Automatic Music Transcription using Convolutional Neural Network (CNN) and Constant-Q Transform (CQT)

Authors:

Yohannis Kifle Telila

Prof. Tommaso Cucinotta

Prof. Davide Bacciu

Introduction

Automatic music transcription (AMT) is the problem of analyzing an audio recording of a musical piece and detecting notes that are being played.

Transcribing music has many significant applications.

- Transcribed music helps musicians avoid memorization.
- Guitarists can use a guitar pickup and AMT to play various sounds from their guitar.
- Real-time AMT systems can be used for education, such as an app that listens to a student playing piano and spots mistakes in real-time.

CONT . . .

Key challenges:

- Polyphonic music is a complex mix of instruments and vocals.
- Multi-pitch estimation can be difficult as the individual harmonics of notes played simultaneously can cancel each other out.
- Lack of a general and sizable dataset for training and evaluation.

Motivation

In general, AMT is considered a single and unified system responsible for transcribing all the notes/chords in a musical piece.

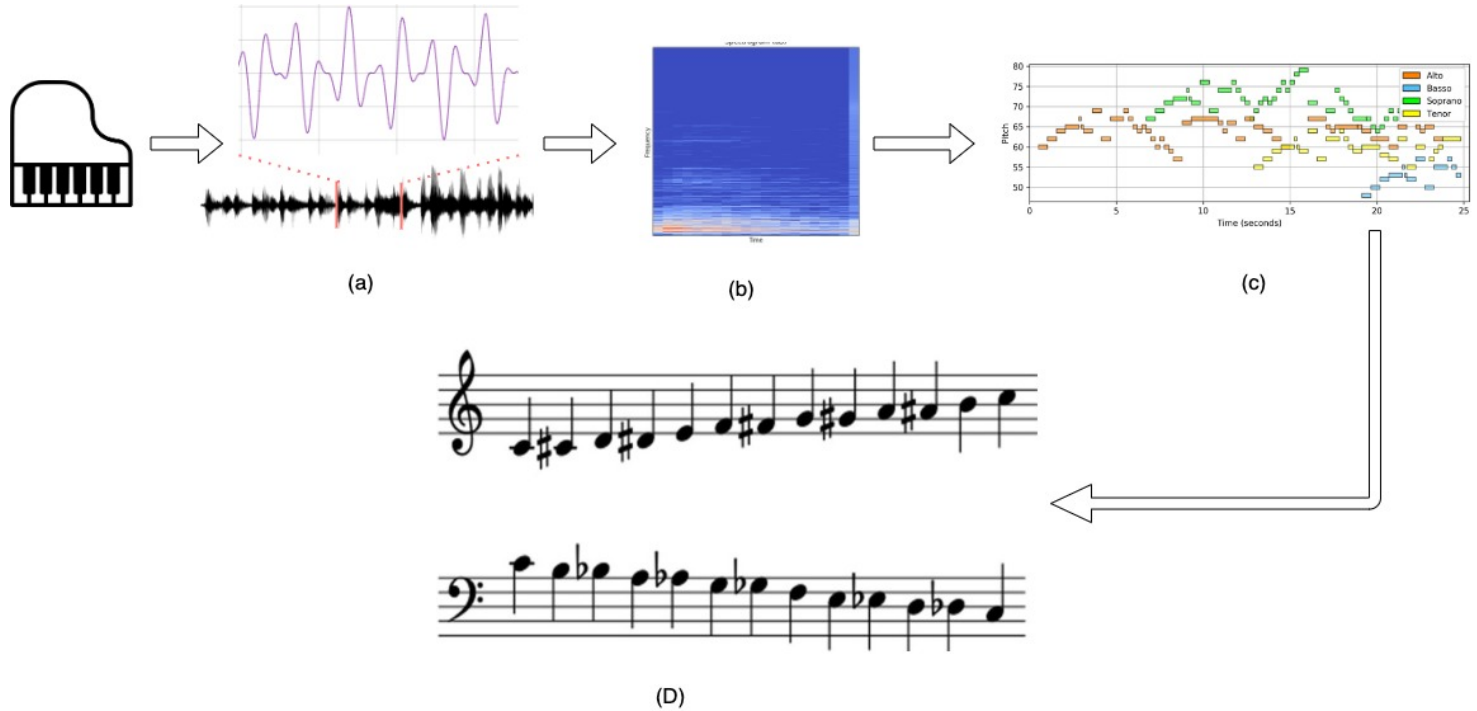

Figure: General automatic music transcription stages.

