CS4347

Sound and Music Computing

L4: Automatic Music Transcription (AMT)

Wang Ye

www.comp.nus.edu.sg/~wangye

wangye@comp.nus.edu.sg

Office: AS6-04-08

Topics to Cover (*selective approach*)

Part A: The Core

- Introduction
- Review of DFT, Audio Representation, and Machine Learning
- Music Representation, Analysis and Transcription
- Automatic Music Transcription (AMT)
- Automatic Speech Recognition (ASR)
- Generative Models for Text-to-Speech (TTS) & Singing Voice Synthesis (SVS)

Midterm break

Part B: The Breadth

- Singing voice processing
- Music production audio effects
- Automatic Music Generation
- Synthesis of sound & music – a DSP approach
- Project presentations/demo

Topics Today

Part A: Automatic music transcription (AMT)

- The AMT task
- Application of AMT systems

Part B: Singing voice transcription

- Signal processing method: Yin
- Neural network methods

Part C: Piano music transcription

- Multipitch challenge
- Frame- and note-level methods

The AMT task

Automatic Music Transcription (AMT)
= Converting audio to notations

Music recordings

Transcription

Music notations

Easier to appreciate

More convenient for analysis

Which form is more important to you? Why?

Draw an analogy between speech and music

Application of AMT systems 1/3

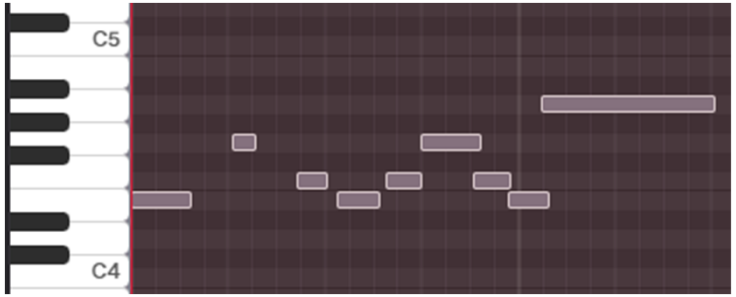
Help to **evaluate singing quality (for entertainment)**

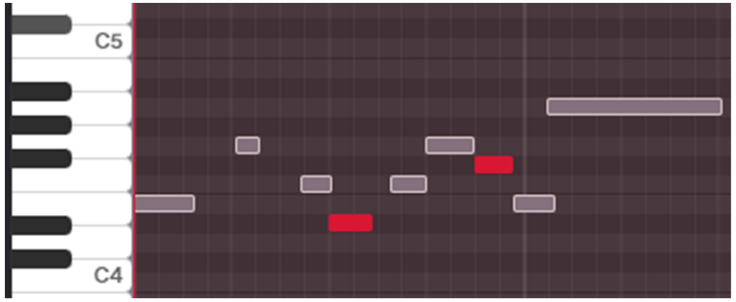
A bit bored?

AMT

Open a karaoke app

Sung Notes





Correct notes

Score:
79/100

That's interesting!

Application of AMT systems 2/3

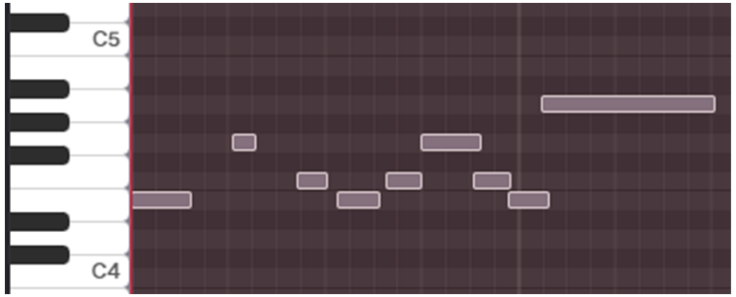
Facilitate **music education**

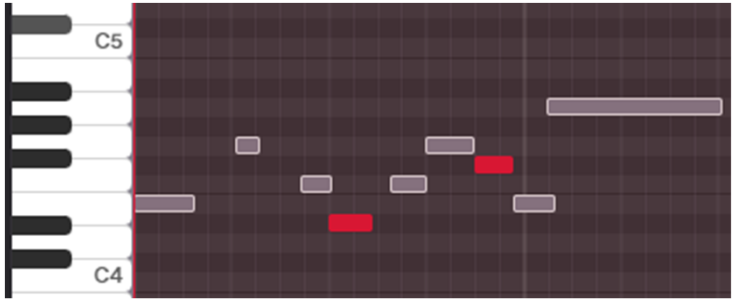
Don’t know how to improve

AMT

Open a guitar tutor app

Played Notes





Correct notes

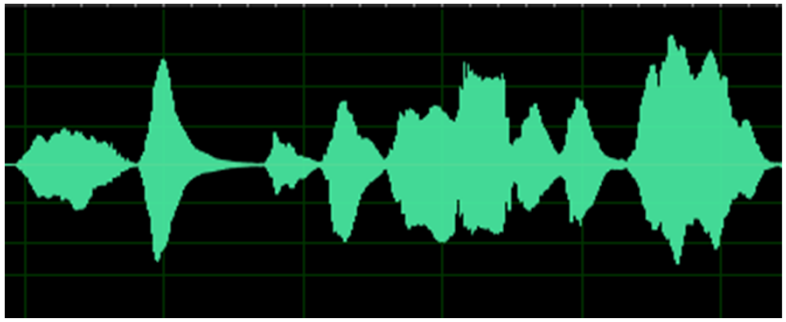
Feedback: The 4th note too high

No wonder!

Application of AMT systems 3/3

Facilitate **audio editing**

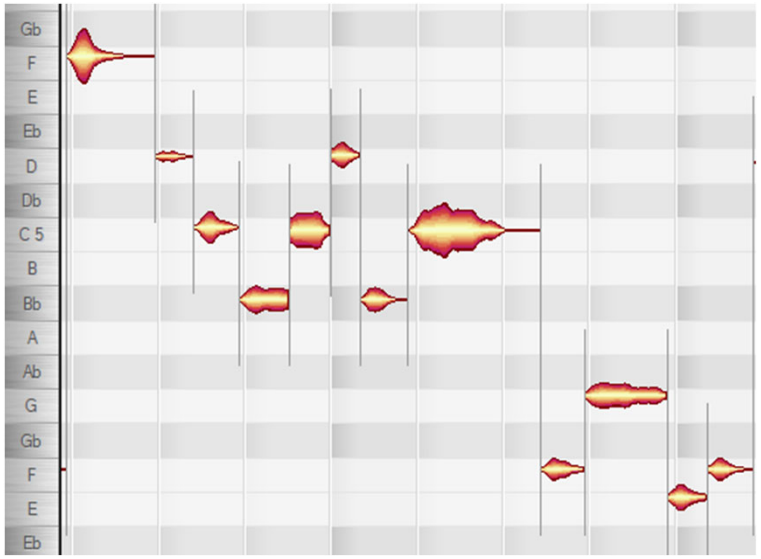
A recording in Audition



- Pitch can't be edited
- Don't know notes' location

AMT

A recording in Melodyne



The audio is visualized to clearly show:

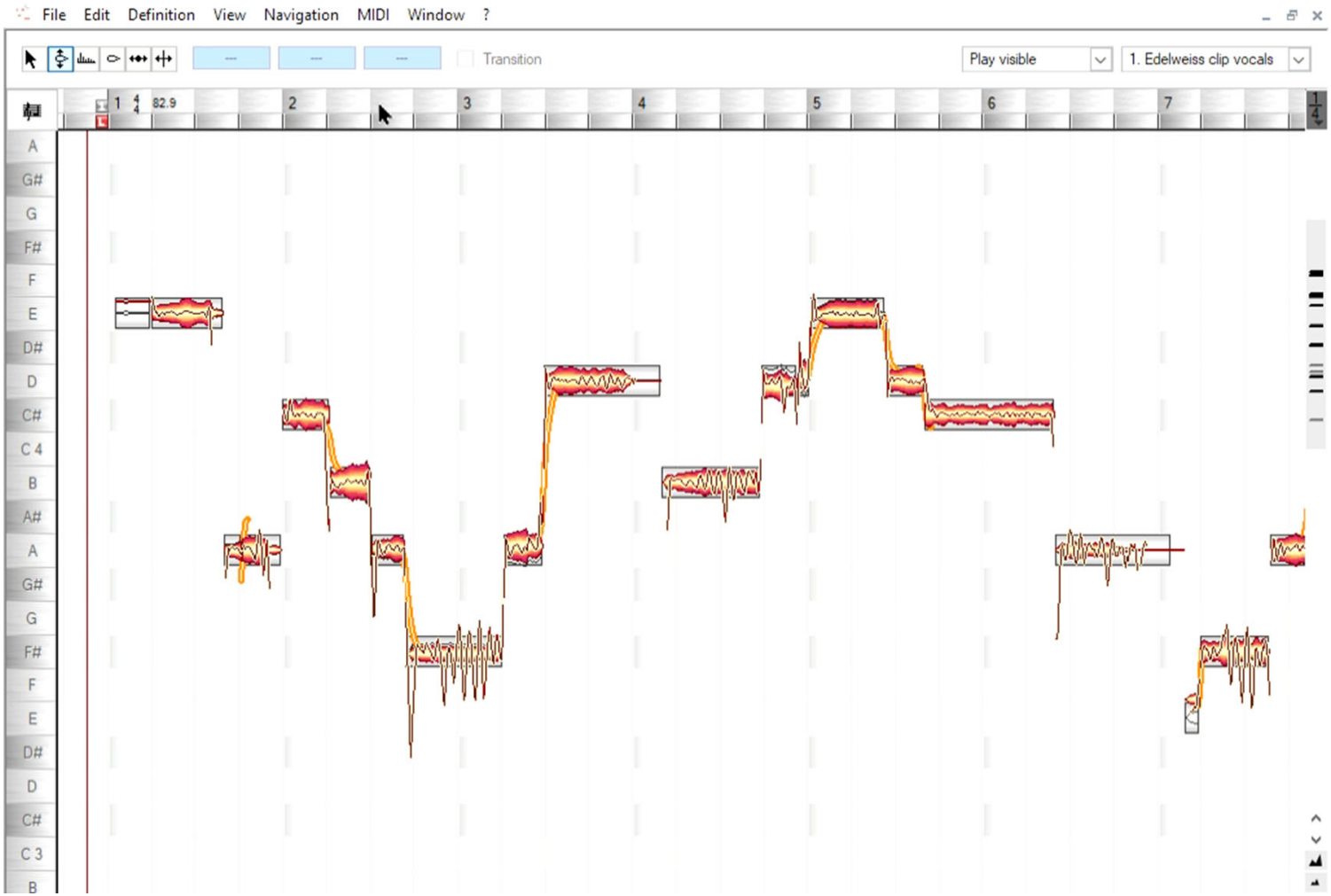
- Where each note starts and ends
- How loud each note is

Then it can be manipulated more easily

Transcription Demo

Recording

Transcription



A quick quiz
On a toy problem

Topics Today

Part A: Automatic music transcription (AMT)

- The AMT task
- Application of AMT systems

Part B: Singing voice transcription

- Signal processing method: Yin
- Neural network methods

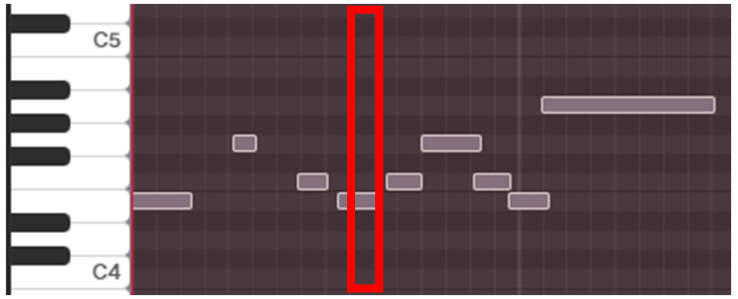
Part C: Piano music transcription

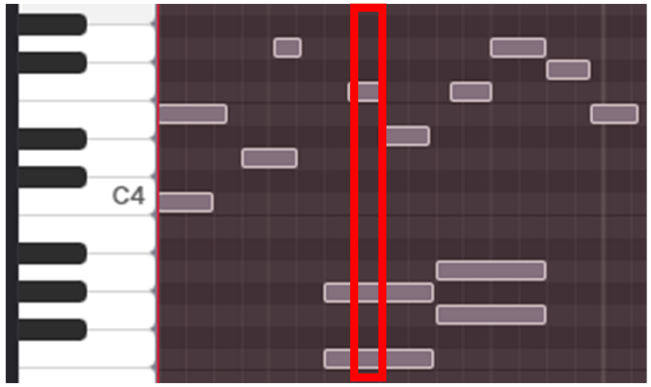
- Multipitch challenge
- Frame- and note-level methods

Let's start from Singing Voice Transcription

In general, **Monophonic** music is easier to transcribe than polyphonic music.

