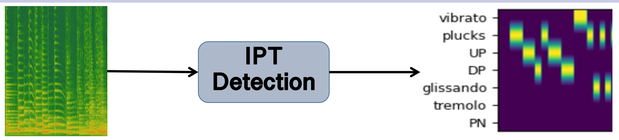
FRAME-LEVEL MULTI-LABEL PLAYING TECHNIQUE DETECTION USING MULTI-SCALE NETWORK AND SELF-ATTENTION MECHANISM

Dichucheng Li, Mingjin Che, Wenwu Meng, Yulun Wu, Yi Yu , Fan Xia and Wei Li
Email: dccli21@m.fudan.edu.cn

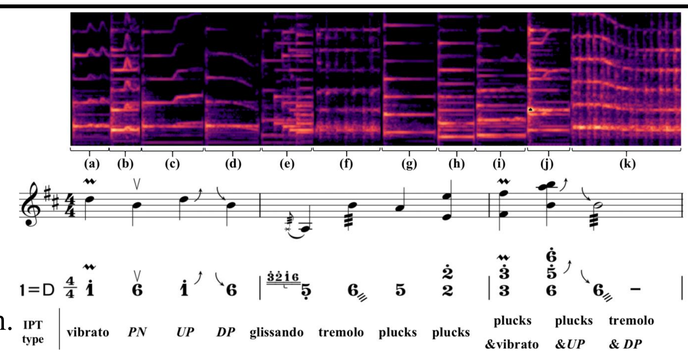
Instrument Playing Technique (IPT) Detection



- classify IPT types & locate the associated boundaries.
- **Application:** complete transcription system, performance analysis.

Contribution

1.Problem:
Frame-level multi-label classification problem.
2.Dataset:
recorded and annotated Guzheng solo pieces (polyphonic).
3.Method:
a new model using multi-scale network and self-attention mechanism.



Paper

GitHub

Dataset

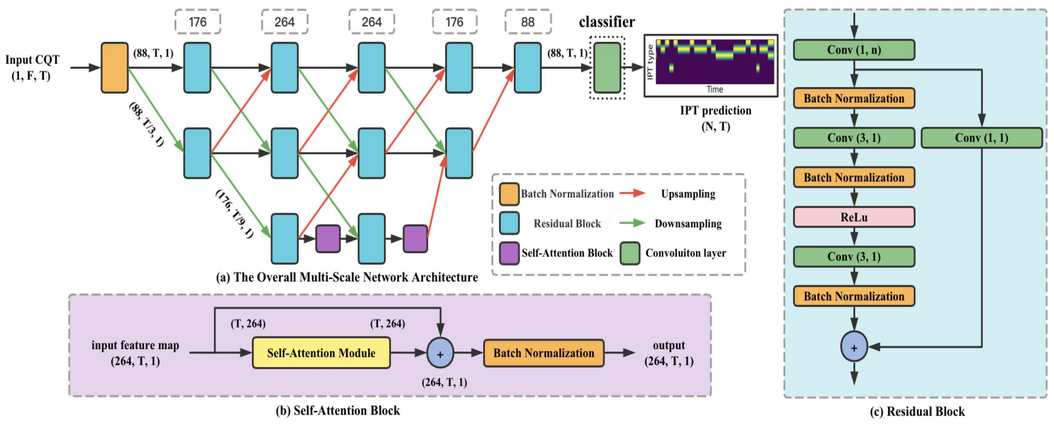
IPT	num	length (seconds)			
		sum	mean	max	min
vibrato	1994	1650.31	0.83	4.37	0.21
UP	756	544.12	0.72	3.84	0.10
DP	208	126.56	0.61	3.44	0.19
PN	209	153.12	0.73	3.24	0.23
glissando	734	67.54	0.09	0.39	0.03
tremolo	77	152.75	1.98	4.67	0.21
plucks	11860	7066.19	0.60	6.82	0.07

IPT description:
- Mixed IPTs, Overlapping IPTs.
- 7 independent IPTs.

Data Collection and Labelling:
- 99 Guzheng solo recordings, 9064.6 seconds long.
- 2 professional Guzheng players.
- label the onset, offset, pitch and IPTs of every note, 63,352 annotated labels in total.

Dataset splitting:
- 79, 10, 10 songs (8:1:1).
- control the distribution of IPT types and performers.

Method



Multi-Scale Network:
- Long-range features are crucial for long IPTs, high-resolution features are necessary for short IPTs.
- down /upsampling the feature to different scales, fuse features with different resolution repeatedly.

Self-Attention Block:
- apply to the feature maps at the coarsest scale.
- capture interactions between different frames on the feature maps.
- further enhance the extraction of global features.

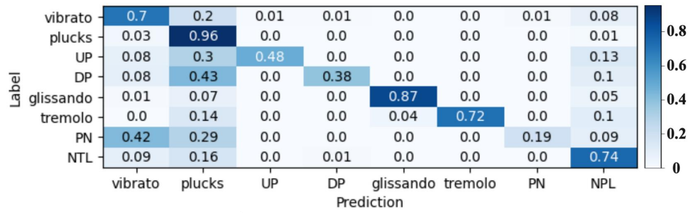
Experiment

Method	precision	recall	F1-score
w/o self-att	85.43	82.64	84.01
w/o res	86.38	83.45	84.89
Single-scale	79.04	75.02	76.98
Proposed	**87.62**	**85.48**	**86.54**

▶ Ablation studies with frame-level precision, recall and F1-score.

Method	precision	recall	F1-score
GZFNO [12]	69.49	68.00	68.73
EHFCN [11]	69.55	67.77	68.65
CNN+Res [20]	75.97	66.23	70.77
Proposed	**87.62**	**85.48**	**86.54**

▶ Results of the proposed and baseline methods with frame-level precision, recall and F1-score.



▶ The confusion matrix

- plucks has the highest true positive proportion.
- DP (Downward Portamento) is often misclassified as plucks.
  - data imbalance; DP can be overlapped with plucks or mixed with tremolo.
- PN (Pitch Note) is prone to misclassification as vibrato.
  - PN can be regarded as a special type of vibrato with only one pitch change.