Proposed Method

The paper proposes a *structured approach* to transcribe piano music using CNN and CQT [1].

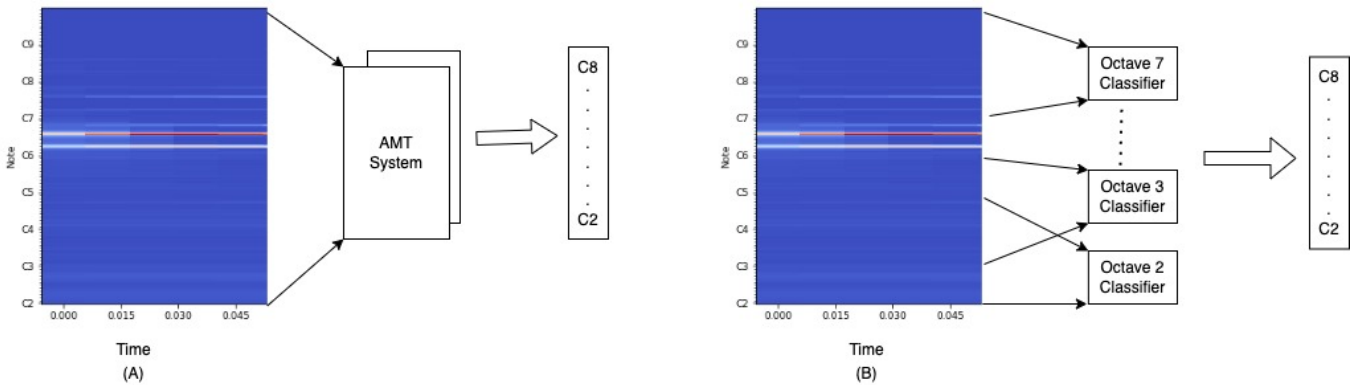

Figure: A) Holistic approach of AMT B) Structured Approach of AMT

Structured Approach of AMT

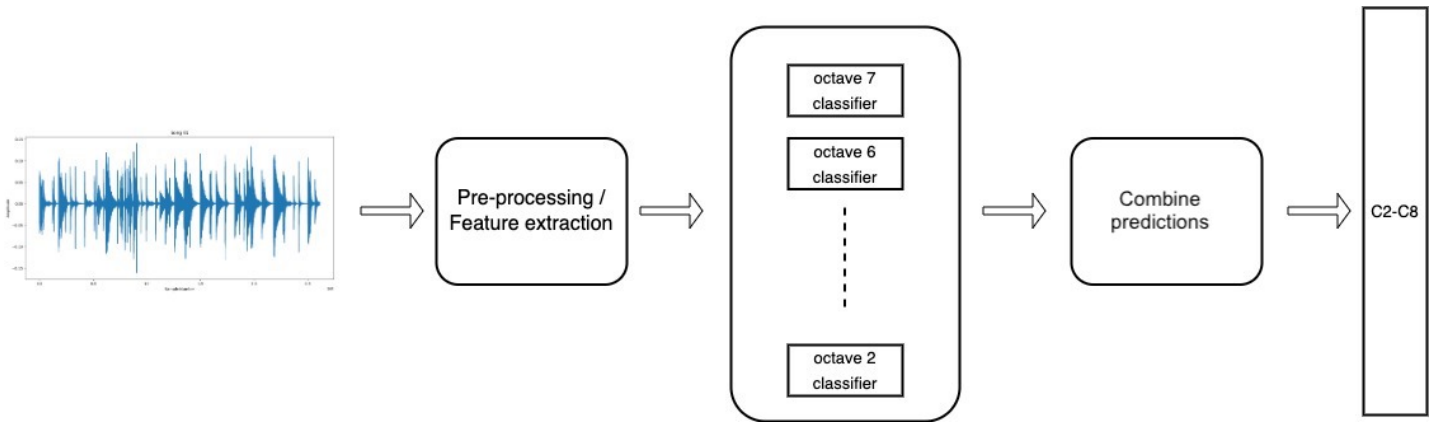

Figure: General architecture of the proposed system.

$$FocalLoss = -\sum_{i=1}^{i=n}(1-p_i)^{\gamma} log_b(p_i) \qquad [2]$$

Focal Loss reduces the influence of easy examples on the loss function, resulting in more attention being paid to hard training example.