Only one pitch is "voiced" at a time





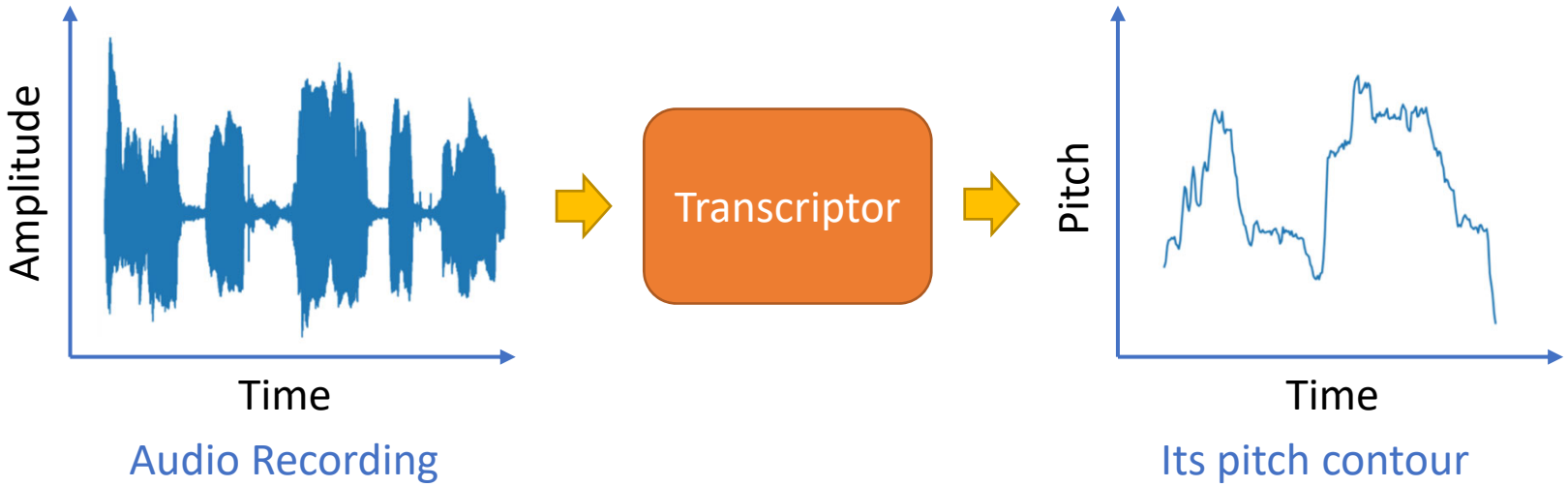
Frame-Level Transcription of Monophonic Music:

- Also be referred to:

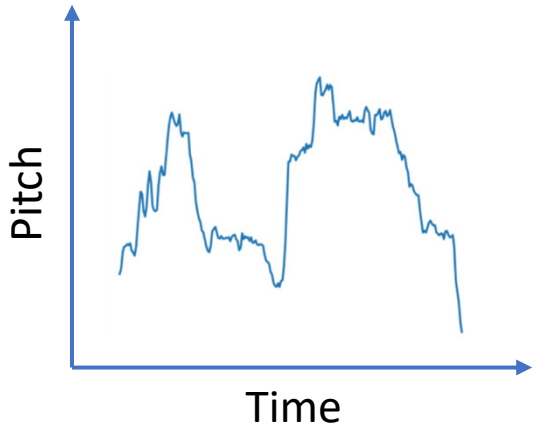
Pitch/F0/melody estimation/detection/tracking/extraction

Singing Voice Transcription (frame-level): Input and Output

- Input: Vocal recording (no accompaniment)
- Output: **Pitch Contour**, the fundamental frequency (or pitch) at every timestep



Pitch Contour



- y-axis can have different unit
  - Frequency in Hz as a real number
  - or Pitch number: an integer in range [0, 127]
    - Quantized pitch values used in the MIDI format

Evaluation Metrics for frame-level transcription

- **Gross Pitch Error (or accuracy)**:

The Gross Pitch Error (GPE) is the proportion of frames, where the decisions of both the pitch tracker and the ground truth are voiced, for which the relative error of F0 is higher than a threshold of 20%.

- Fine Pitch Error:

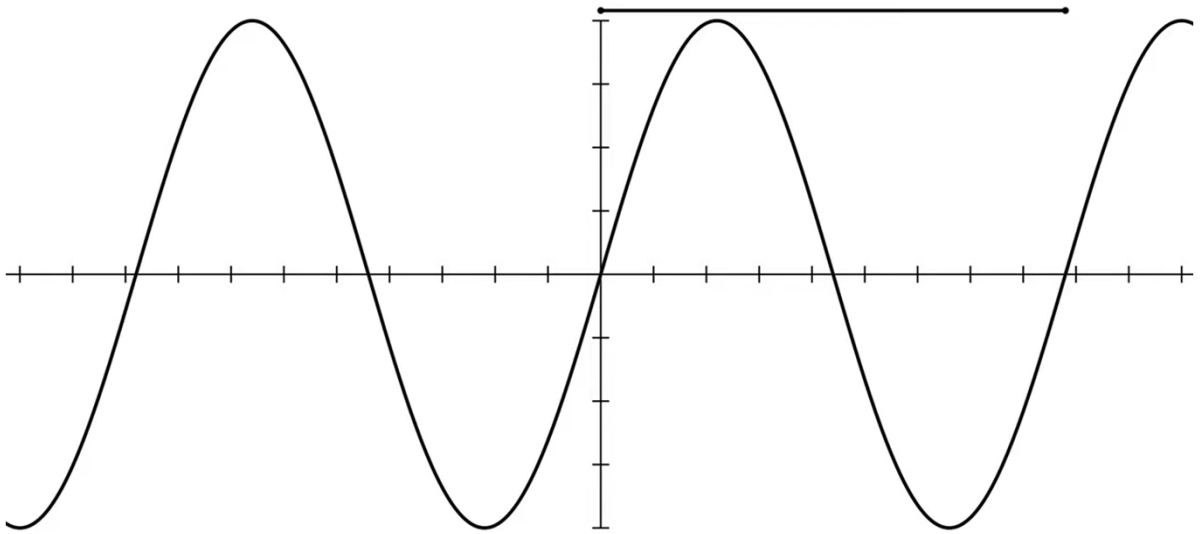
The Fine Pitch Error (FPE) is defined as the standard deviation (in %) of the distribution of the relative error of F0 for which this error is below a threshold of 20%.

T. Drugman, A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics", Interspeech 2011

YIN, a Traditional Method

- **Name**: from oriental philosophy “阴阳”
- Recall a periodic signal, the signal remains unchanged if we shift it by its period

$$T = 9\ samples \Rightarrow f = \frac{1}{duration\ of\ 9\ samples}$$



YIN, a Traditional Method

- The YIN algorithm is a time domain method which is still widely used today. Can we use autocorrelation function here? Pros/cons?

- For periodic signal:

$$x[t] = x[t+T]$$
$$x[t] - x[t+T] = 0$$

The quadratic sum of errors over a window of W

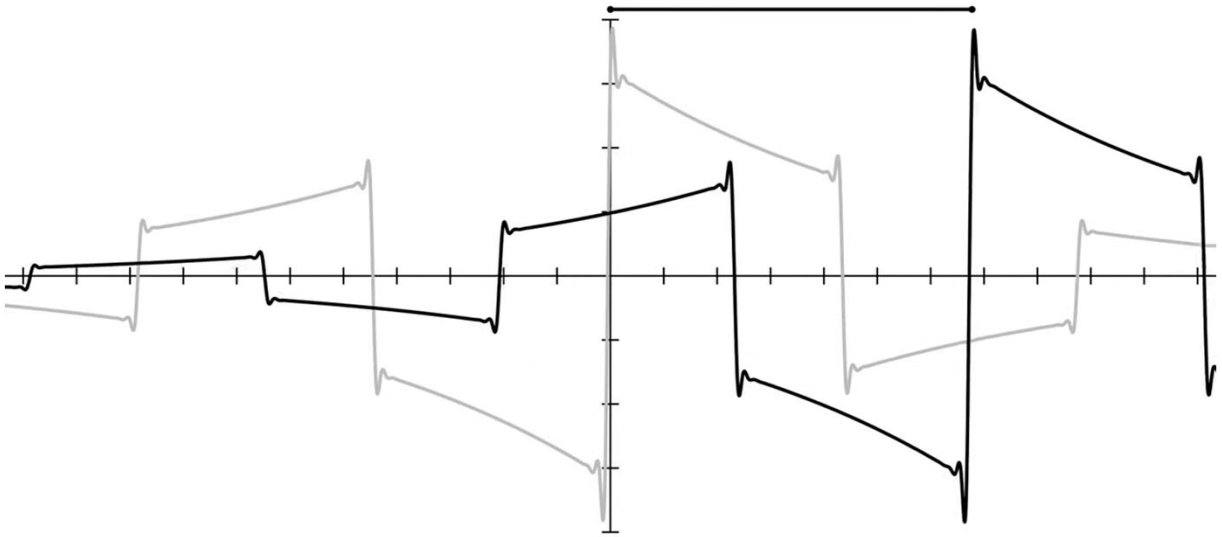
$$\sum_{i=t}^{t+W} (x[i] - x[i+T])^2 = 0$$

YIN, a Traditional Method

- However, for **aperiodic signal**
  the equation

$$\sum_{i=t}^{t+W}(x[i]-x[i+\tau])^2 = 0$$

  does not hold true for any τ value



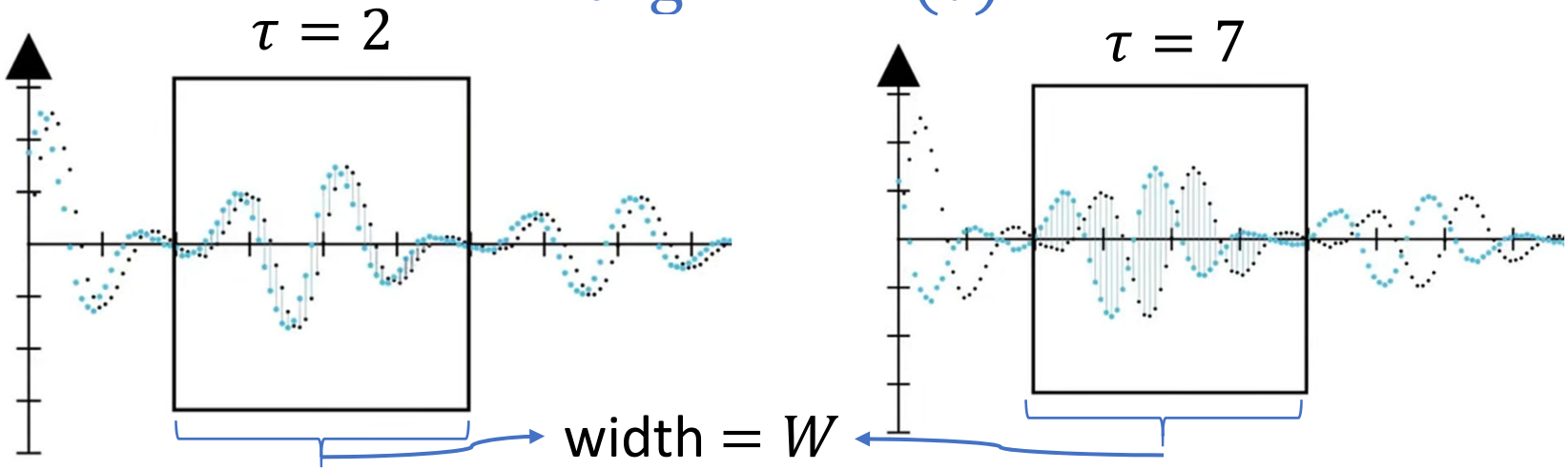
YIN, a Traditional Method

- **The Main idea**: For aperiodic signals, shift the signal for τ samples, search for the τ that minimally changes the original signal.
- Define a difference function (DF)

$$DF(\tau) = \sum_{i=t}^{t+W} (x[i] - x[i+\tau])^2$$

- Then search for the τ that minimizes the DF

$$\operatorname{argmin} DF(\tau)$$



YIN, a Traditional Method

- There should be a **lower and higher bound** for the searching value of τ, defining the highest and lowest frequency
- The value of difference function is **normalized by cumulative mean** to reduce the impact of formants with higher frequency

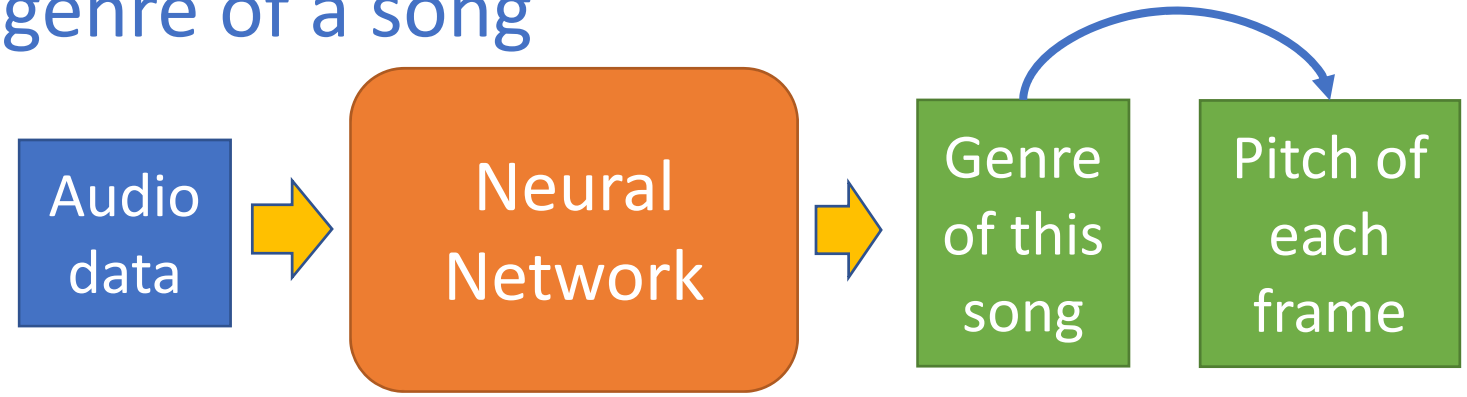
$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \Big/ \left[(1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{otherwise.} \end{cases}$$

- Refer to [1] for more tricks to stabilize recognition

[1] de Cheveigne, Alain, and Hideki Kawahara. "YIN, a Fundamental Frequency Estimator for Speech and Musica)." *J. Acoust. Soc. Am.*, vol. 111, no. 4, 2002, p. 14.

