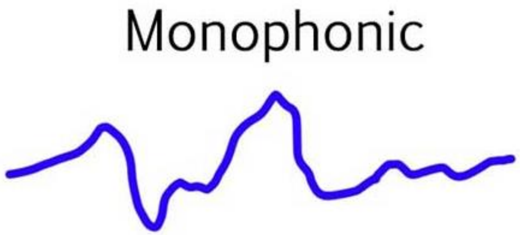
Deep salience representations for f0 estimation in polyphonic music

Bittner, Rachel M., et al. "Deep Salience Representations for F0 Estimation in Polyphonic Music." ISMIR. 2017.

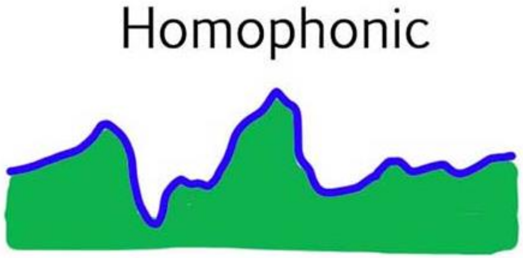
경영과학연구실 이태헌
2023.07.09

Music Texture

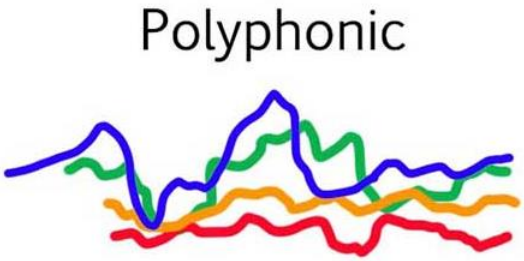
- **Depending on the structure of the music and the combination of sounds, music can be categorized into monophonic, homophonic, and polyphonic**



- Music that emphasizes only one pitch (i.e., one pitch or note) at a time



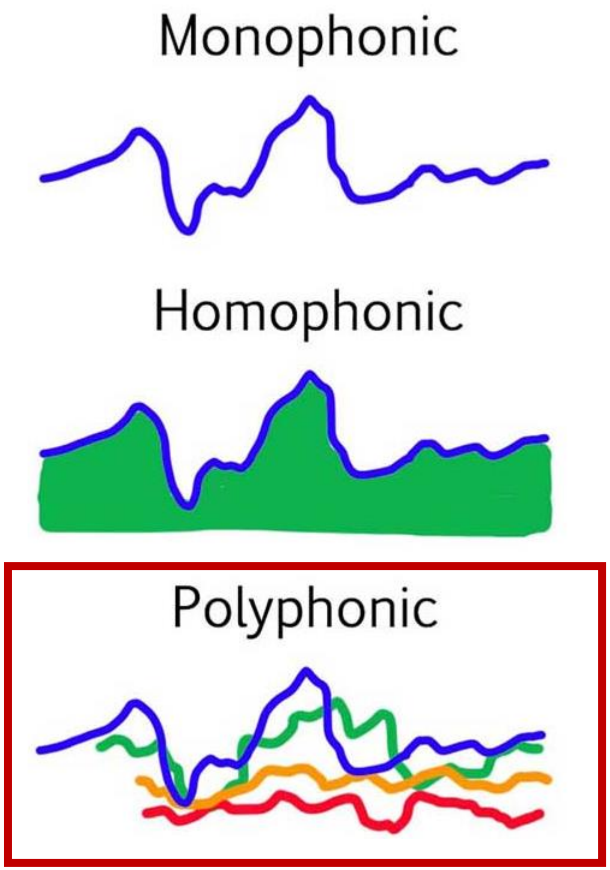
- It consists of multiple tones played simultaneously, forming specific "harmony" or "chords."
- One melody plays a prominent musical role, while the other tones constitute the background for this melody



- Each of these melody lines can have its own individual theme
- Characterized by its complexity and richness, as it is composed of multiple independent melodies interacting with each other

Music Texture

- **Depending on the structure of the music and the combination of sounds, music can be categorized into monophonic, homophonic, and polyphonic**