Silent Frames

Piano note sound energy decay affects AMT accuracy.

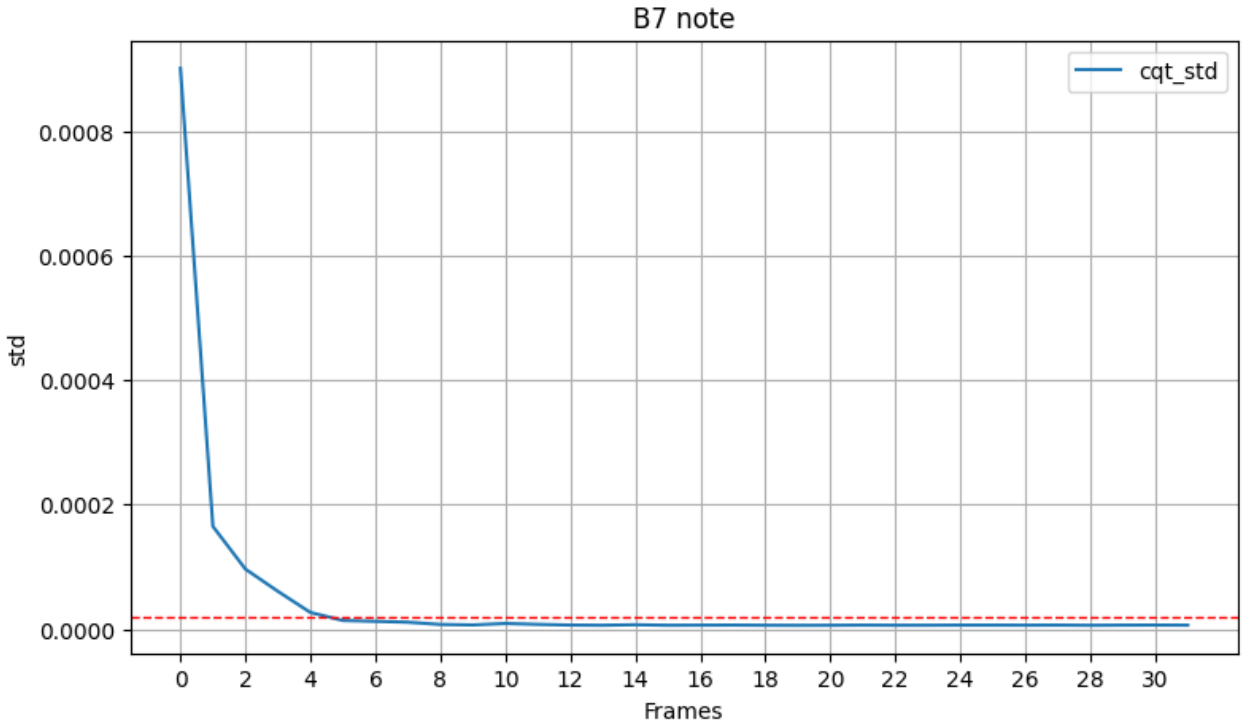

Figure: standard deviation of frames for note B7 - Threshold = 0.00001

Decomposing AMT task

Decomposing AMT task requires careful consideration

- Structured v1 (*2 octaves above*)
- Structured V2 (*Structured V1 with no silent frames*)
- Structured V3 (*2 octaves above, 1 octave below and no silent frames*)
- Holistic v1 (*with silent frames*)
- Holistic V2 (*no silent frame*)

