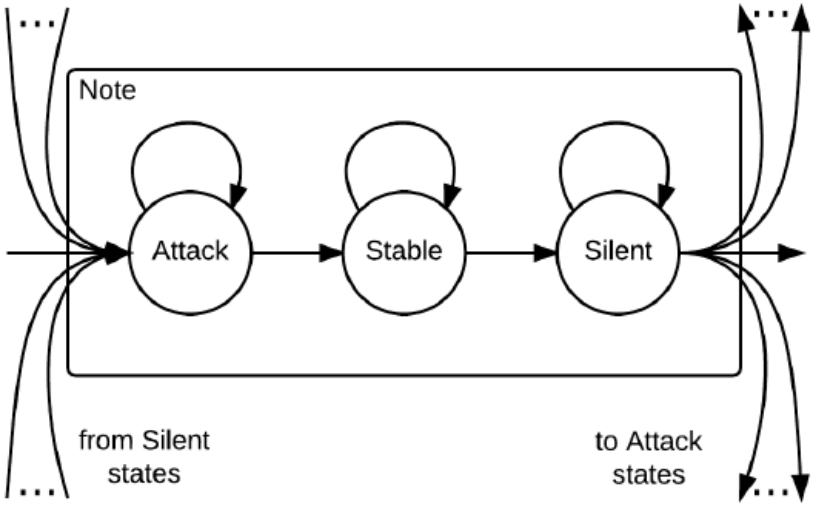


From:2010 臺灣傳統音樂年鑑
http://rimh.ncfta.gov.tw/2010_taiwan_traditional_music_yearbook/tw/review/review_f3.html

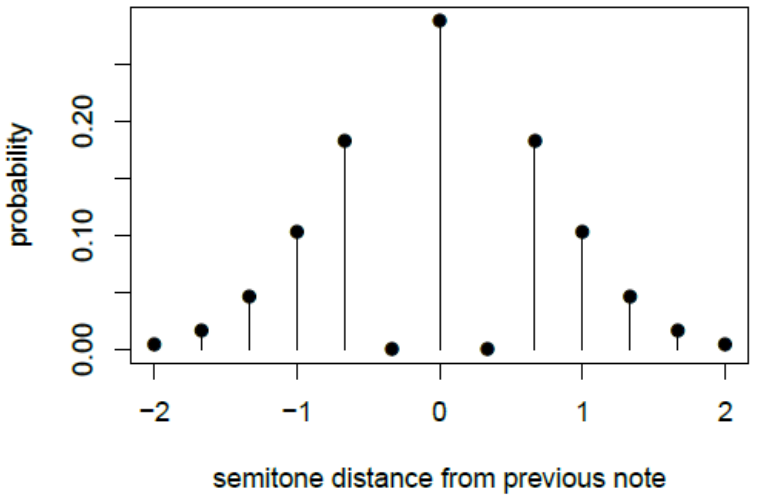


Note tracking (1)

- Hidden Markov Models (HMM) as post filtering
- Example: Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency (Mauch et al., 2016)



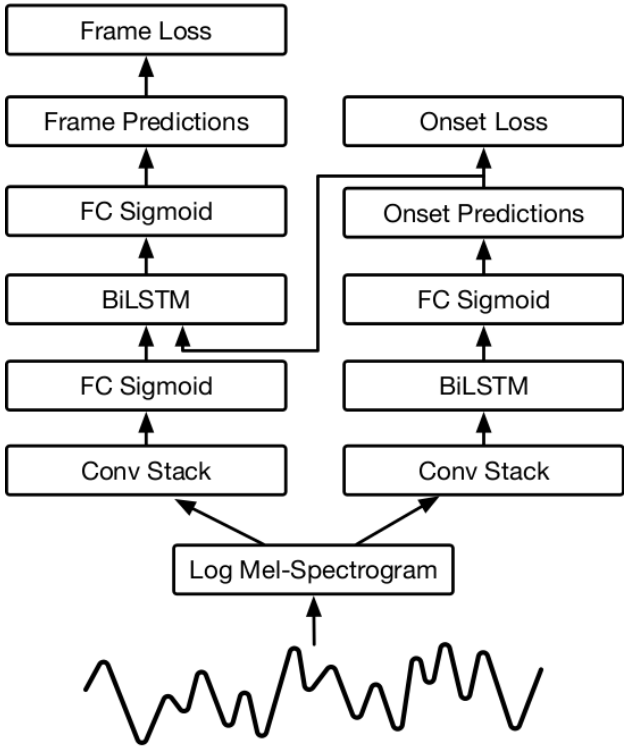
(a) Excerpt of the pYIN note transition network.