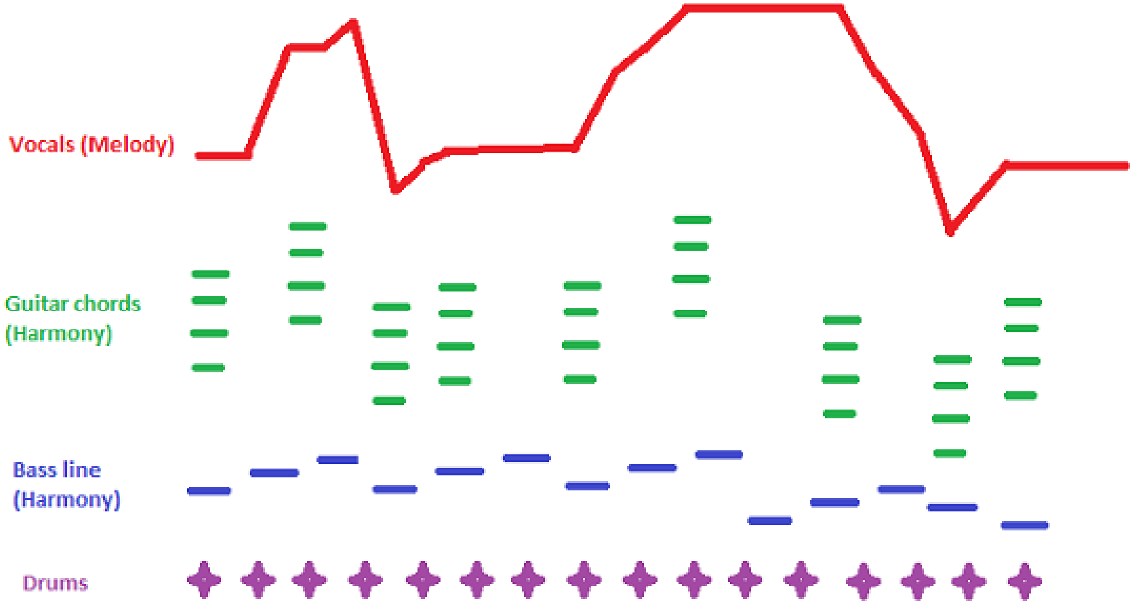


- Each of these melody lines can have its own individual theme
- Characterized by its complexity and richness, as it is composed of multiple independent melodies interacting with each other

The features and differences of multiple f0 estimation and melody Extraction

- **The F0 (Fundamental Frequency) line refers to the line or curve that represents the fundamental pitch of a melody or vocal line in music**

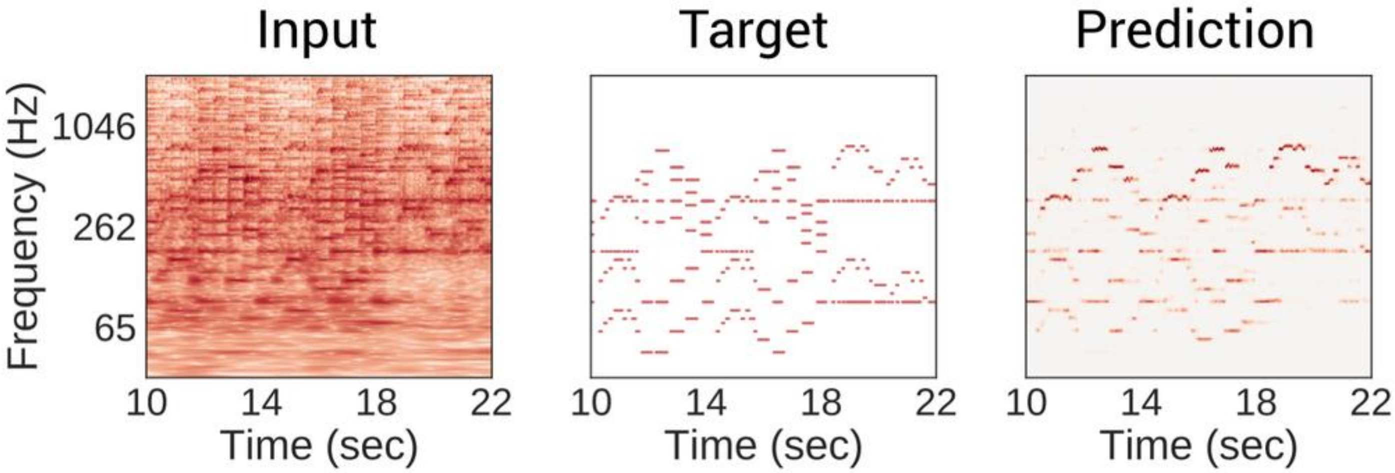
- **Multiple f0 Estimation** : estimating the fundamental frequency (F0) of all simultaneously played pitches in music
- **Melody Extraction** : Tracking the pitch (i.e., frequency) of the main melody line in music



Salience Representation

- **Salience is a concept that describes how important or prominent certain information is compared to the surrounding information**
- **It is used to identify and emphasize important features in various types of data and contexts**

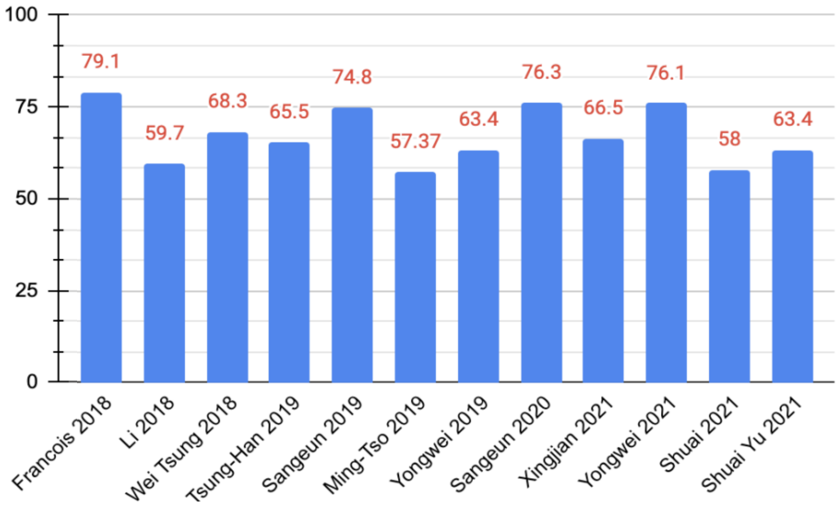
- Music and Speech Processing
- Computer Vision and Image Processing
- Natural Language Processing
- Cognitive Science