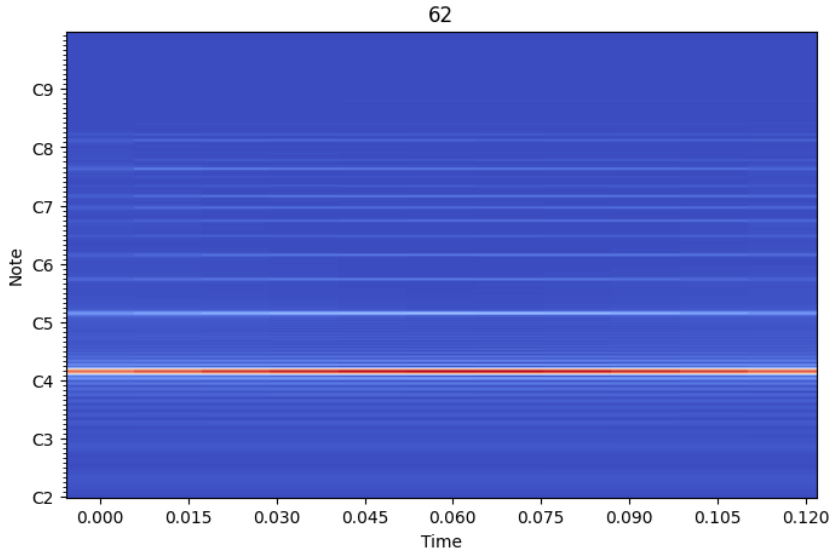

Figure: CQT output of note 62(D4) frame.

Training Procedure

Synthetically generated piano note/chord progression is used for training.

Chord selection included major and minor triads, 7th chords, 6th chords, diminished triads, suspended chords, and augmented chords.

830 3-minute piano notes/chord progressions generated. 80% for training, 20% for validation, and 166 3-minute progressions for testing.