Neural Network Method

Recall: the neural network can classify the genre of a song

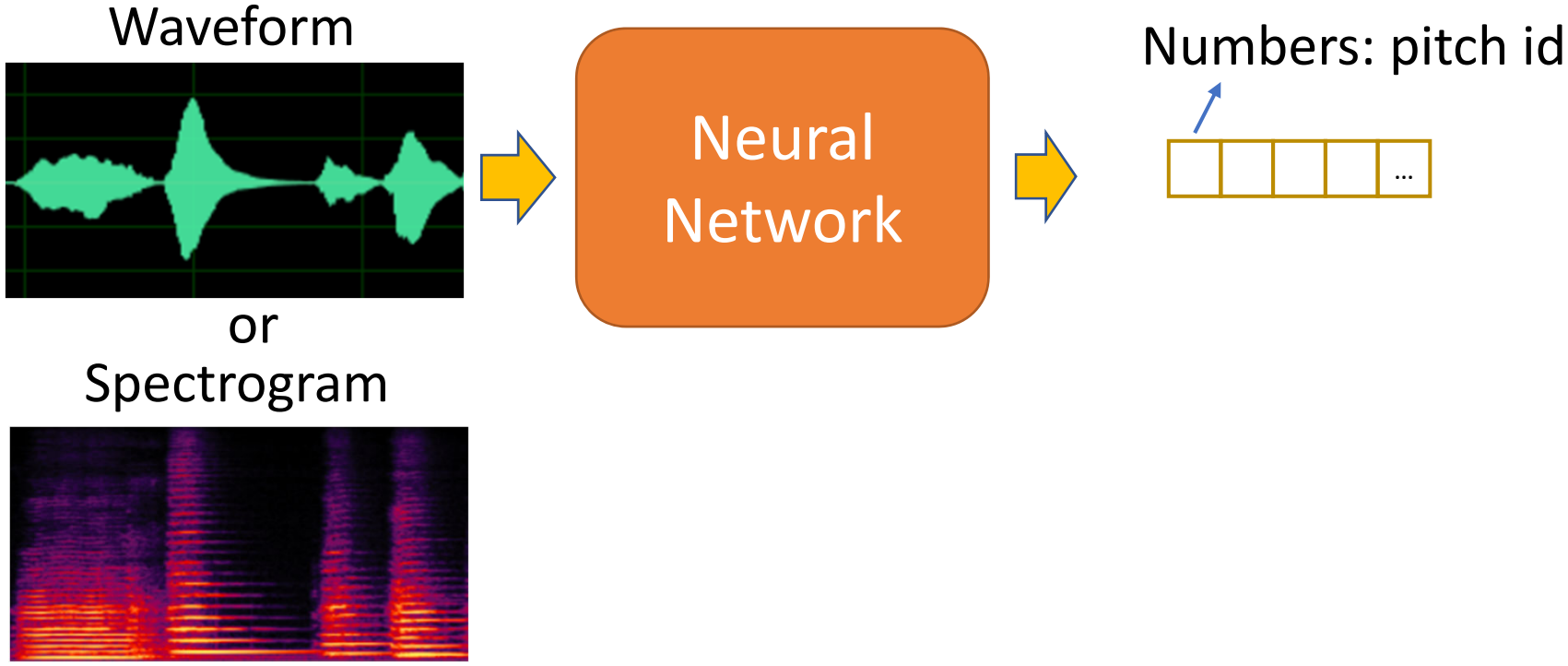


Major difference with genre classification:

	Genre classification	Pitch detection
Level of classification	Song-level	Frame-level
Target	Genre	Pitch

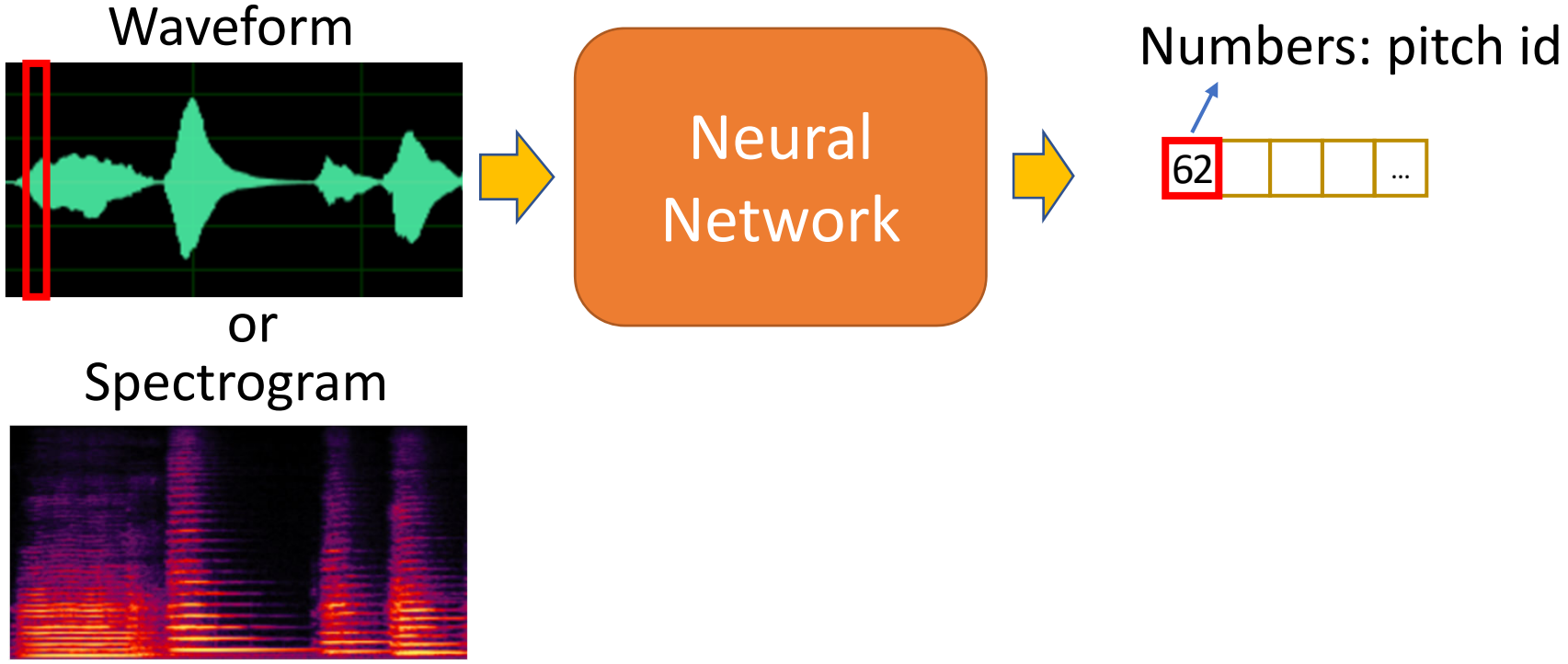
Neural Network Method: Input and Output

- Main idea: Frame-level **classification**
- Input: waveform or spectrogram of an audio **clip**
- Output: predicted pitch class of each frame



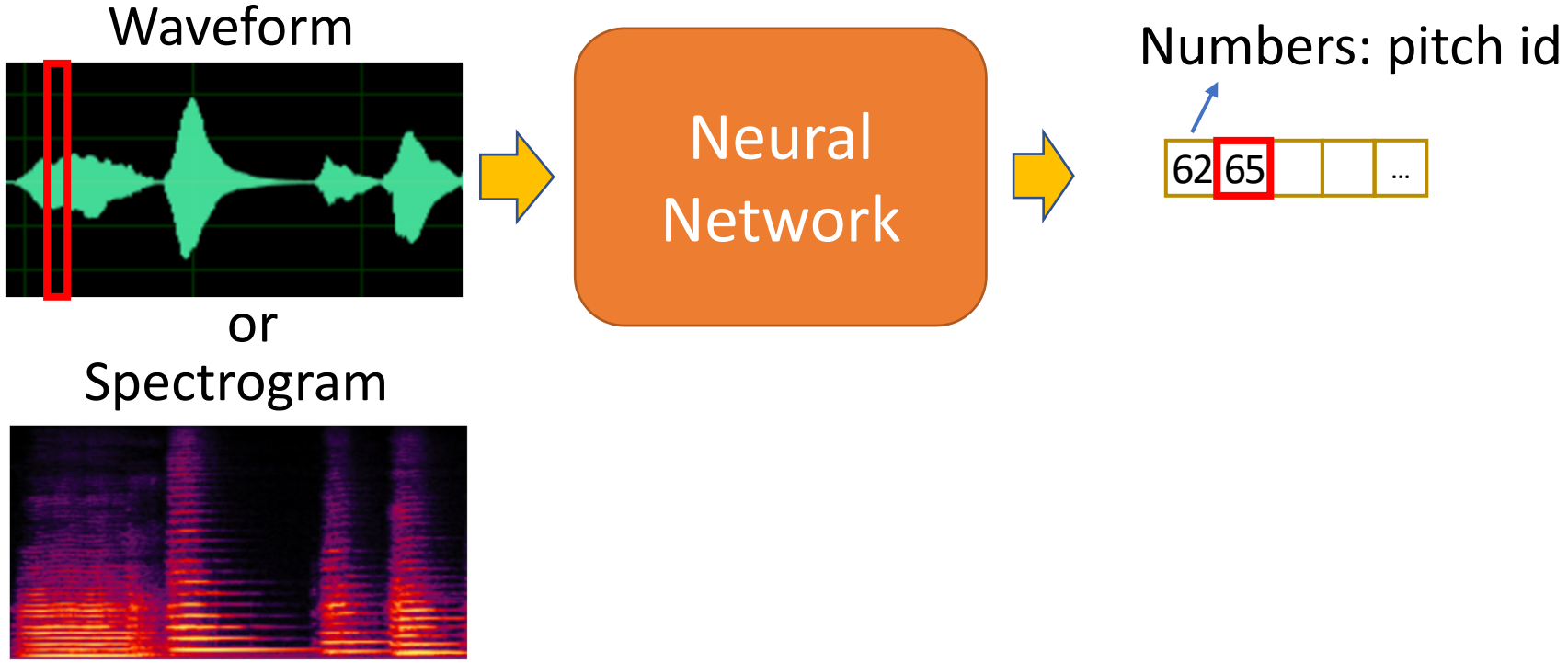
Neural Network Method: Input and Output

- Main idea: Frame-level **classification**
- Input: waveform or spectrogram of an audio clip
- Output: predicted pitch class of each frame



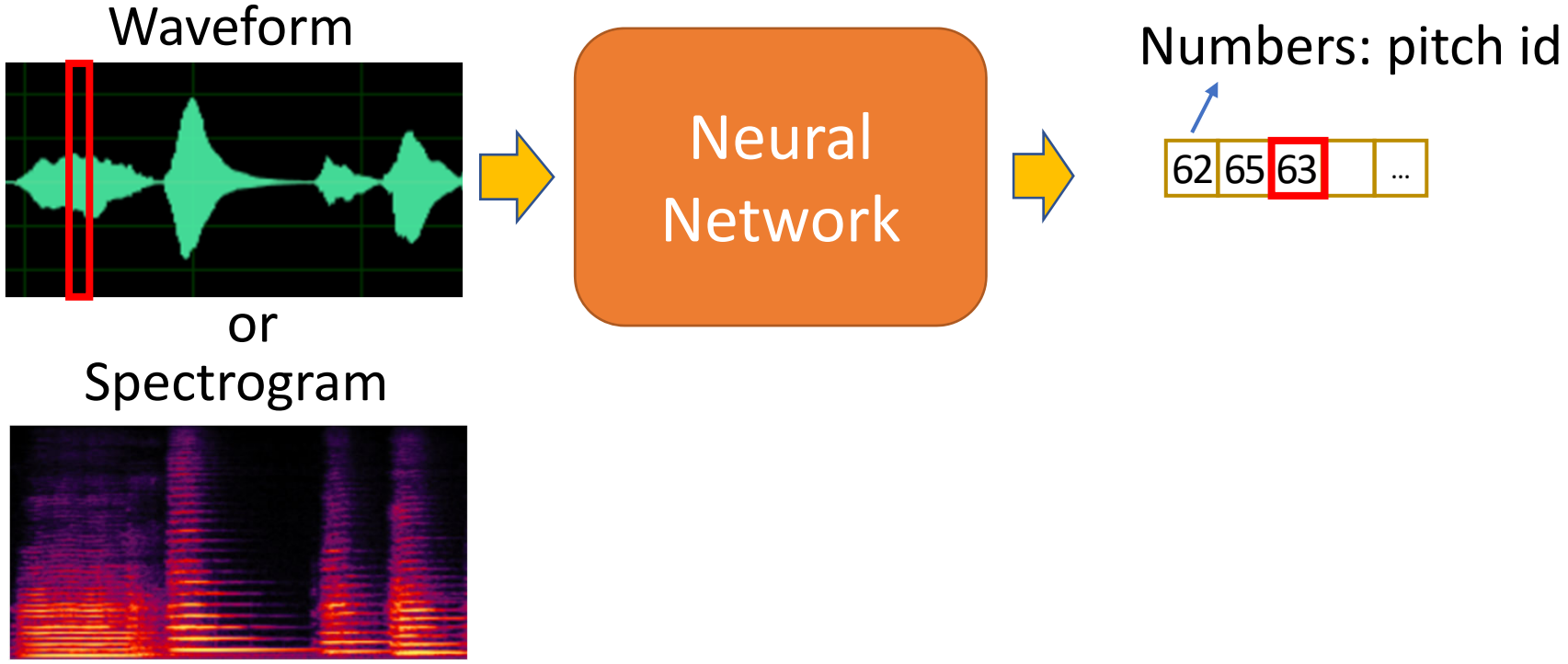
Neural Network Method: Input and Output

- Main idea: Frame-level **classification**
- Input: waveform or spectrogram of an audio clip
- Output: pitch classification of each frame



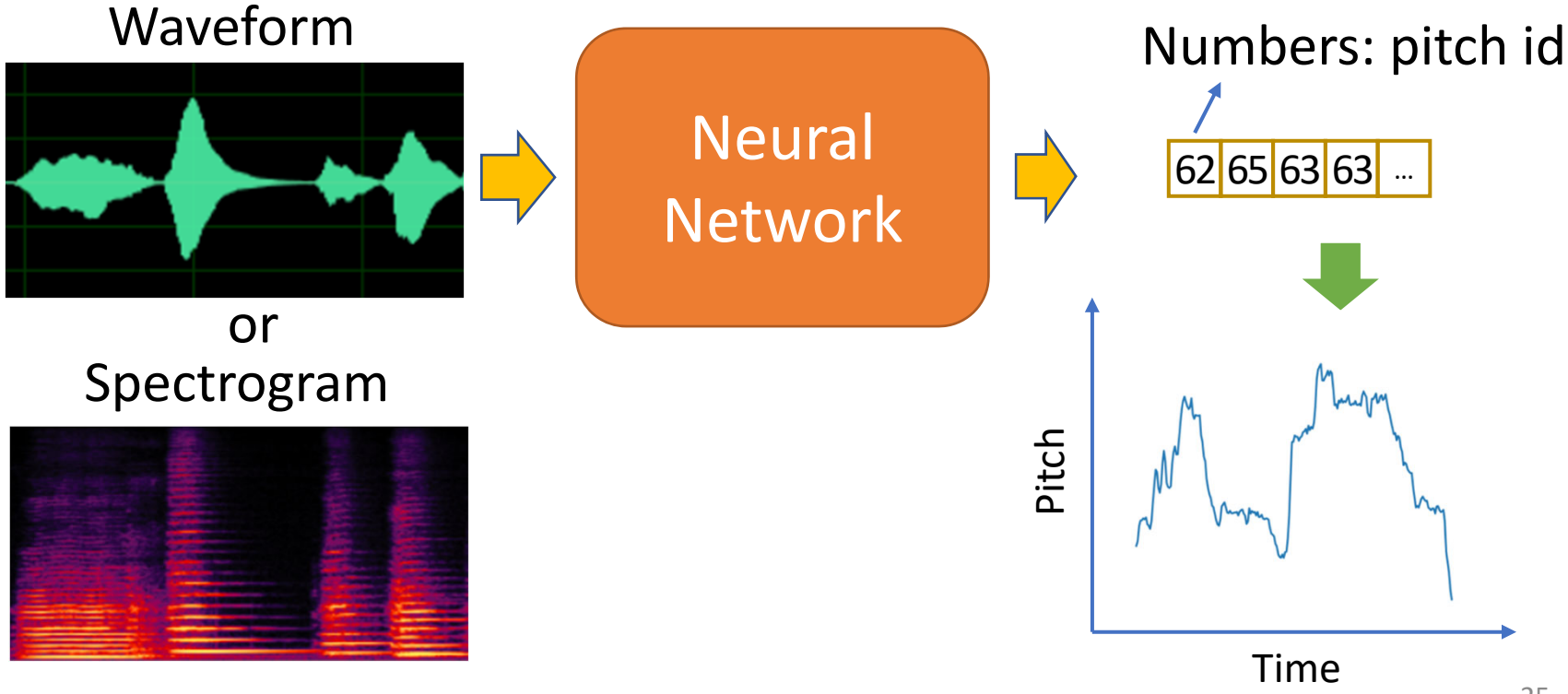
Neural Network Method: Input and Output

- Main idea: Frame-level **classification**
- Input: waveform or spectrogram of an audio clip
- Output: pitch classification of each frame