The difficulties in multiple-f0 estimation and melody extraction

- The performance of models used for melody extraction has been low
- The Melody DB dataset is comprised of complex music tracks designed for melody extraction

Raw pitch accuracy of the melody extraction models on Melody DB dataset

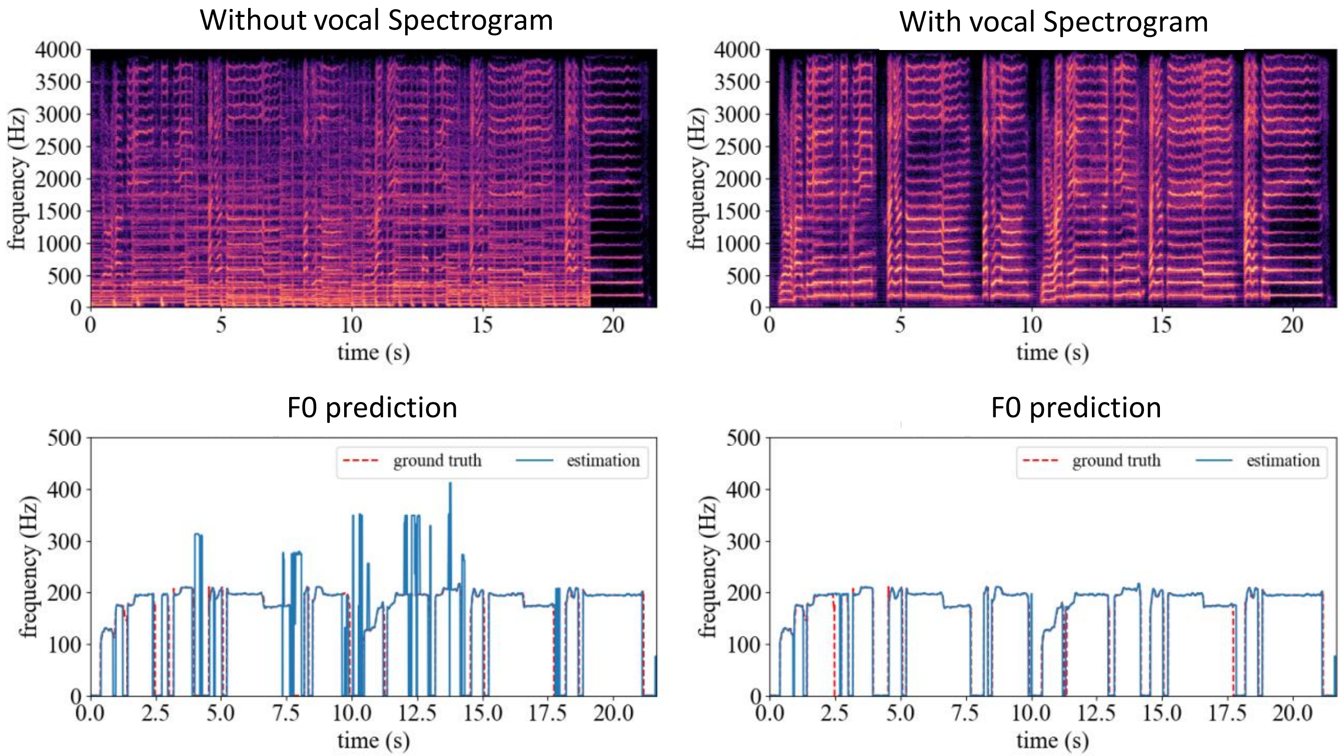


Distinguishing and tracking individual notes in polyphonic music is a highly complex task

Rao, K. Sreenivasa, and Partha Pratim Das. "Melody extraction from polyphonic music by deep learning approaches: A review." *arXiv preprint arXiv:2202.01078* (2022)

F0 Estimation and Melody extraction are relatively easier in music that includes vocals

- **In music with vocals, the primary melody (F0) is determined by the vocals**
- **The vocalist establishes and guides the main melody, providing a clear reference for the fundamental pitch**