(b) Central part of the note transition probability function.

Note tracking (2)

- Multiple objective functions
- Example: Onsets and Frames from Google Magenta Team



$$L_{total} = L_{onset} + L_{frame} \quad (1)$$

$$L_{onset} = \sum_{p=p_{min}}^{p_{max}} \sum_{t=0}^{T} CE\left(\mathbf{I}_{onset}(p,t), \mathbf{P}_{onset}(p,t)\right) \quad (2)$$

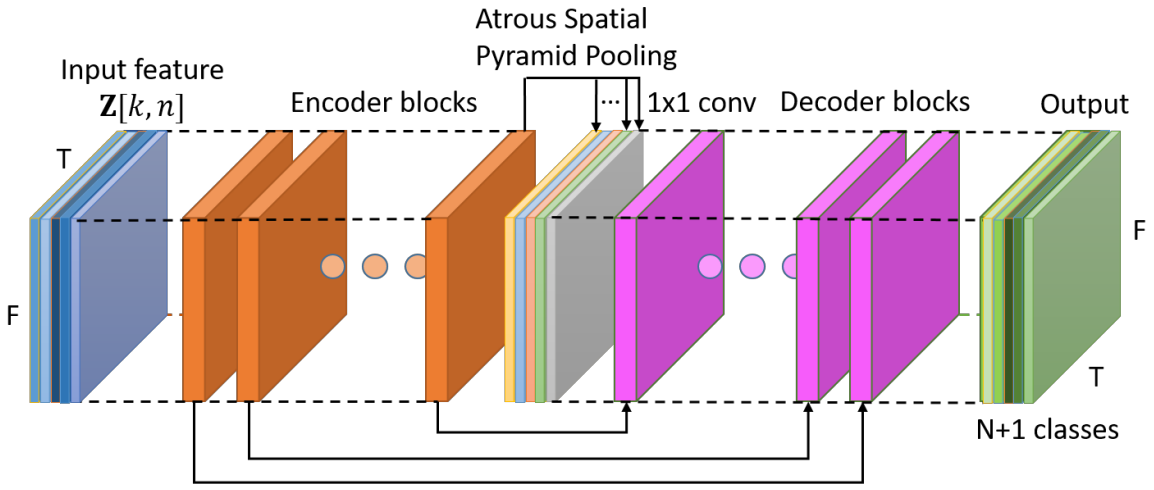
$$L_{frame} = \sum_{p=p_{min}}^{p_{max}} \sum_{t=0}^{T} CE\left(\mathbf{I}_{frame}(p,t), \mathbf{P}_{frame}(p,t)\right) \quad (3)$$

$$L_{frame}(l,p) = \begin{cases} cL'_{frame}(l,p) & t_1 \le t \le t_2 \\ \frac{c}{t-t_2} L'_{frame} & t_2 < t \le t_3 \\ L'_{frame}(l,p) & elsewhere \end{cases} \quad (4)$$

Multi-pitch streaming

用時抓音高與樂器(聲部)
簡單版：一樂器是一聲部
超版：同樂器有好多聲部

- In symbolic music: voice separation
- Constrained clustering of timbre feature
- Multi-task learning: joint detecting pitch and instrument
- Example: Yu-Te Wu, Berlin Chen, Li Su, "Polyphonic Music Transcription with Semantic Segmentation," in ICASSP 2019



Conclusion

- Deep learning wins with its flexibility
- Things before and after learning becomes the focus
  - Before learning: sensors and signals
  - After learning: meanings and applications
- That means, music becomes the focus