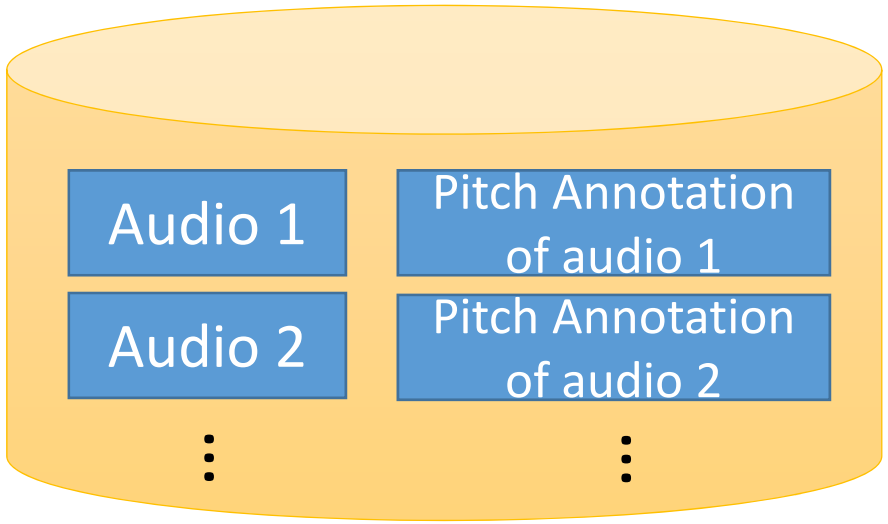
Neural Network Method: Input and Output

- Main idea: Frame-level **classification**
- Input: waveform or spectrogram of an audio clip
- Output: pitch classification of each frame



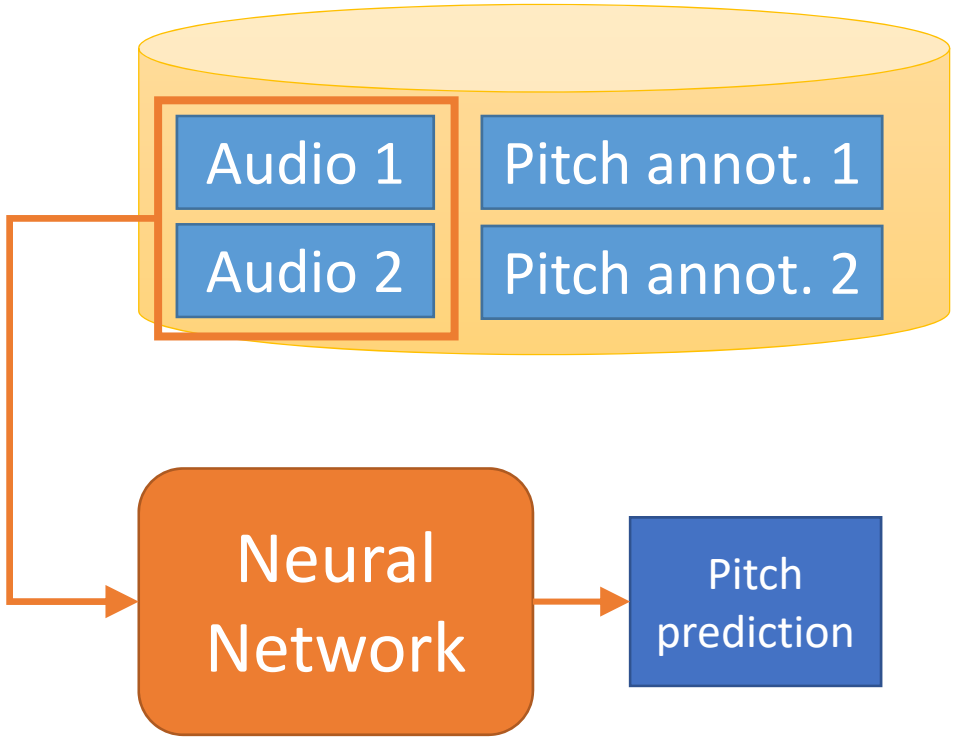
Neural Network Method: Network Training

- Training scheme:
  - In a **Supervised** manner
  - Which need **labeled dataset** (audio and frame-level pitch annotation), as shown below:



Neural Network Method: Network Training

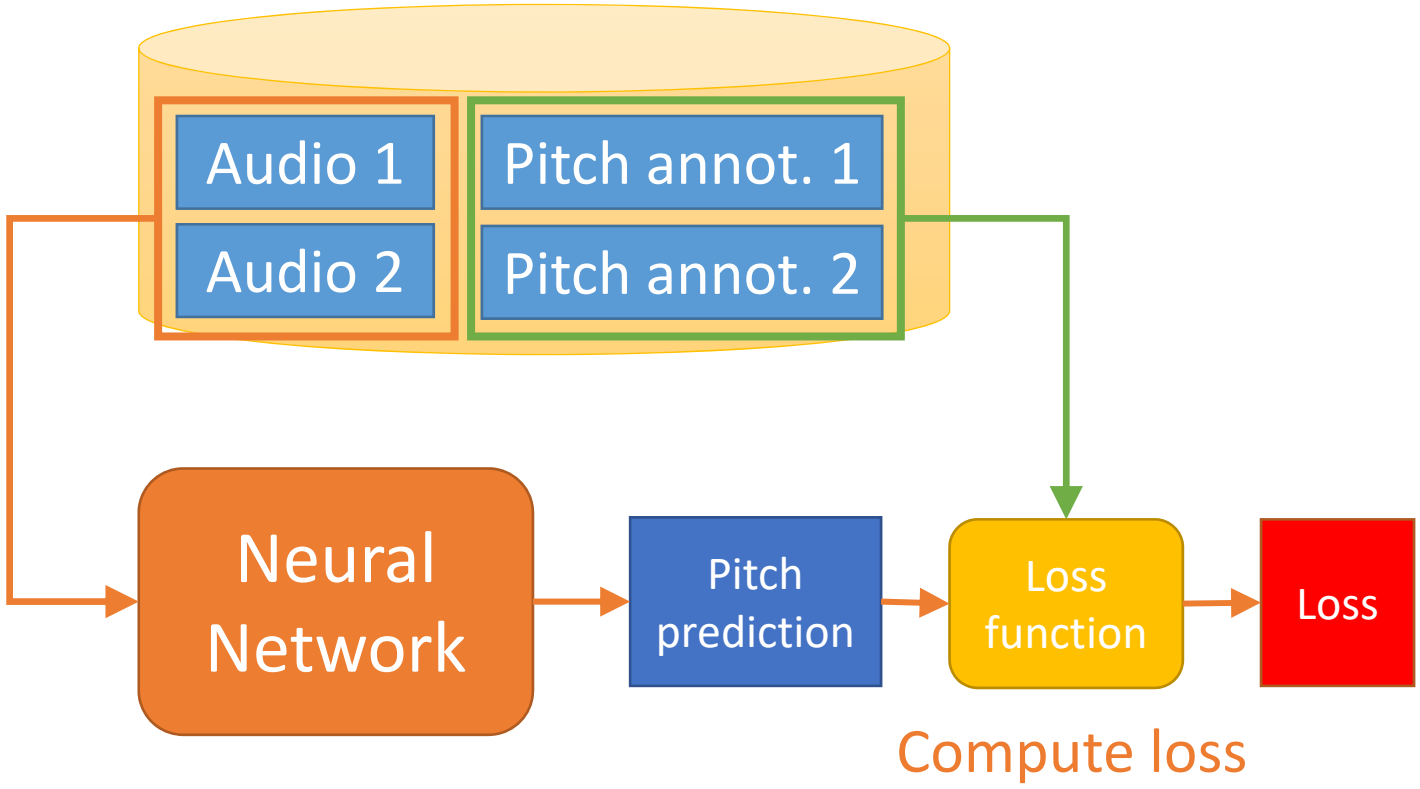
- Training procedure:



Forward pass

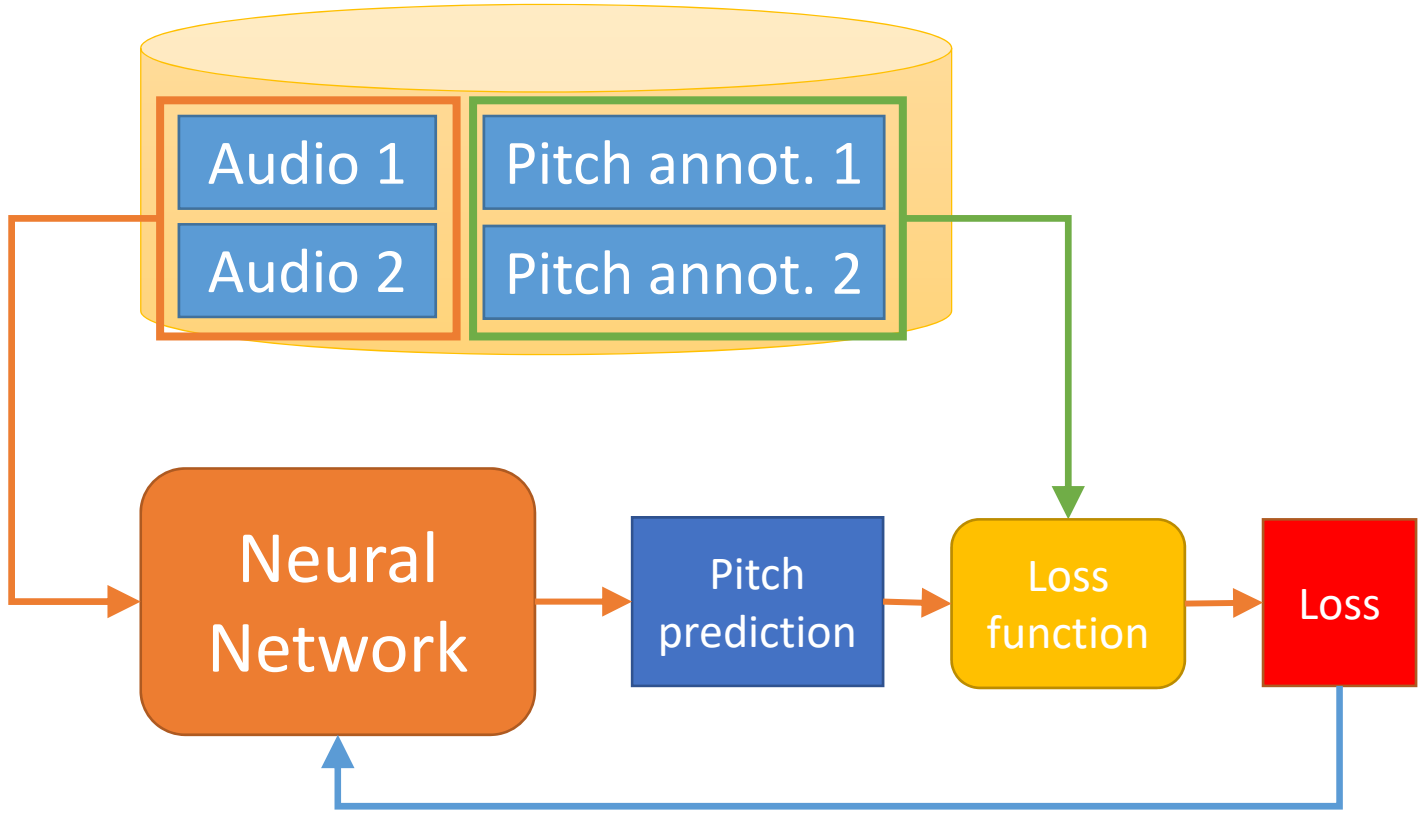
Neural Network Method: Network Training

- Training procedure:



Neural Network Method: Network Training

- Training procedure:



Loss Computing: Cross Entropy (CE)

Probability that the network believes the pitch is …

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Raw output

Pitch	1	2	3	4	5
C3	9	0.2	2	3	-2
D3	0.5	0.8	1	0.5	1.5
E3	0.2	0.1	4	4.5	3.5

Frame

Softmax

Pitch	1	2	3	4	5
C3	1	0.27	…		
D3	0	0.49	…		
E3	0	0.24	…		

Frame

Each column's summation is 1

Similarity score, higher score means more similar, range (-inf, inf)

$$CE(pred, label) = -\sum_{i=1}^{num\ of\ classes} label_i \cdot \log(pred_i)$$

$$loss = \sum_{j=1}^{num\ of\ frames} CE(pred_{Frame\ j}, label_{Frame_j})$$

Sum of loss from all columns

Loss Computing: Binary Cross Entropy

Probability that the network believes the pitch is …

$$S(x) = \frac{1}{1 + e^{-x}}$$

Raw output

Pitch	1	2	3	4	5
C3	9	0.2	2	3	-2
D3	0.5	0.8	1	0.5	1.5
E3	0.2	0.1	4	4.5	3.5

Frame

Sigmoid

Pitch	1	2	3	4	5
C3	0.99	0.55			
D3	0.62	0.69			
E3	0.54	0.53			

Frame

Probability of each pitch is independent

$$BCE(pred, label) = -[label \cdot \log(pred) + (1 - label) \cdot \log(1 - pred)]$$