Without vocal Spectrogram

With vocal Spectrogram

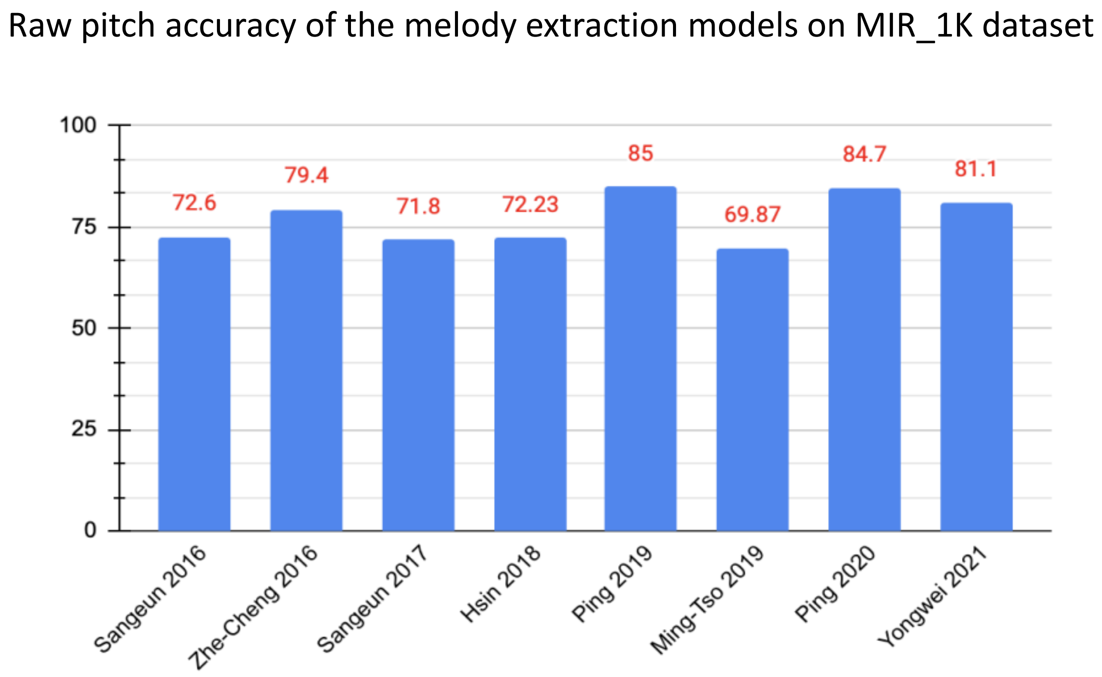
F0 prediction

F0 prediction

Gao, Yongwei, Xulong Zhang, and Wei Li. "Vocal melody extraction via hrnet-based singing voice separation and encoder-decoder-based f0 estimation." *Electronics* 10.3 (2021)

F0 Estimation and Melody extraction are relatively easier in music that includes vocals

- **The MIR_1K dataset is a dataset that includes both vocals and background music**

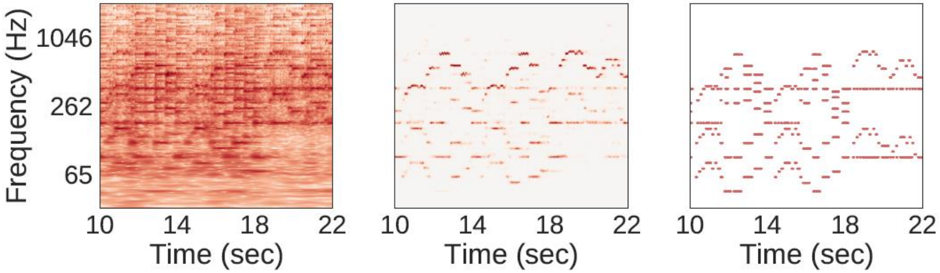


Rao, K. Sreenivasa, and Partha Pratim Das. "Melody extraction from polyphonic music by deep learning approaches: A review." *arXiv preprint arXiv:2202.01078* (2022)

Problem statement

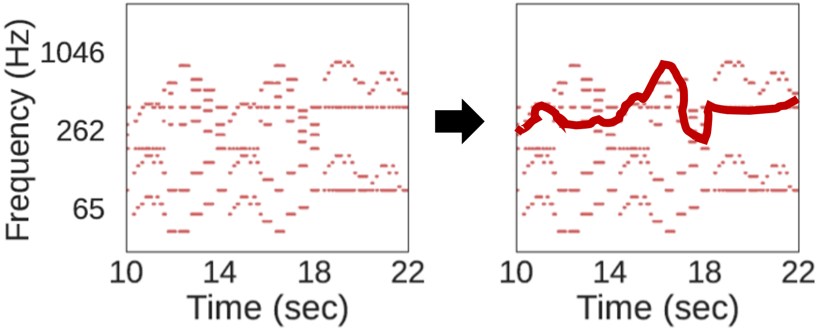
- **The objective is to address the problems of multiple-F0 estimation and melody extraction in polyphonic music using deep salience representation**

- Multiple-f0 estimation



- Melody extraction

Extracting the F0 line with the highest salience among the estimated multiple F0s



Key idea

Deep Salience representation using CNN model

- Training a CNN model to learn a salience representation that can accurately detect melodies (or fundamental frequencies, F0) despite the complexity of the music

The Harmonic Constant-Q Transform (HCQT) is used as the input

- HCQT is used to generate the time-frequency
- HCQT is effective in directly measuring harmonics in each frequency band, which allows for better emphasis and detection of melodies