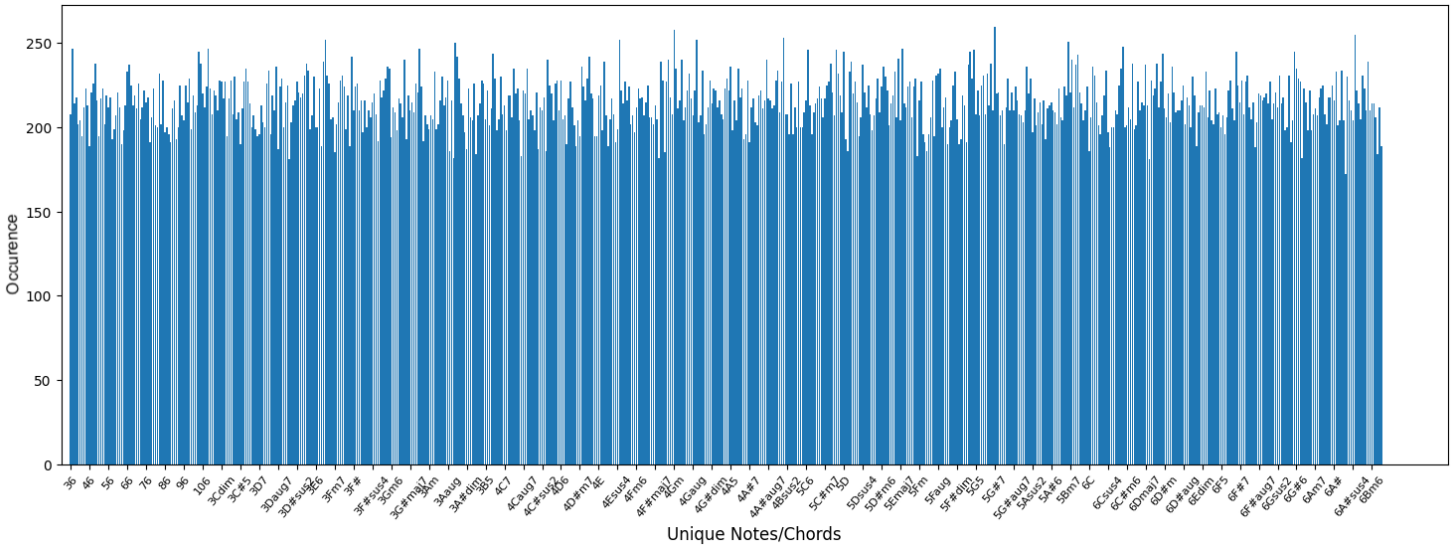

Figure: Dataset histogram

Results and comparison

Model	Training time(minutes)	Inference time(msec)	Accuracy(%)
Holistic - v1	793.6	660	88.2
Holistic - v2	720.2	664	92.5
Structured - v1	118.3	230	82.1
Structured - v2	116.3	215	84.0
Structured - v3	**246.7**	**241**	**86.4**

Table: Model result summary

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i)$$

Accuracy vs Model Parameter Count

Comparison of the model complexity and accuracy for each model.