Sum of loss from all cells

$$loss = \sum_{j=1}^{\#frames} \sum_{i=1}^{\#pitches} BCE(pred_{Frame_j, Pitch_i}, label_{Frame_j, Pitch_i})$$

Neural Network Method: Network Design

Large degrees of freedom

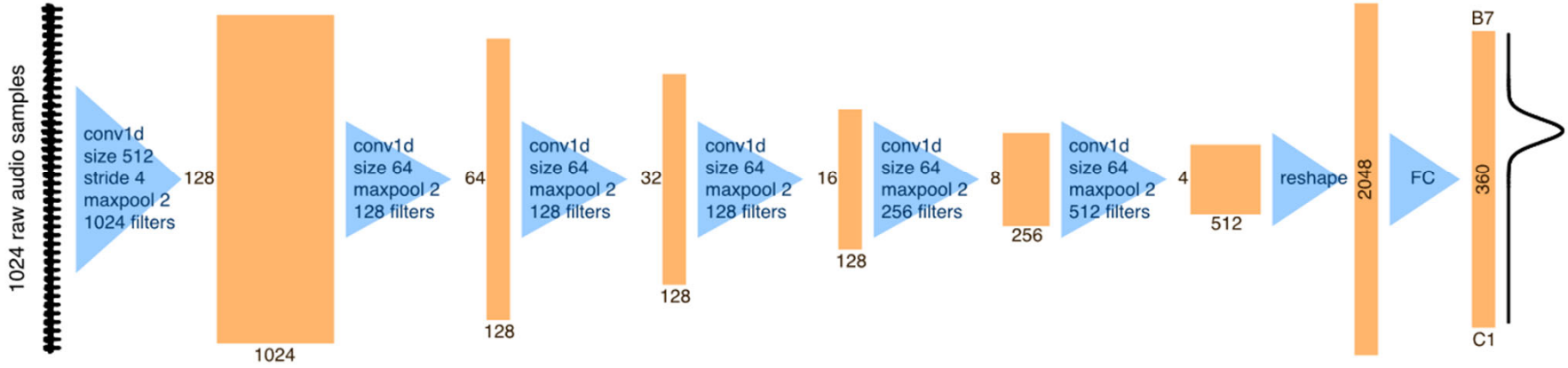
- Network structure:
  - CNN, RNN, Transformer, …
  - Number of neurons, CNN kernals, layers, …
- Loss function:
  - Sigmoid + Binary Cross Entropy loss
  - Softmax + Cross Entropy loss
  - Mean absolute error, as a regression problem
- How to get the network that works better?
  - Trial and error
  - Imitate the design of SOTAs

Neural Network Method

State-of-the-art Methods

Highest accuracy for monophonic music:

- Crepe[1]



[1] Kim, Jong Wook, et al. "Crepe: A convolutional representation for pitch estimation." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Neural Network Method

State-of-the-art Methods

Highest accuracy for monophonic music:

- Crepe[1]

Other works:

- Detect singing activity and pitch together[2]
- Classify Tone & Octave separately[3]
- Self-supervised learning method[4]

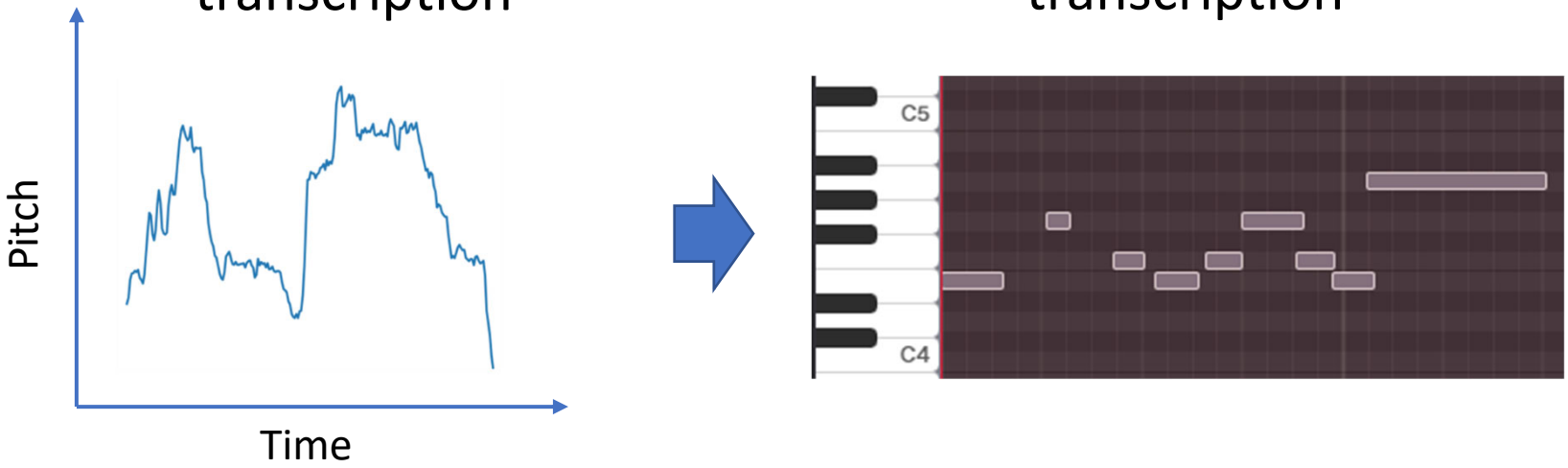
[1] Kim, Jong Wook, et al. "Crepe: A convolutional representation for pitch estimation." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
[2] Kum, Sangeun, and Juhan Nam. "Joint detection and classification of singing voice melody using convolutional recurrent neural networks." *Applied Sciences* 9.7 (2019): 1324.
[3] Chen, Ke, et al. "Tonet: Tone-octave network for singing melody extraction from polyphonic music." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.
[4] Gfeller, Beat, et al. "SPICE: Self-supervised pitch estimation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 1118-1128.

Note-level Methods

Method that **output note-level events**

Frame-level transcription

Note-level transcription



vs. frame level method, note-level method capture information of:

- Onset, start time of a note
- Offset, end time of a note

Evaluation Metrics of Note-level Method

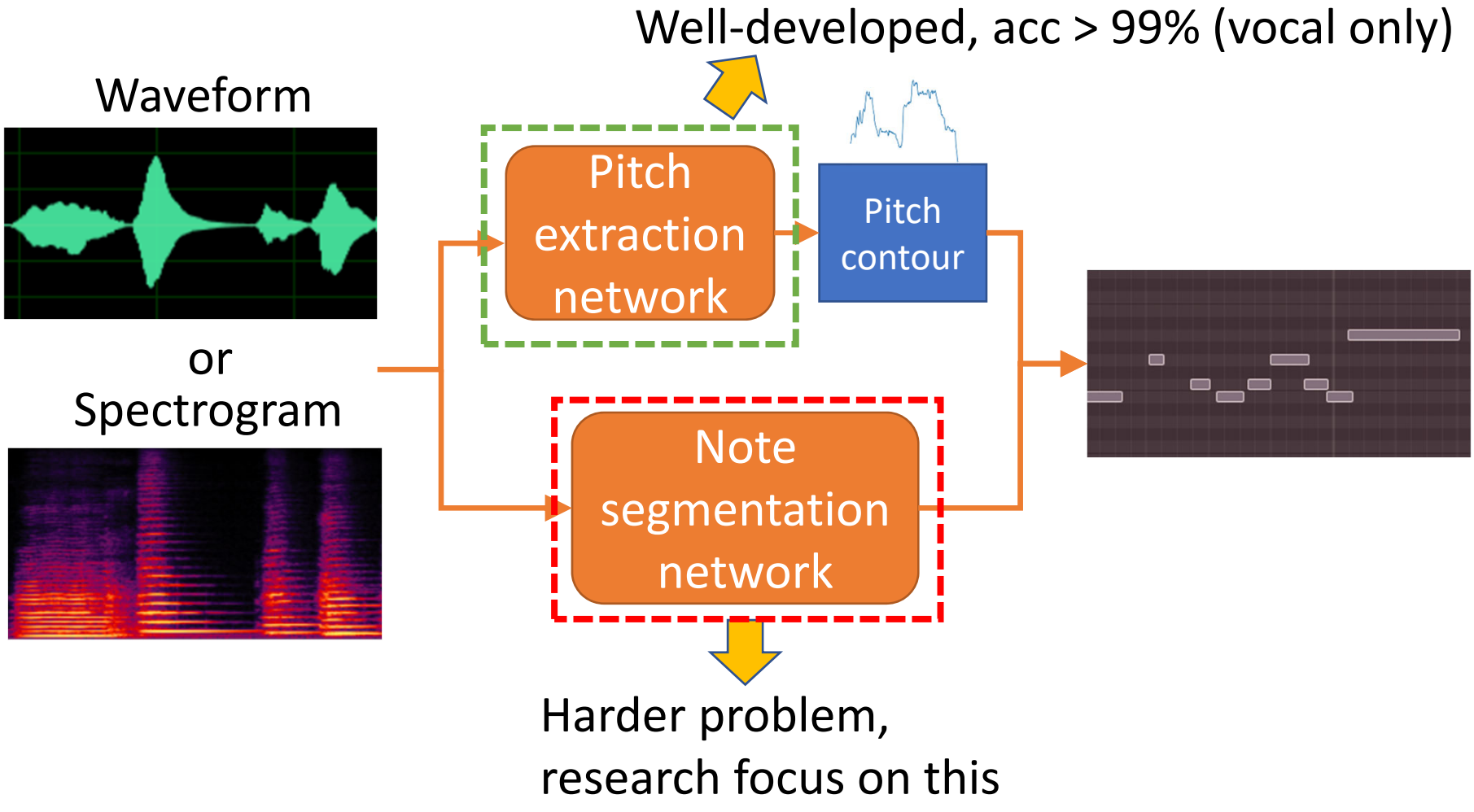
F1 score as the metric (refer to [1] for details)

- Consider different note element
  - onset/offset only
  - onset + offset
  - onset + pitch
  - onset + pitch + offset
- Have tolerance for different element:
  - Pitch: 50 cents
  - Onset: 50 ms
  - Offset: 0.2 × the note's duration

F1:
harmonic mean of
precision and *recall*

[1] Hsu, Jui-Yang, and Li Su. "VOCANO: A Note Transcription Framework for Singing Voice in Polyphonic Music." *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, pp. 293–300,

Note-level Neural Network Method: Overall structure



Note segmentation network: Naïve design

1: this frame is **onset**
-1: this frame is **offset**
0: neither onset/offset

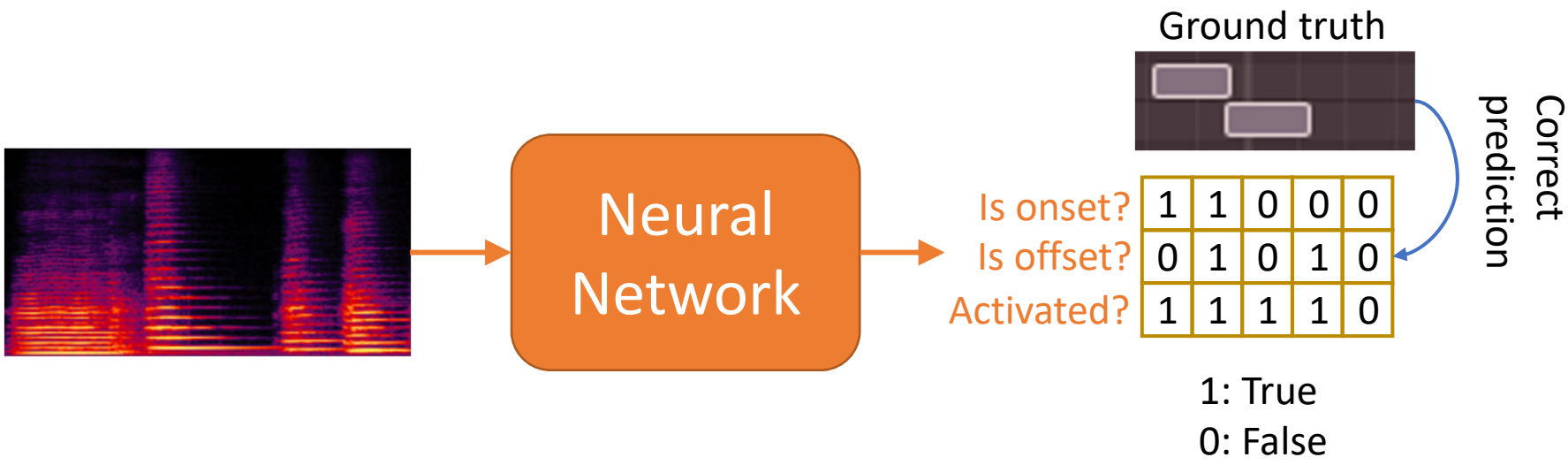
2 Potential problem:

- Ambiguity of onset and offset
  - In the annotation, onset and offset may appear on the same frame
- Data imbalance:
  - Most frames are neither onset nor offset

Note segmentation network: Better structure

Solution to previous problems:

- Allow a frame to be onset and offset at the same time (through **multi-label classification**)
- Predict whether a frame have voice (activation)



Fu, Zih-Sing, and Li Su. "Hierarchical classification networks for singing voice segmentation and transcription." *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*. 2019.