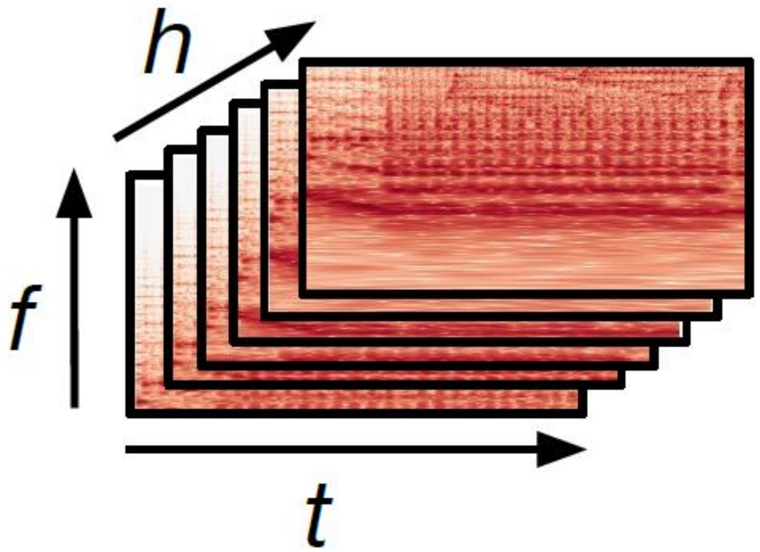
A common framework for multiple F0 estimation and melody extraction

- A common framework is provided for both multiple F0 estimation and melody extraction
- which helps to better emphasize and identify melodies in complex music

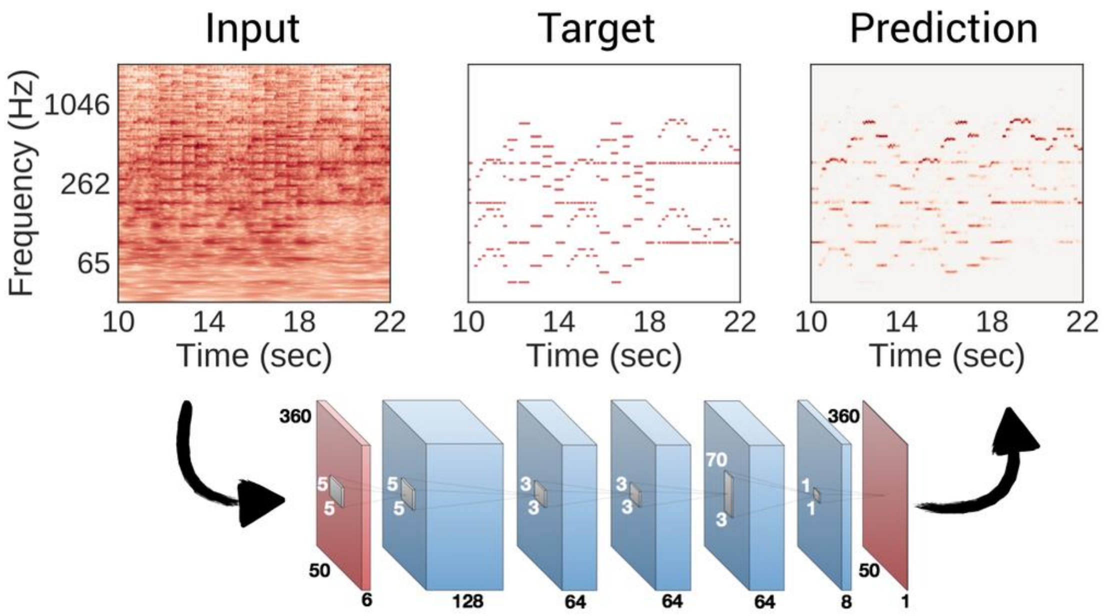
Harmonic constant-Q transform (HCQT)

- **The HCQT is a 3-dimensional array indexed by harmonic, frequency, and time: [h; t; f], measures the h th harmonic of frequency f at time t.**
- **HCQT is effective in analyzing multiple characteristics of simultaneous sounds in complex polyphonic music**

Harmonic constant-Q transform (HCQT)