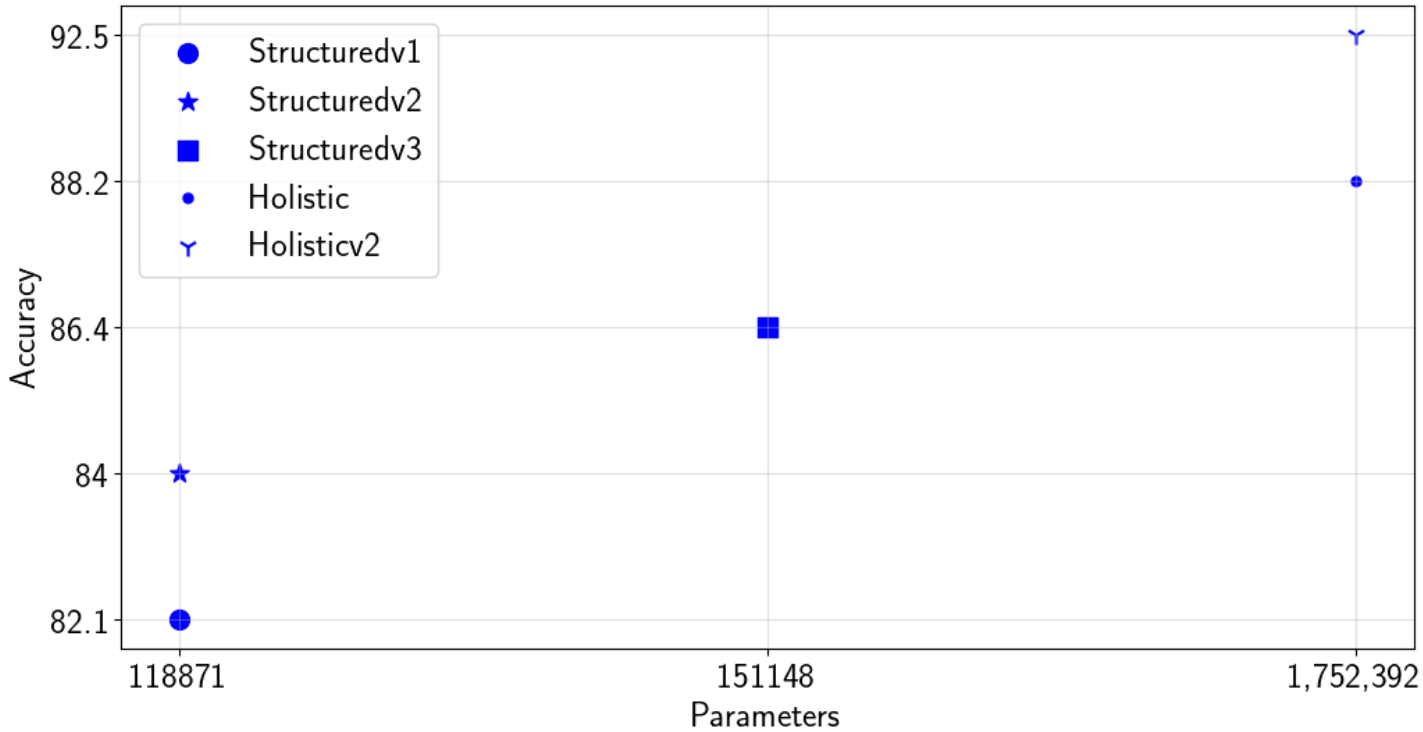

Figure: Models comparison

Octave accuracy

Accuracy comparison of both holistic and structured models across the six octaves.

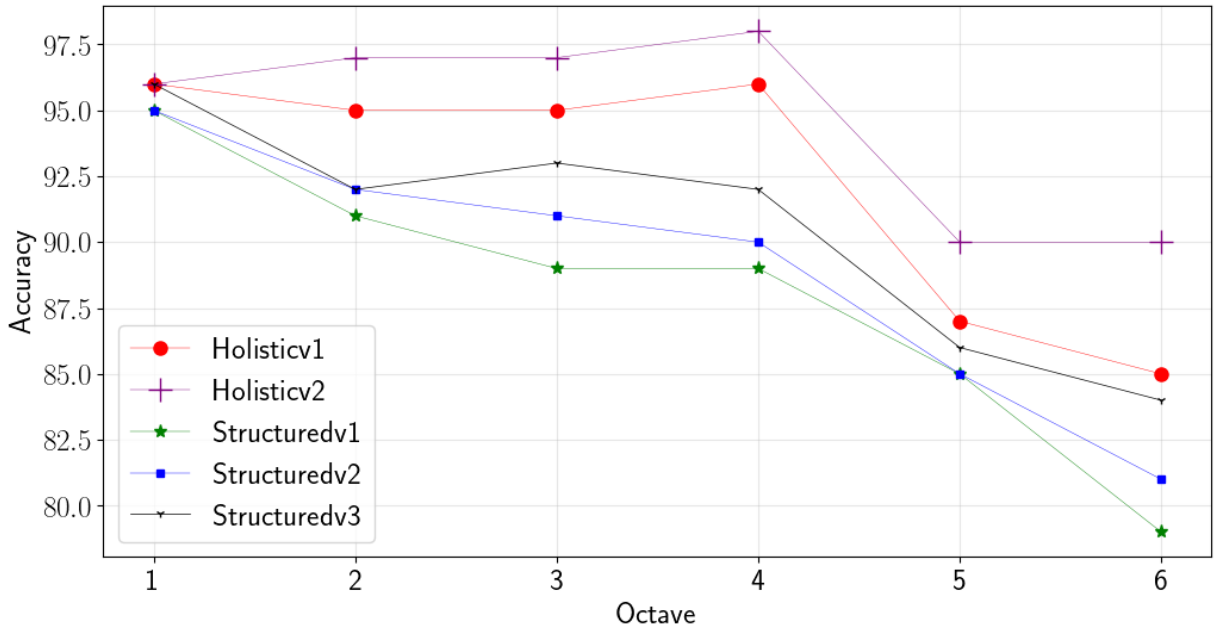

Figure: Prediction accuracy across the octaves.

Chord accuracy

Accuracy comparison of models on different frame types: single notes, 2-note chords, 3-note chords, and 4-note chords.