Further improvement

- Highlight the energy change, by using spectral difference as additional input[1]
- Semi-supervised learning to augment data, better post-processing method[2]
- Transfer learning[3] from frame-level to note-level

[1] Fu, Zih-Sing, and Li Su. "Hierarchical classification networks for singing voice segmentation and transcription." *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*. 2019.
[2] Hsu, Jui-Yang, and Li Su. "VOCANO: A note transcription framework for singing voice in polyphonic music." *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR 2021)*. 2021
[3] Kum, Sangeun, et al. "Pseudo-Label Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

Take a break

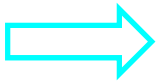
Topics Today

Part A: Automatic music transcription (AMT)

- The AMT task
- Application of AMT systems

Part B: Singing voice transcription

- Signal processing method: Yin
- Neural network methods

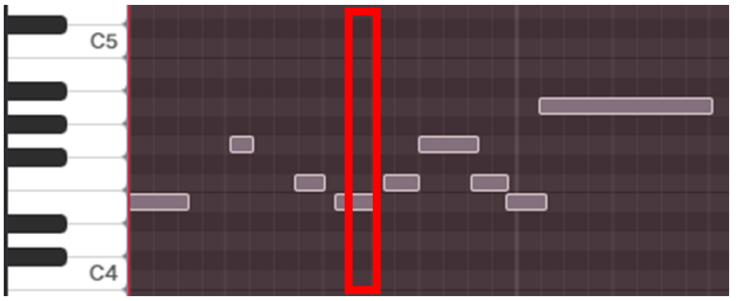


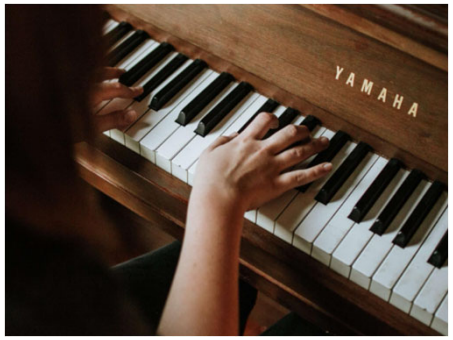
Part C: Piano music transcription

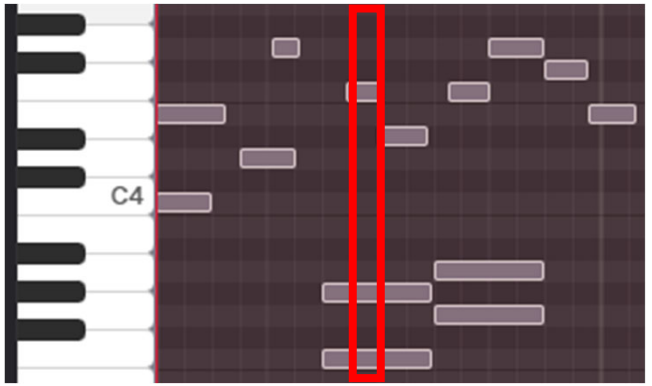
- Multipitch challenge
- Frame- and note-level methods

Piano Music is Polyphonic

Polyphonic music is difficult to transcribe





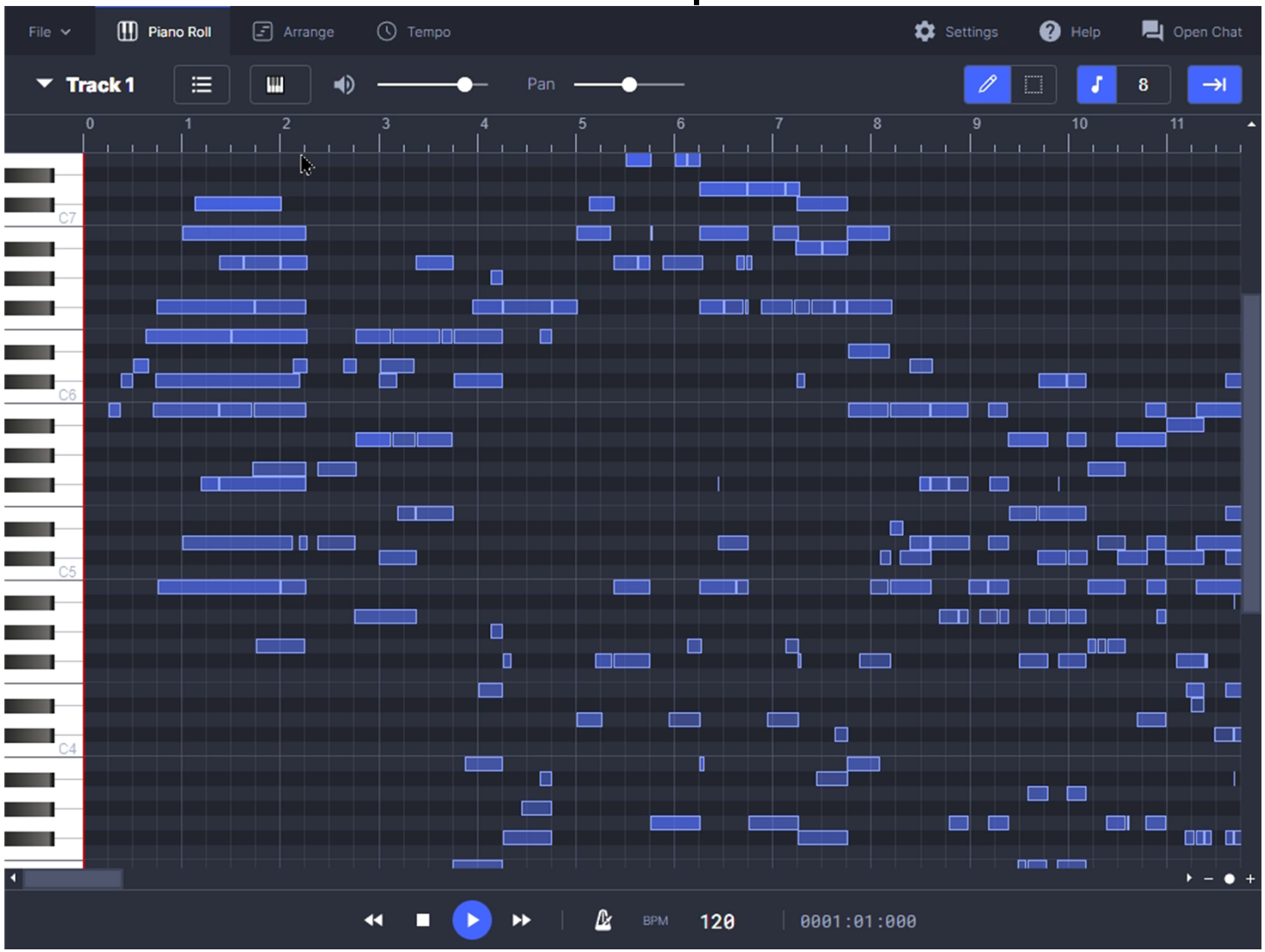


Multiple notes may appear together

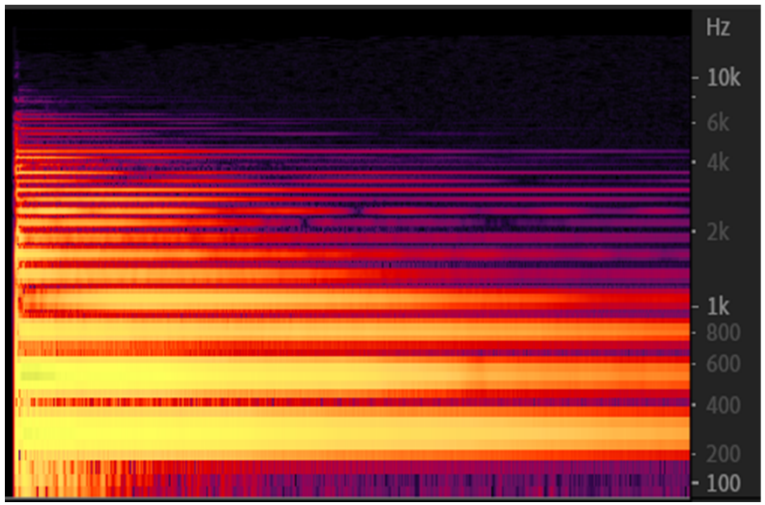
Piano Transcription Demo

Recording

Transcription

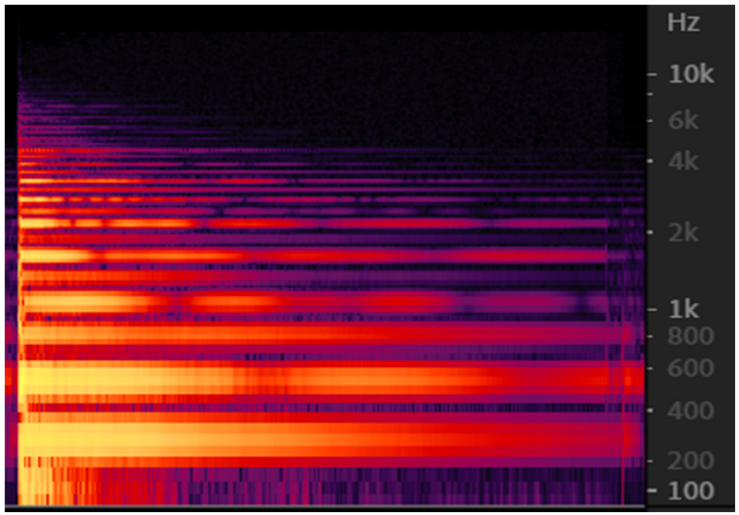


Challenge of polyphonic music: Harmonics overlap in frequency

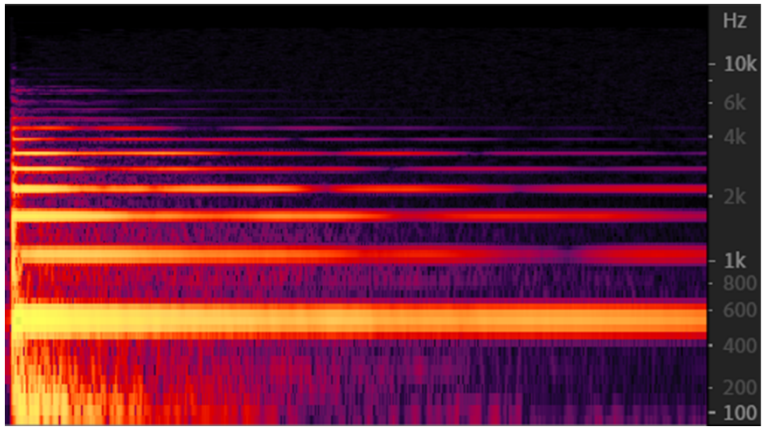


C3

vs.



Mixture



C4

Such a problem occurs when two notes, whose f0s have small integer ratio, are played together

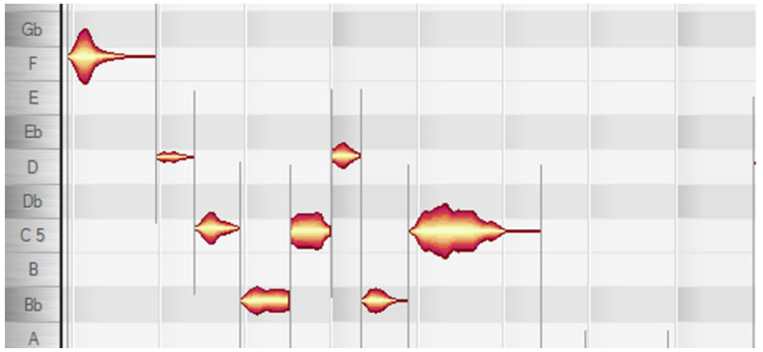
Subtasks and Metrics of Piano Transcription

- Necessary elements of piano music:
  - onset, offset
  - pitch
  - More: velocity, pedal, …
- F1 score as evaluation metric, similar to note-level singing transcription
  - onset/offset only
  - onset + pitch
  - onset + pitch + offset
  - onset + pitch + offset + velocity

Transcribe More Note Elements: Note Intensity

The volume of notes are also an important component for AMT

- It's been done by previously mentioned "velocity detection" module
- "Velocity" refers to the speed of pressing piano keys, which relates to the note intensity/volume
- Make the transcription results more "natural"



Transcribe More Note Elements: Sustain Pedal

The pedal event of is another important component for piano AMT

- Sustain pedal is broadly used in various piano pieces
- Transcribed pedal information help player play the music more easily when using the note-level transcription as the score

Transcription representation: MIDI file

- A standardized format to save music sequences
- Contains one (or more) **list of events** that describe
  - instrument type
  - onset, offset, volume, pitch of each note

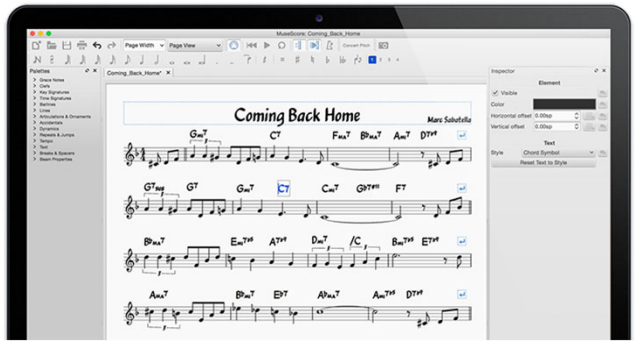

```
(StandardMidiFile
  (MidiHeader 'format "0" 'numberOfTracks "1" 'pulsesPerQuarterNote "480" 'mode "deltaTime")
  (MidiTrack
     (ControlChange 'deltaTime "0" 'info "0:0:0 Bank select MSB" 'channel "1" 'control "0" 'value "0")
     (ControlChange 'deltaTime "0" 'info "0:0:0 Bank select LSB" 'channel "1" 'control "32" 'value "0")
     (ProgramChange 'deltaTime "0" 'info "0:0:0 **GM Piano: Bright Acoustic Piano" 'channel "1" 'number "1")
     (Meta 'deltaTime "0" 'info "0:0:0 Tempo: 80 BPM." 'type "81" "0B 71 B0")
     (NoteOn 'deltaTime "480" 'info "0:1:0 C3" 'channel "1" 'note "60" 'velocity "88" 'duration "202")
     (NoteOn 'deltaTime "119" 'info "0:1:119 C#3" 'channel "1" 'note "61" 'velocity "114" 'duration "292")
     (NoteOn 'deltaTime "131" 'info "0:1:250 D3" 'channel "1" 'note "62" 'velocity "87" 'duration "201")
     (NoteOn 'deltaTime "119" 'info "0:1:369 D#3" 'channel "1" 'note "63" 'velocity "114" 'duration "290")
     (NoteOn 'deltaTime "131" 'info "0:2:20 E3" 'channel "1" 'note "64" 'velocity "114" 'duration "290")
     (NoteOn 'deltaTime "131" 'info "0:2:151 F3" 'channel "1" 'note "65" 'velocity "87" 'duration "199")
```