Model Architecture



CNN Model

Example image of the output salience map

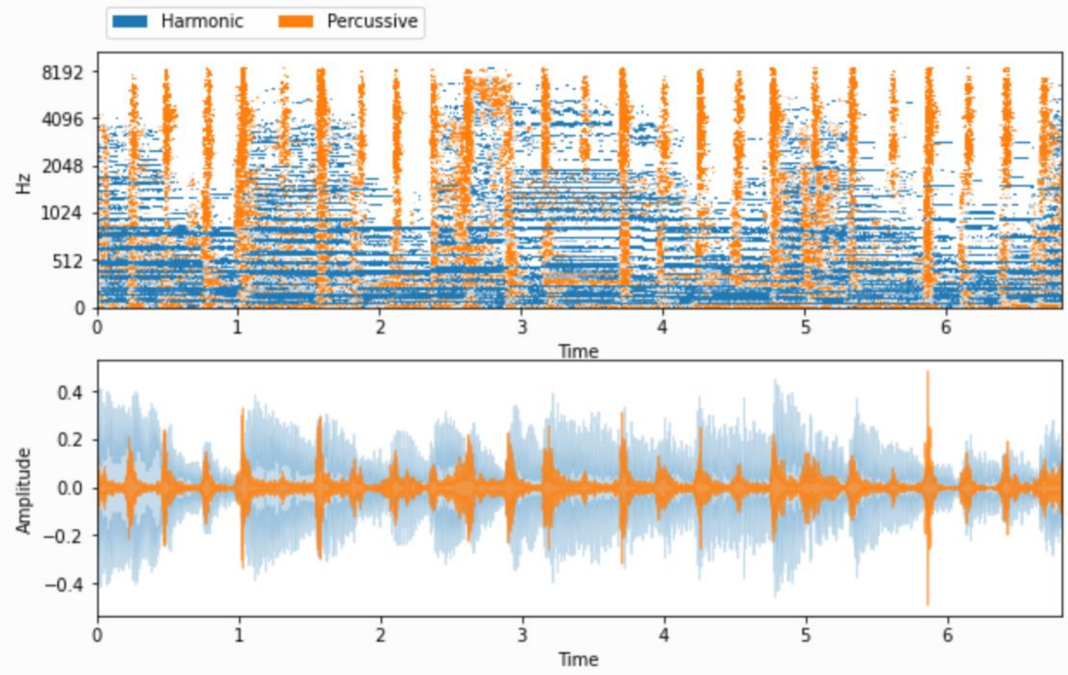
bin	0	0	0	0	0	0
0	1	0	0	0	0	0
1	1	1	0	0	1	1
1	0	1	1	0	1	1
0	0	0	0	1	0	1
0	0	0	0	1	0	0

Bin : Each pixel in the time-frequency representation of the signal

Salience representation

- Computations of salience representations usually perform two functions:
  (1) de-emphasize un-pitched or noise content
  (2) emphasize content that has harmonic structure

- Using a CNN allows for the joint learning of parameters for both the noise reduction stage and the harmonic enhancement stage



Dataset

- **The usage and validation datasets are the datasets used for evaluating the performance of the melody extraction algorithm**

Training dataset

- **Melody DB**

The dataset used for training is the Melody DB dataset, which provides music tracks spanning various genres and instruments

Validation dataset

- **Melody DB**

- **Bach 10**

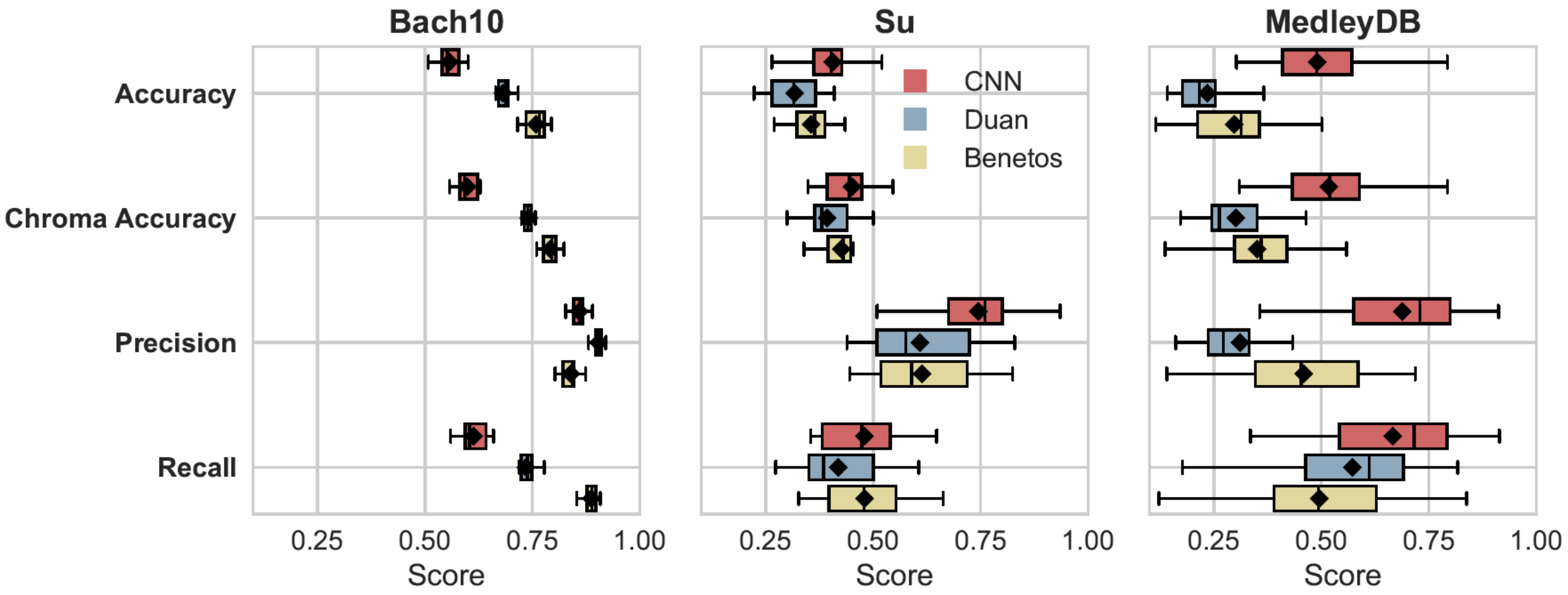
Bach10 is used for evaluating the performance of melody extraction algorithms in classical music

