PLAYING TECHNIQUE DETECTION BY FUSING NOTE ONSET INFORMATION IN GUZHENG PERFORMANCE

Dichucheng Li
Yulun Wu
Qinyu Li
Jiahao Zhao
Yi Yu
Fan Xia
Wei Li

Sichuan Conservatory Of Music

Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

Overview

What is Guzheng & What is instrument playing technique (IPT) detection & Why IPT detection

- The Guzheng (古箏), which is also known as the Chinese zither, is a plucked 21-string Chinese musical instrument with diverse playing techniques.
- Instrument playing technique detection aims to classify the types of playing techniques and locate the associated playing technique boundaries in an audio clip.
- IPT detection can be utilized to form a more complete Automatic Music Transcription system that contains IPT information.

Motivation:

- Most of the existing works provide no assurance in the generalization as they rely on a single sound bank for training and testing;
- Existing methods have computational redundancy caused by the sliding window;
- Guzheng is a plucked string instrument which has a significant note onset and existing works show low accuracy at the boundary of adjacent notes.

Contribution:

- Create a new dataset from multiple sound banks and real-world recordings for Guzheng performance analysis;
- the first end-to-end method that can be applied to variable-length audio for Guzheng IPT detectiont ;
- A new decision fusion method fuses onset and IPT prediction to make a note-level prediction.

GZ_IsoTech Dataset

The traditional Guzheng playing techniques can be divided into two classes: plucking string with the right hand and bending string with the left hand. GZ_IsoTech contains 8 IPTs:
Left-hand IPT: vibrato, upward portamento (UP), downward portamento (DP), returning portamento (RP).
Right-hand IPT: glissando, tremolo, harmonic, plucks.

Data collection and labelling: we collected audio clips of single playing techniques from real-world recordings (RW) and 2 virtual Guzheng sound banks (VSB), covering 8 IPTs and all the tones, 3838.7 seconds in total. We use the short clips sampled in VSB as the training split and the real-world recording notes as the test split.

IPT	VSB		RW	
	num	seconds	num	seconds
vibrato	192	261.1	42	51.5
UP	488	752.0	48	94.4
DP	333	508.5	51	87.5
RP	272	327.0	94	94.7
glissando	316	595.3	42	95.9
tremolo	205	259.8	23	59.7
harmonic	318	305.8	61	42.0
plucks	204	204.0	135	99.7

▲ Quantity of data in the virtual sound banks (VSB) and real-world (RW) recordings.

Generating audio sequence: we concatenated the short clips one after another randomly until the length of the concatenated audio sequence is greater than 12.8 seconds. We cut the audio sequence in the training set directly to 12.8 seconds long and use the unsplit audio sequence for testing. Meanwhile, we generate the onset lables and the IPT labels.

Method

IPT Detector:

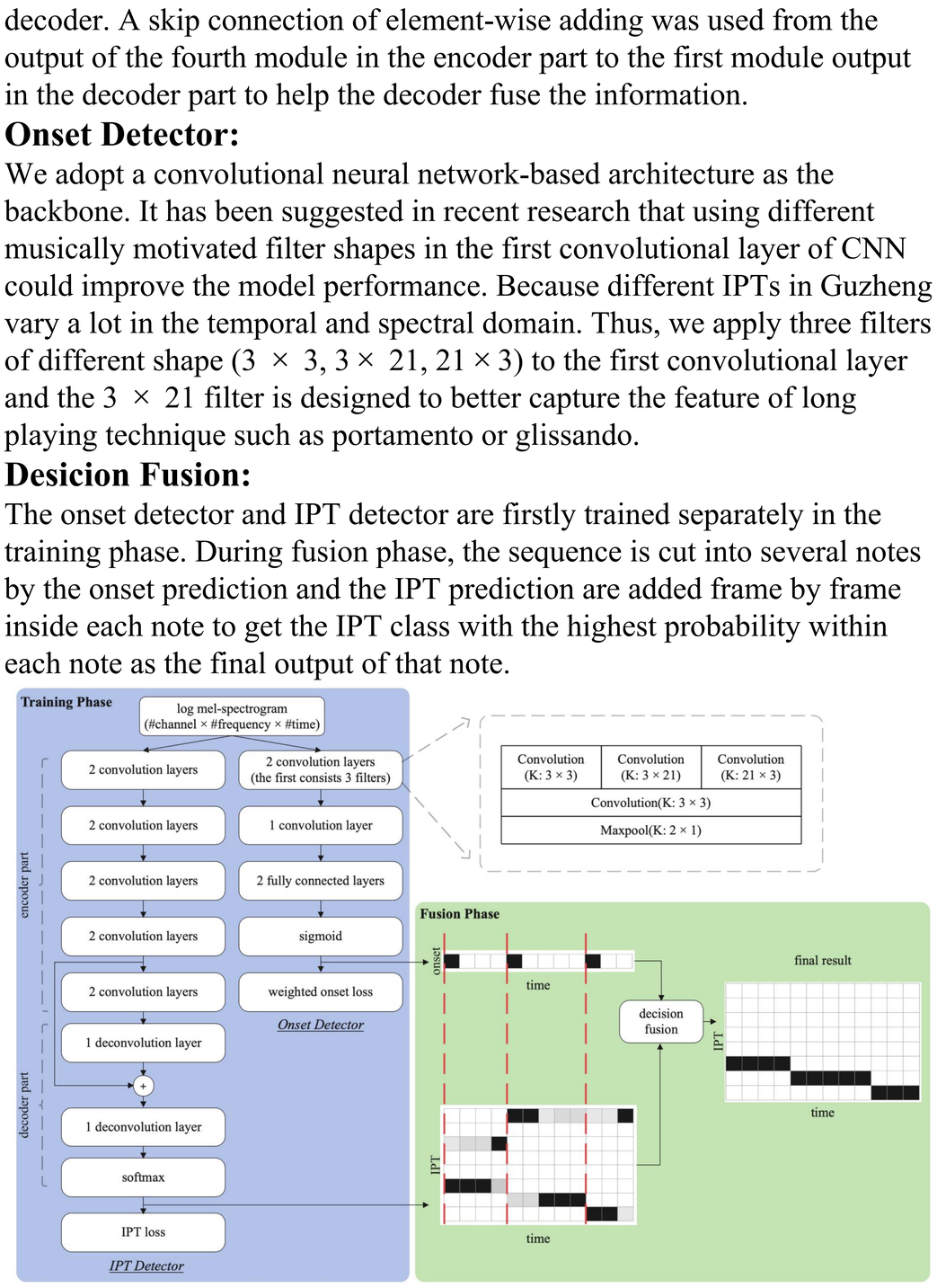
We can treat the IPT detection task as a semantic segmentation task since there is no overlap of playing techniques in our data and frames in music are similar to pixels in images. Inspired by the success of fully convolutional networks (FCN) in semantic segmentation, we applied FCN in our task. The IPT detector can be divided into the encoder and the decoder. A skip connection of element-wise adding was used from the output of the fourth module in the encoder part to the first module output in the decoder part to help the decoder fuse the information.