Transcription representation: MIDI file

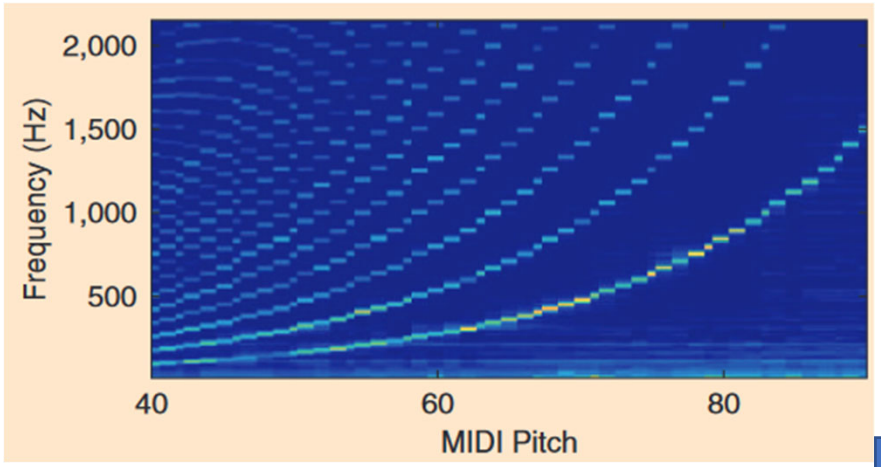
- Recorded by MIDI controller
- Can be "played" by MIDI synthesizer
- Can be edited by digital audio workstation (e.g., FL studio, Cubase)

opened by DAW

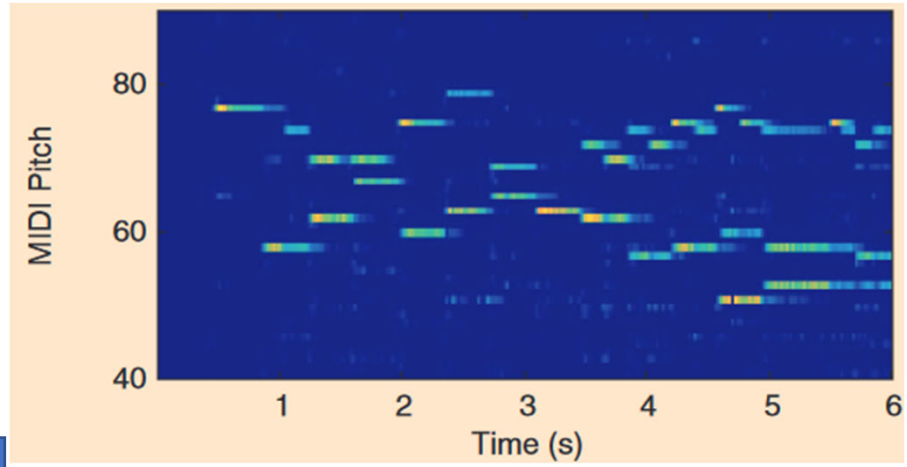
opened by MuseScore



Traditional method:
Non-negative matrix factorization (NMF)

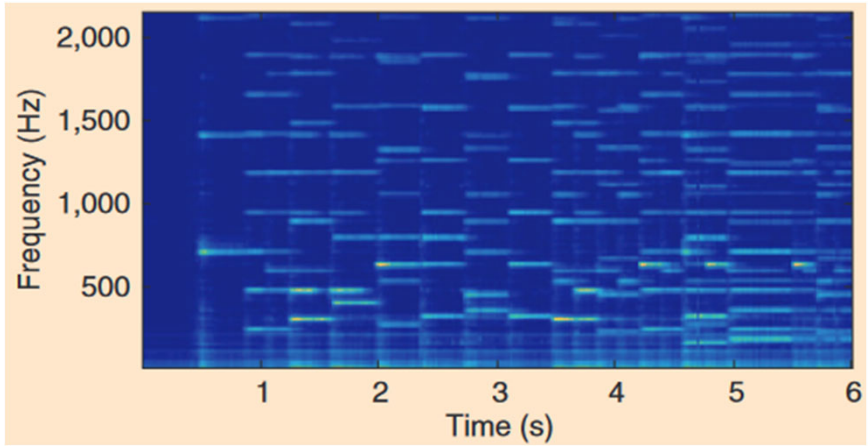


Instrument



Score

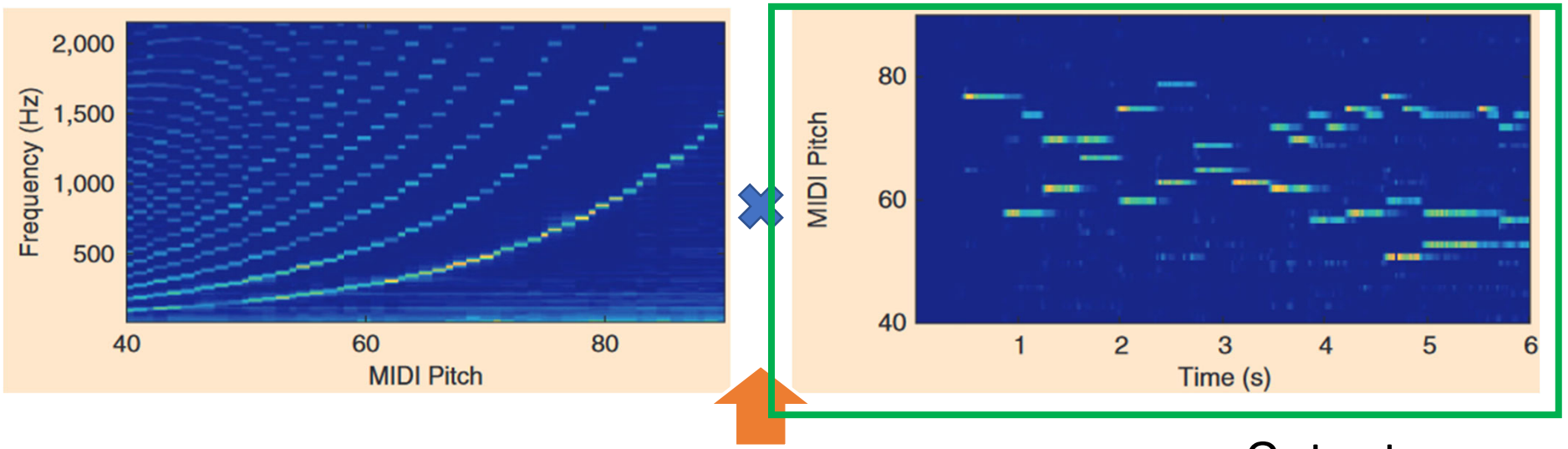
Performance



Music
(spectrogram)

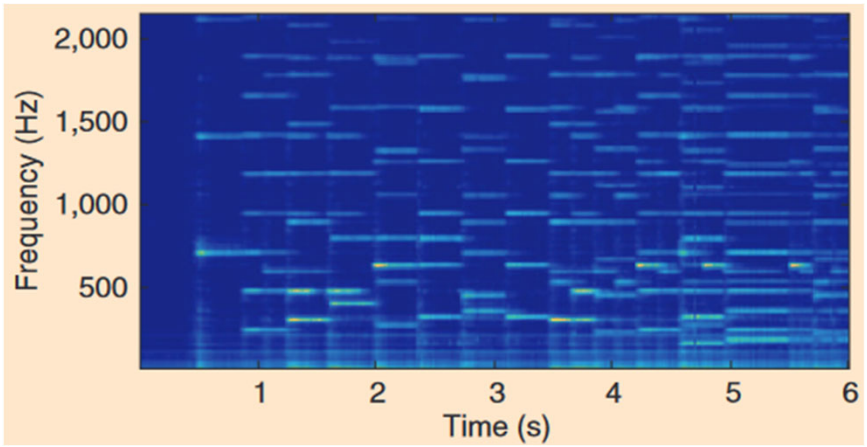
Figure from:
Benetos, Emmanouil, et al. "Automatic music transcription: An overview." *IEEE Signal Processing Magazine* 36.1 (2018): 20-30.

Traditional method:
Non-negative matrix factorization (NMF)



Transcribe:

Factorize the spectrogram under non-negative constraints



Output

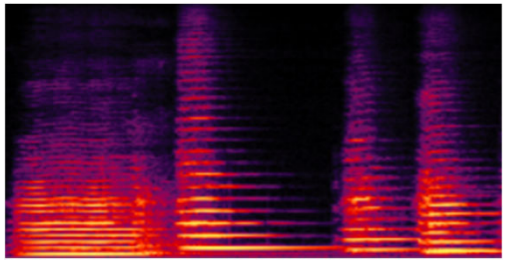
Spectrogram as input

More details might be covered later in a guest lecture.

Methodology of Frame-level NN method

Multi-label Frame classification

Spectrogram

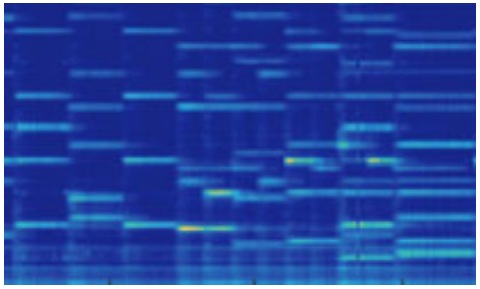


Neural Network

Singing data:
One label at a time

62	65	63	63	…

Spectrogram



Neural Network

Piano data:
Multiple labels per frame

Pitch 63 exists?	1	1	0	0	0
Pitch 64 exists?	0	1	0	1	0 …
Pitch 65 exists?	1	1	1	1	0

(88 pitches in total)

1: True
0: False

Methodology of Frame-level NN method Training and datasets

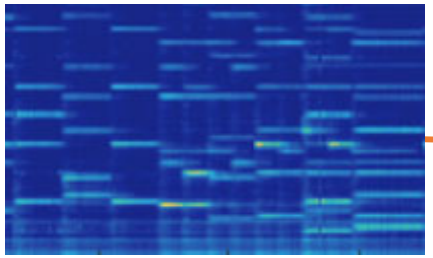
- Training scheme:
  - In a supervised manner
- Loss function:
  - Sigmoid + Binary Cross Entropy loss
    ~~Softmax + Cross Entropy~~
- Datasets:
  - Easier to obtain compared to singing
  - Commonly used: MAPS[1] (18h), MAESTRO[2] (172h)

[1] Emiya, Valentin, et al. "MAPS-A piano database for multipitch estimation and automatic transcription of music." (2010): 11.
[2] Hawthorne, Curtis, et al. “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset.” 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019

Pitch classification for piano AMT

System structure

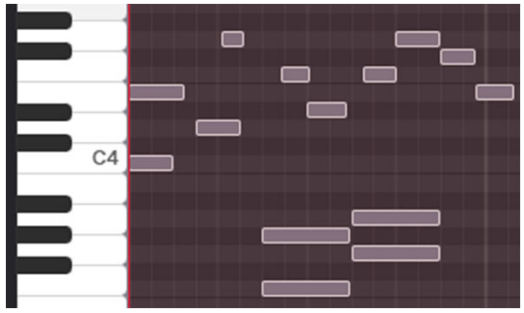
Last layer:
a linear layer



Network can be:
- CNN,
- RNN,
- Transfor-mer

Pitch 63 exists?	1	1	0	0	0
Pitch 64 exists?	0	1	0	1	0
Pitch 65 exists?	1	1	1	1	0

Visualize



Piano roll

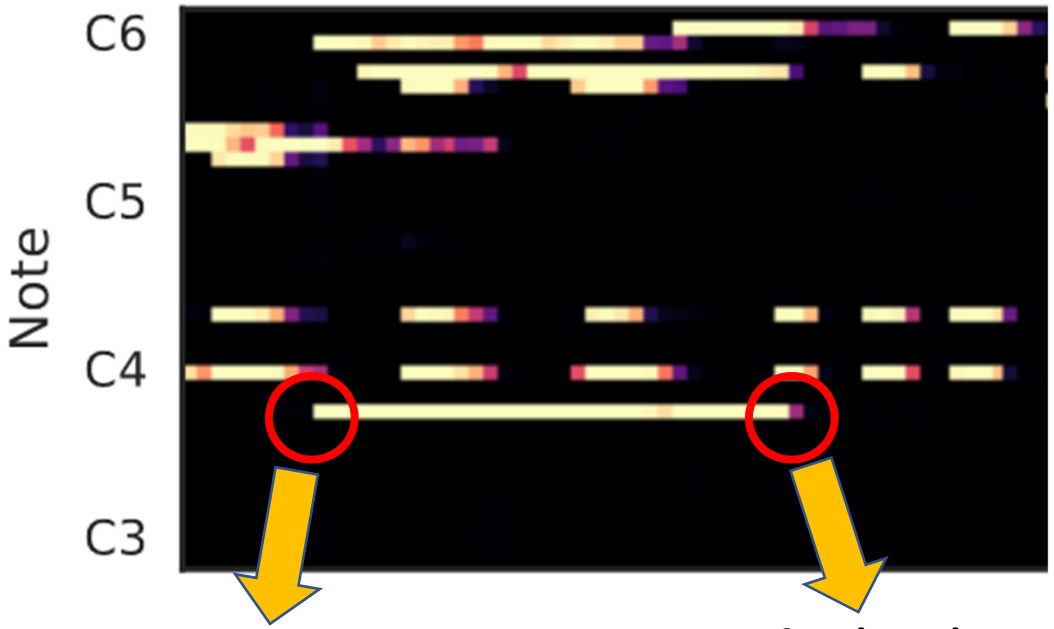
Spectrogram obtained through:
- Constant-Q transform
- Or short-time FT

Kelz, Rainer, et al. "On the Potential of Simple Framewise Approaches to Piano Transcription." *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*

Post-processing to get note-level result

- Onset/offset can be inferred from frame-level result, because the transitions between notes are easier to distinguish, compared to AMT for singing

E.g.