- **Su**

Su dataset consists of multi-track music extracted from Western pop music

MULTIPLE-F0 estimation Experiments

- **Benetos and Duan, used for comparison are models developed for multiple F0 estimation**
- **Overall, the proposed model in the paper demonstrates good precision and stable chroma accuracy**



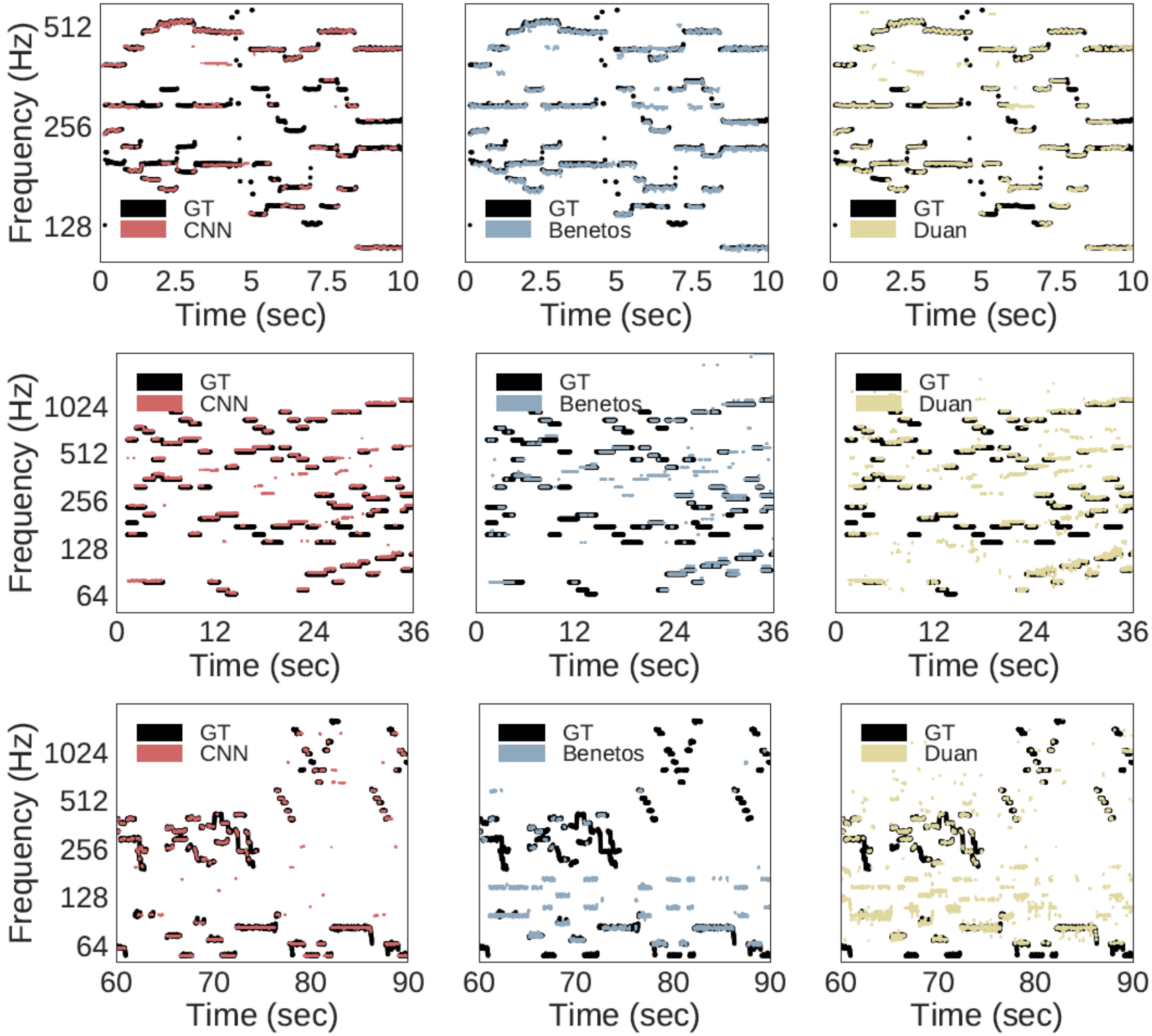
Chroma accuracy : Measures the model's ability to accurately estimate the chroma information of music. It indicates the ratio of correctly estimated chroma information to the total number of samples