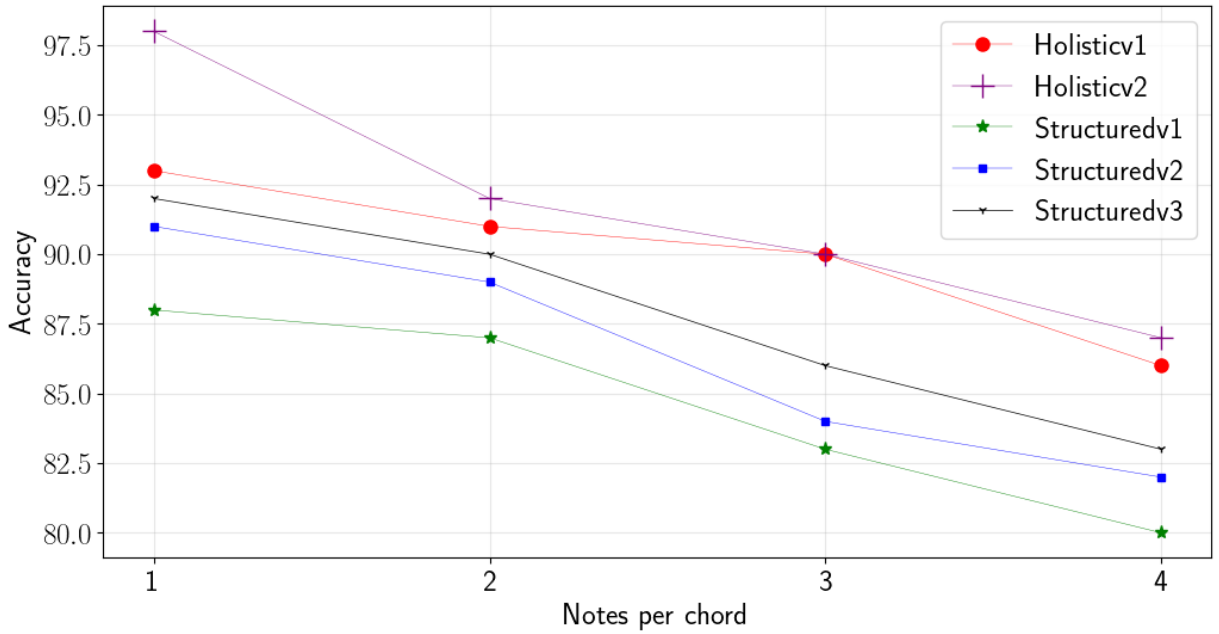

Figure: Number of notes in a chord vs prediction accuracy

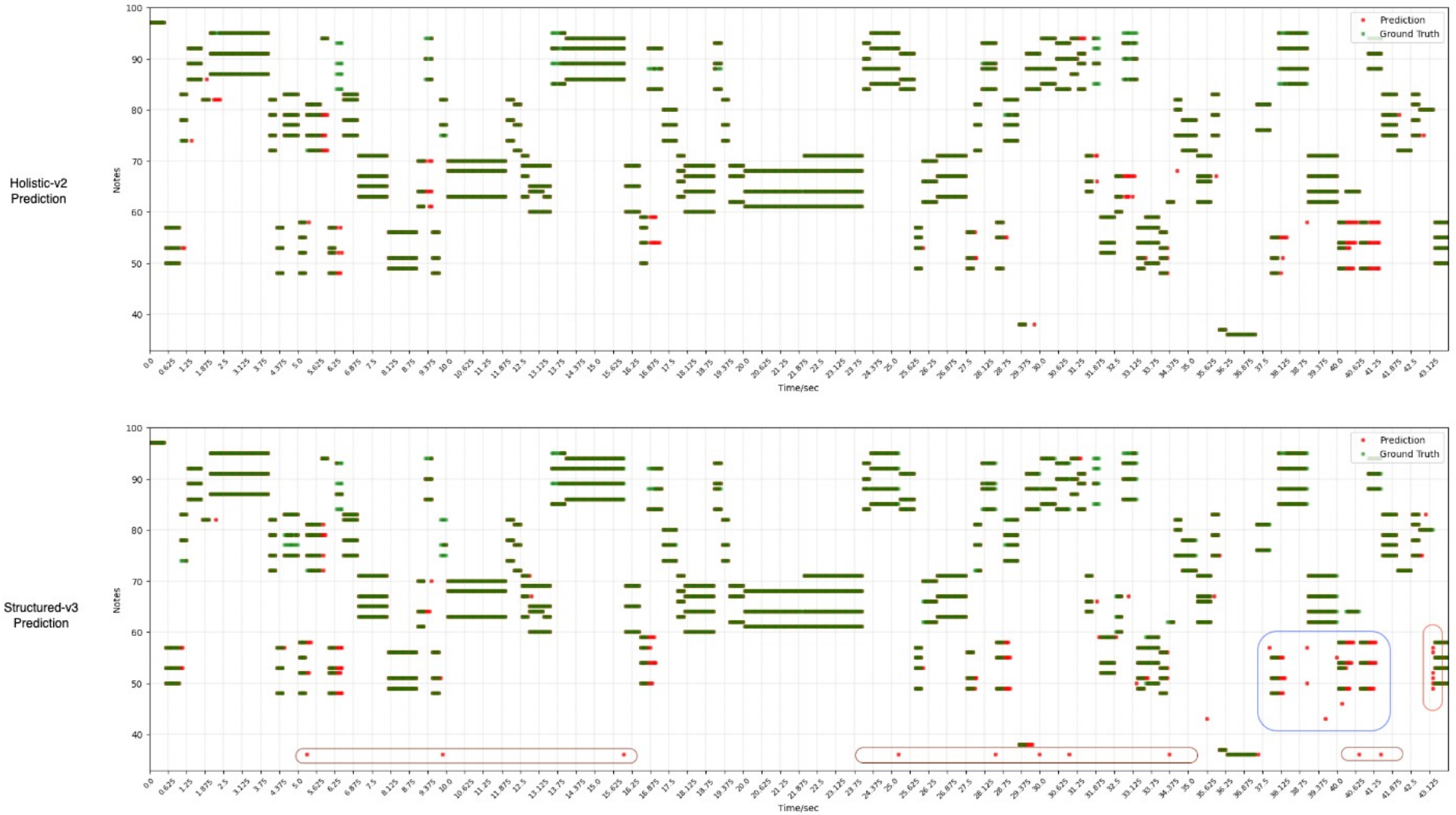

Figure: Piano roll prediction visualization

Conclusion

This study aimed to evaluate the feasibility of breaking down music transcription into smaller and manageable tasks.

Structured models offer a good balance between accuracy and efficiency in decomposing the music transcription task into smaller models.

Holistic models were slightly more accurate, but structured models were deemed more practical and efficient for real-time music transcription tasks.

This approach could be beneficial for developing more efficient and effective music transcription systems in the future.

Future Works

Training the AMT system on real-world piano audio data can enhance its performance in practical applications by providing a more diverse set of training examples.

Exploring other models like RNN to improve accuracy by capturing longer dependencies between notes and chords.

An attention mechanism can also be used to effectively model long-term characteristics without increasing the network size [3] [4].

References (I)

[1] Christian Schörkhuber. Constant-q transform toolbox for music processing. 2010.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. `https://arxiv.org/pdf/1708.02002.pdf`, 2018.

[3] Sehun Kim, Tomoki Hayashi, and Tomoki Toda. Note-level automatic guitar transcription using attention mechanism. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 229–233, 2022. doi: 10.23919/EUSIPCO55093.2022.9909659.

[4] Julian Salazar, Katrin Kirchhoff, and Zhiheng Huang. Self-attention networks for connectionist temporal classification in speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2019. doi: 10.1109/icassp.2019.8682539.

THANK YOU!