Jump from silent to activation: onset detected

Activation gradually disappear: offset detected

Recent Neural Network Method: Convolutional Neural Network

Network structure:

ConvNet
Input 5x229
Conv 32x3x3
Conv 32x3x3
BatchNorm
MaxPool 1x2
Dropout 0.25
Conv 64x3x3
MaxPool 1x2
Dropout 0.25
Dense 512
Dropout 0.5
Dense 88

Performance

- 71.60% frame-level F1 score
- 23.14% note-level F1 score

Kelz, Rainer, et al. "On the Potential of Simple Framewise Approaches to Piano Transcription." *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*

Recent Neural Network Method: Onset & Pitch (1/3)

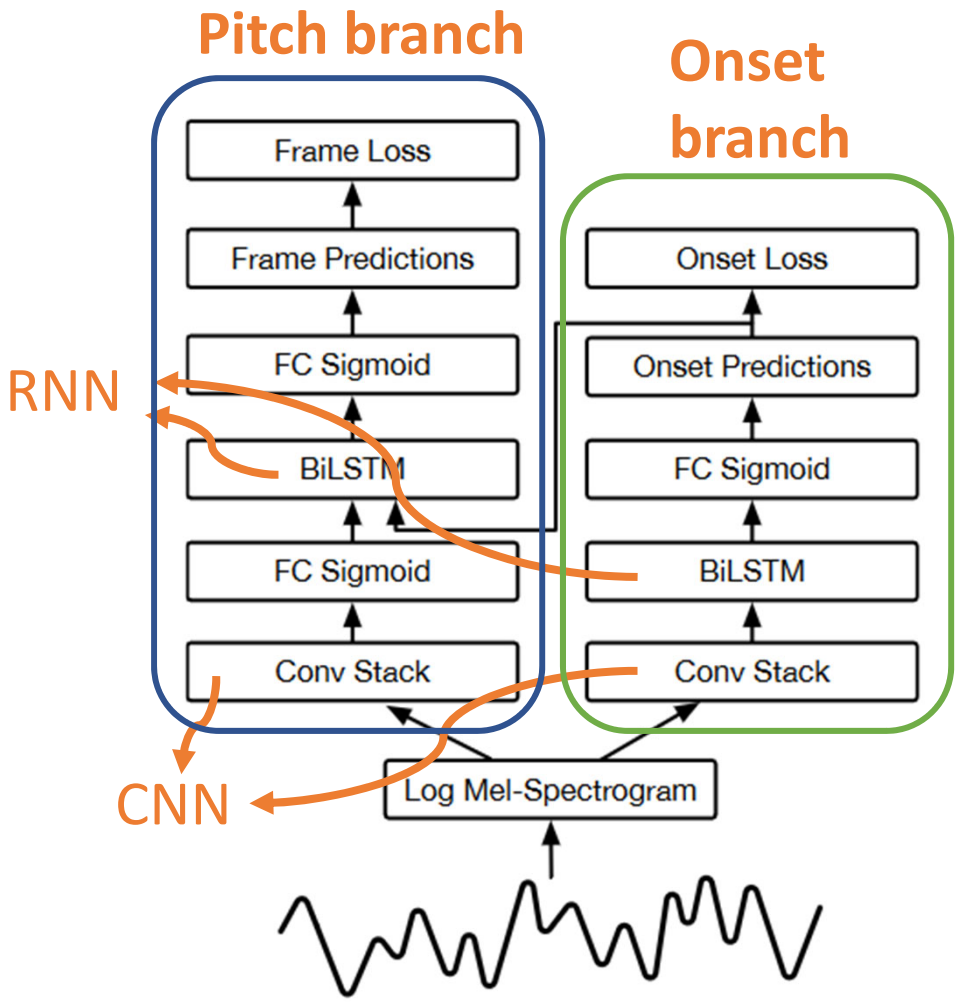
- **Problem**: The onset inferred from the pitch output may be inaccurate
- **Solution**: design network that jointly detect note pitch and onset events
- Why onset?
  - It's more important, we care more about it
  - It's easy to detect than offset

Hawthorne, Curtis, et al. "Onsets and Frames: Dual-Objective Piano Transcription." *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pp. 50–57

Recent Neural Network Method: Onset & Pitch (2/3)

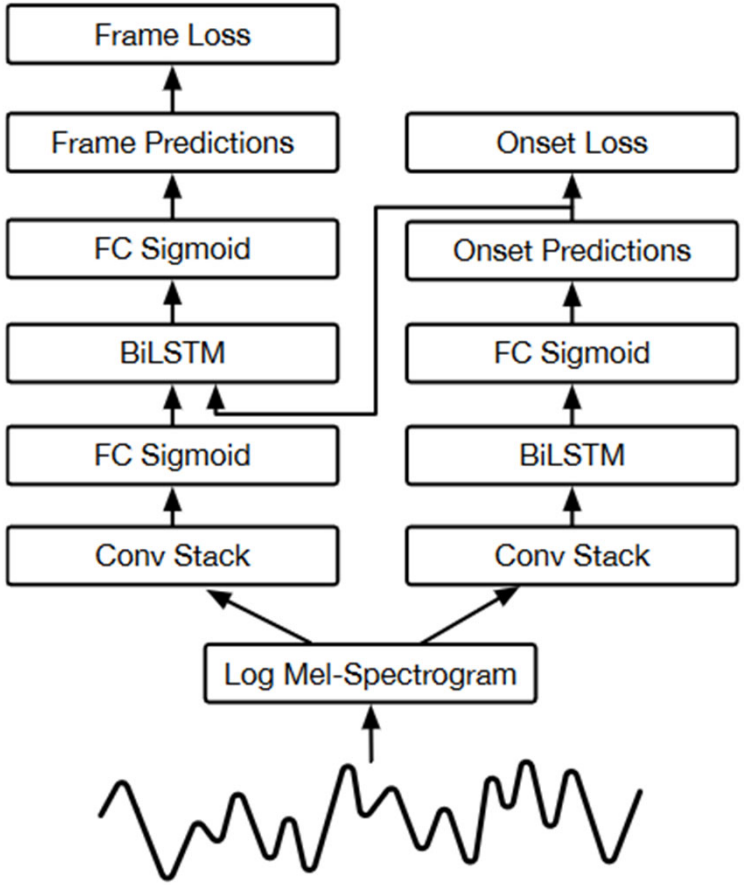
Structure:

- Combination of CNN and RNN
- Two branches, responsible for detecting onsets and pitches, respectively.



Recent Neural Network Method: Onset & Pitch (3/3)

- When doing forward pass:
  1. Onsets are first detected by onset branch
  2. Onset output served as additional input to pitch classifier
- Performance (on MAPS):
  - 78.30% Frame F1
  - **50.22**% Note F1 (a huge leap, vs. previous 23.14%)



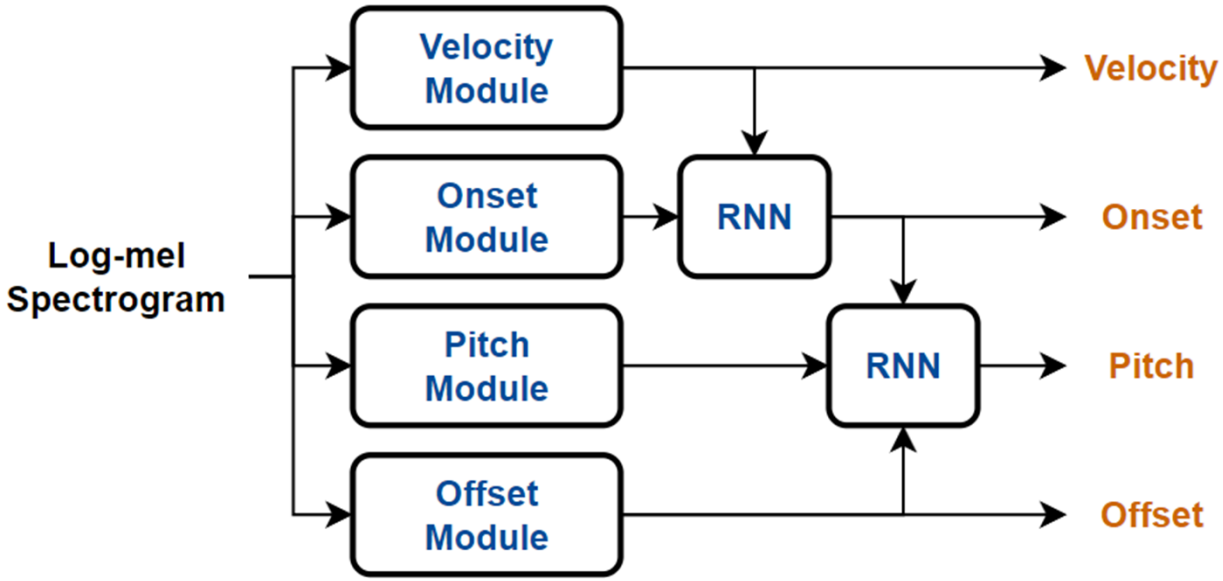
Recent Neural Network Method: High-Resolution Piano Transcription (1/2)

- **Problem**: The time resolution is restricted by the length of frame
  - We can only know a frame has onset or not, don't know it's exact location inside the frame
- **Solution**: Let network predict the distance between
  - the middle of each frame
  - and the nearest onset

Kong, Qiuqiang, et al. "High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times." IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, 2021, pp. 3707–17.

Recent Neural Network Method: High-Resolution Piano Transcription (2/2)

- Structure: Additional modules for
  - Offset detection
  - and velocity detection
- Performance:
  - 89.62% Frame F1
  - 80.92% Note F1



More Elegant Solution: Direct Note-level Transcription

Advantages over frame-level models:

- A single objective function during training, which is consistent with the **note-level** AMT objective (that we may be more interested in)
- Simplify the transcribe procedures
- More compact network structure

More Elegant Solution: Direct Note-level Transcription

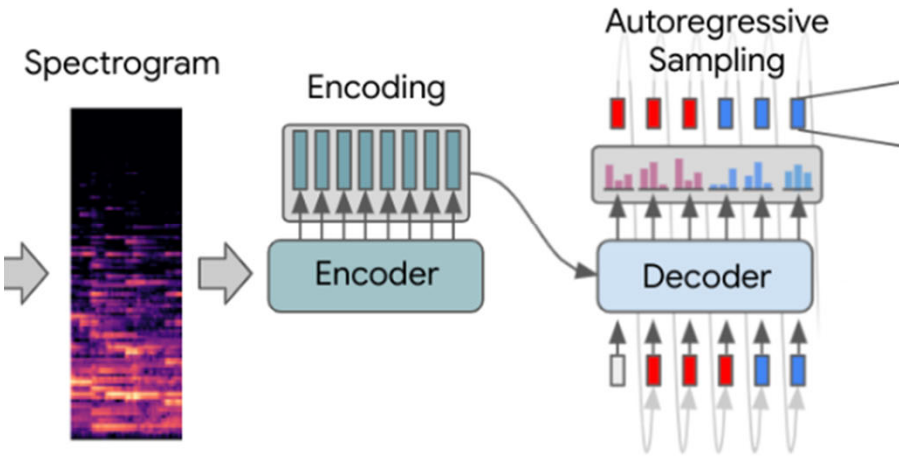
Sequence-to-sequence piano transcription

Hawthorne, Curtis, et al. “Sequence-to-Sequence Piano Transcription with Transformers.” *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*

More Elegant Solution: Direct Note-level Transcription

Sequence-to-sequence piano transcription

- Model: Encoder-decoder Transformer

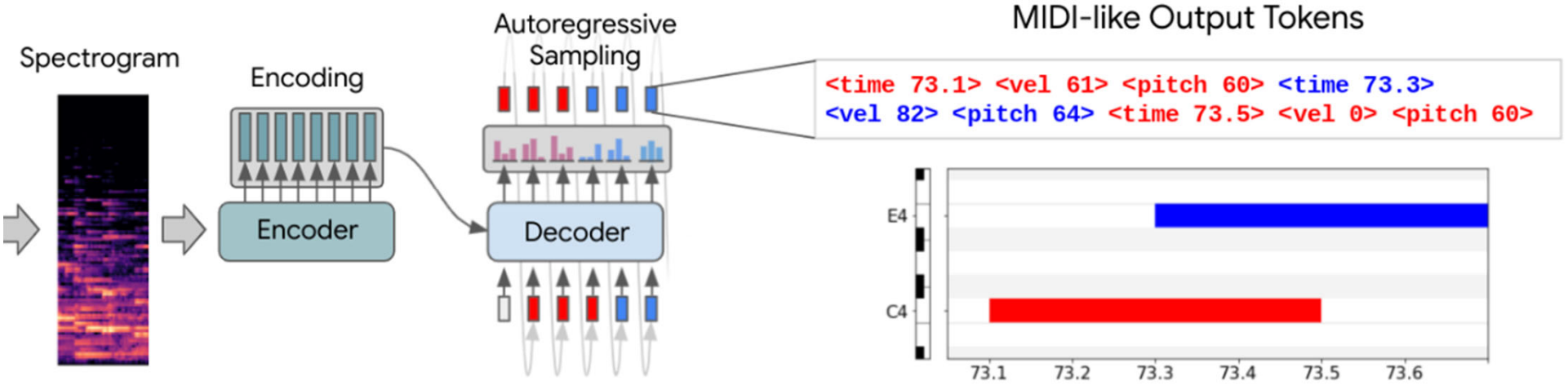


Hawthorne, Curtis, et al. "Sequence-to-Sequence Piano Transcription with Transformers." *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*

More Elegant Solution: Direct Note-level Transcription

Sequence-to-sequence piano transcription

- Model: Encoder-decoder Transformer
- Model output: MIDI-like note event
  - (Time, velocity, pitch) describe one onset/offset



Hawthorne, Curtis, et al. "Sequence-to-Sequence Piano Transcription with Transformers." *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*

Towards More instruments, and multi-instrument transcription

Multi-task Multitrack Music Transcription

- An extension of seq-to-seq piano transcription
- **Multi-task**: able to transcribe audio **from different instruments**, even a combination of them.
- **Multitrack**: when transcribing a multi-instrument piece, it output the transcription for different instruments **with only one pass**.

Gardner, Josh, et al. "MT3: Multi-Task Multitrack Music Transcription." *ICLR 2022*.

Future Direction of AMT

- Music Language Model
  - Some specific note sequence are more likely to appear, others are not
- Context-specific transcription
  - If we know the timbre of the instrument beforehand, how to take advantage of this to obtain more accurate transcription
- Transcribing expressive techniques
  - E.g., Portamento, vibrato
- etc.

Benetos, Emmanouil, et al. "Automatic music transcription: An overview." *IEEE Signal Processing Magazine* 36.1 (2018): 20-30.

Main Takeaways

- Automatic music transcription systems have broad applications.
- The methodology to solve singing voice transcription problem, including both signal processing and neural networks.
- Polyphonic music transcription remains a challenging problem which requires innovative solutions.