Multiple f0 output for each of the 3 algorithms

Bach 10

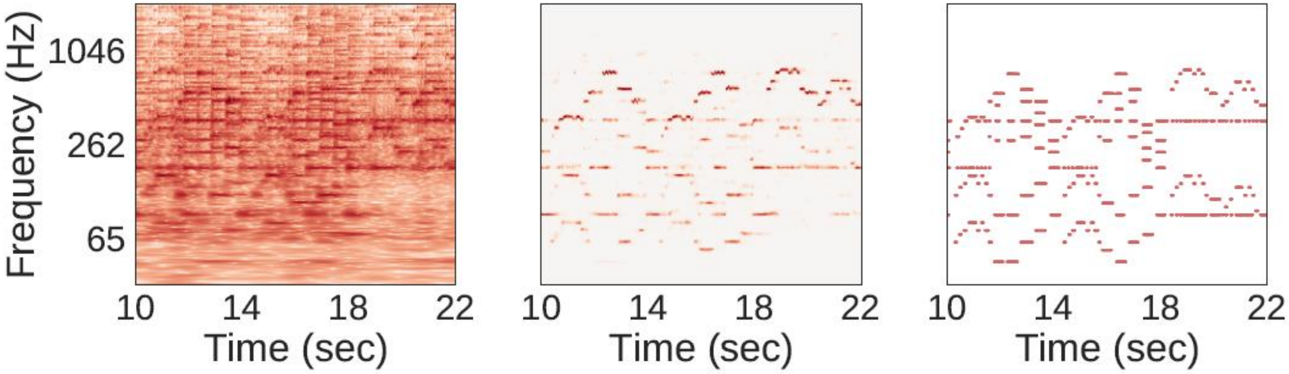
Su

Melody DB



Multipe-f0 estimation Salience representation result

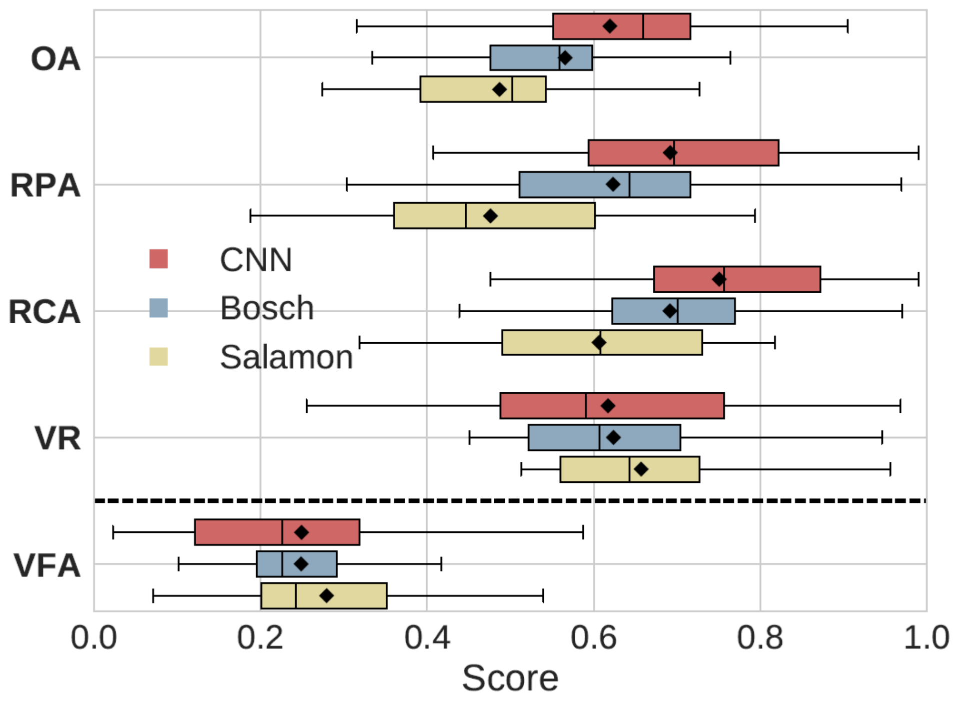
- **CNN's output for unseen tracks in the Su dataset**



(left) Input $\mathcal{H}[1]$, (middle) predicted output, (right) ground truth annotation for an unseen track in the Su dataset.

Melody Extraction Experiments result

- The outputs of the CNN-based system are compared with these two baseline Melody extraction algorithms to assess its performance
- Salamon is a heuristic algorithm that has maintained a high level of performance in melody extraction
- Bosch combines heuristic rules with the salience function to achieve the highest level of performance



Overall Accuracy (OA)

Raw Pitch Accuracy (RPA)

Raw Chroma Accuracy (RCA)

Voicing Recall (VR)

Voicing False Alarm (VFA)

Conclusion & Further work

Conclusion

- In this paper, a complete convolutional neural network (CNN) model is proposed to learn the salience representation for multiple F0 tracking and melody extraction
- The model demonstrates that by simply decoding the salience representation, state-of-the-art results can be achieved in multiple F0 tracking and melody extraction.

Further work

- If a sufficient amount of training data is provided, this architecture can be useful for related tasks such as bass, piano, guitar, and more
- To further improve the performance of the system, data augmentation techniques can be employed to diversify the training set and balance the class distributions

Q & A