Onset Detector:

We adopt a convolutional neural network-based architecture as the backbone. It has been suggested in recent research that using different musically motivated filter shapes in the first convolutional layer of CNN could improve the model performance. Because different IPTs in Guzheng vary a lot in the temporal and spectral domain. Thus, we apply three filters of different shape (3×3, 3×21, 21×3) to the first convolutional layer and the 3×21 filter is designed to better capture the feature of long playing technique such as portamento or glissando.

Desicion Fusion:

The onset detector and IPT detector are firstly trained separately in the training phase. During fusion phase, the sequence is cut into several notes by the onset prediction and the IPT prediction are added frame by frame inside each note to get the IPT class with the highest probability within each note as the final output of that note.



▲ System architecture

Experiments

Results:

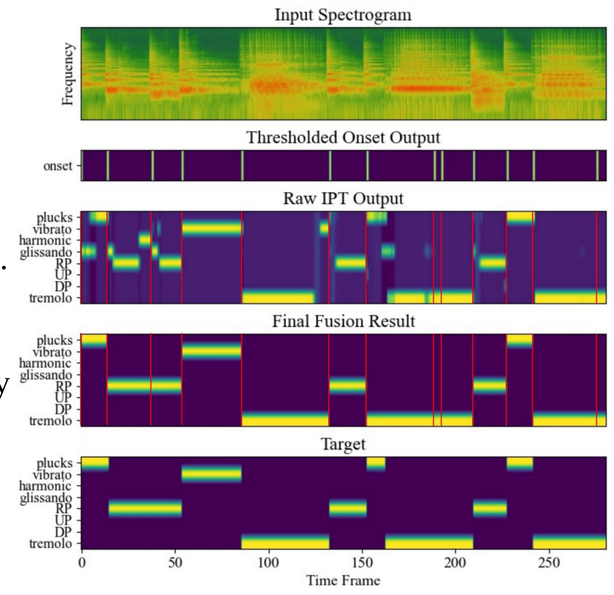
model	Frame-level	Note-level		
	accuracy	precision	recall	F1-score
FCN+Onsets	**87.97**	**78.20**	**83.78**	**80.76**
Z.Wang's model [10]	69.44	35.18	47.45	39.53
J.-F. Ducher's model_Reproduced [8]	66.93	19.37	21.31	20.05
B.Liang's model_Reproduced [12]	53.98	41.38	44.50	42.81

▲ Frame-level accuracy and note-level precision, recall and F1-score on GZ_IsoTech dataset.

Our FCN+Onsets model reaches 87.97% in frame-level accuracy and 80.76% in note-level F1-score, producing the best scores in both frame-level and note-level metrics. The results show that our model offers good generalization capabilities.

Sample Visualization:

The visualization of the input spectrogram, thresholded onset output, raw IPT output, final fusion result and the label for a recording from the test set. The vertical red lines in the third and fourth images are the onset predictions. The accuracy of the predictions in the "Final Fusion Result" image is far exceed that in the "Raw IPT Output" image, indicating the effectiveness of the decision fusion procedure.



▲ The visualization of the inference process on an audio in the test set.

Ablation Study:

model	Frame-level	Note-level		
	accuracy	precision	recall	F1-score
FCN+Onsets	**87.97**	78.20	**83.78**	**80.76**
CNN+Onsets	60.80	45.01	48.30	46.51
No Onset Fusion	76.48	27.74	50.55	35.47
Onset3×3	87.70	72.67	82.76	77.19
No Skip Connection	82.45	71.05	76.15	73.39
No Weighted Loss	86.26	**80.88**	79.46	80.03

▲ Ablation studies with frame-level accuracy and note-level precision, recall and F1-score on GZ_IsoTech dataset

"CNN+Onsets" use the same CNN architecture for the IPT detector as presented in the onset detector. "No Onset Fusion" is the output of the IPT detector without the decision fusion. "Onset3 × 3" replace the whole first convolutional layer of the onset detector by 3 × 3 filters. "No Skip Connection" remove the skip connection in IPT detector. "No Weighted Loss" is the model that uses ordinary binary cross entropy (BCE) Loss in onset detector. The results prove the effectiveness of our idea.