P5-08

Exploring Isolated Musical Notes as Pre-training Data for Predominant Instrument Recognition in Polyphonic Music

APSIPA ASC 2023

Lifan Zhong [1], Erica Cooper [2], Junichi Yamagishi [2], Nobuaki Minematsu [1]

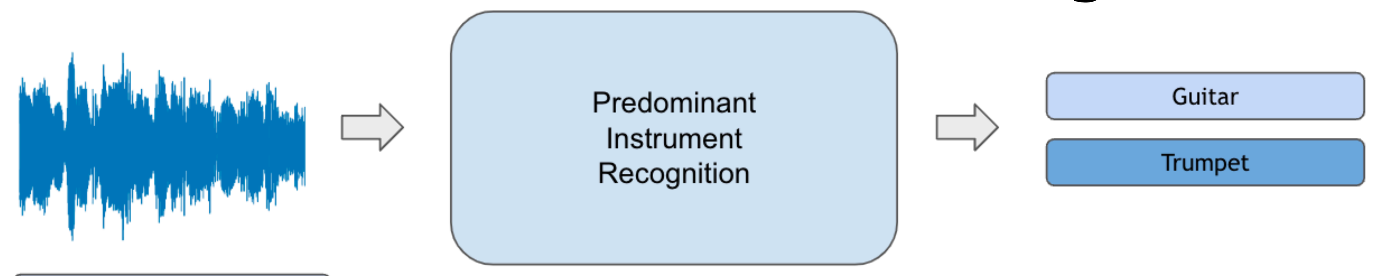
[1] Graduate School of Engineering, The University of Tokyo, Japan

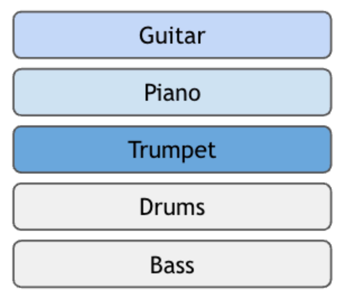
[2] National Institute of Informatics, Japan

Introduction

- **Automatic instrument recognition** has various applications in music recommendation, music transcription, etc.
- We propose a **robust end-to-end instrument recognition system for polyphonic multi-instrument music**, using isolated musical notes as pre-training data.

Predominant Instrument Recognition

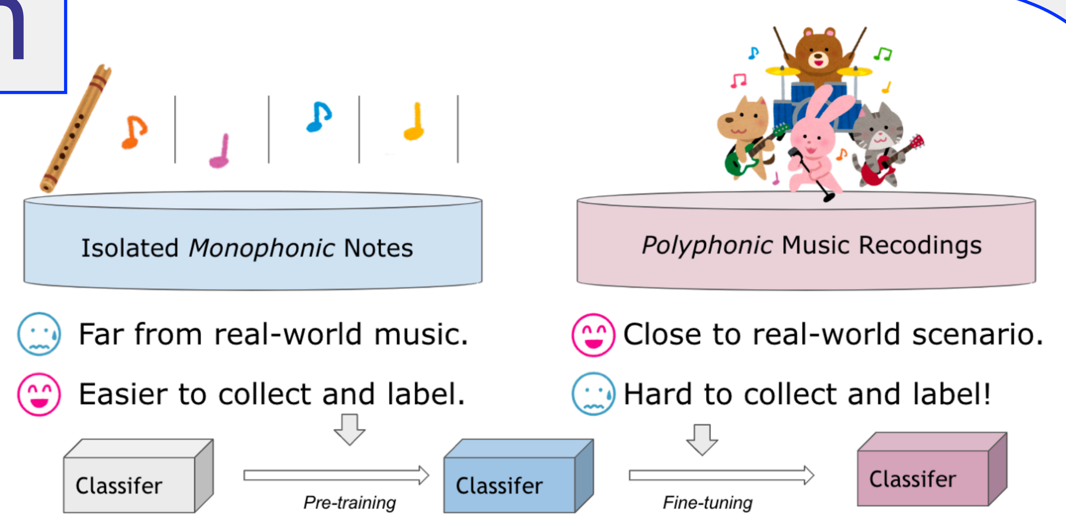




- Automatic recognition of the predominant or lead instrument(s)
- Input music signals can be **polyphonic** and **multi-instrumental**

Motivation

A lack of well-annotated polyphonic musical data has been a constraint, because:

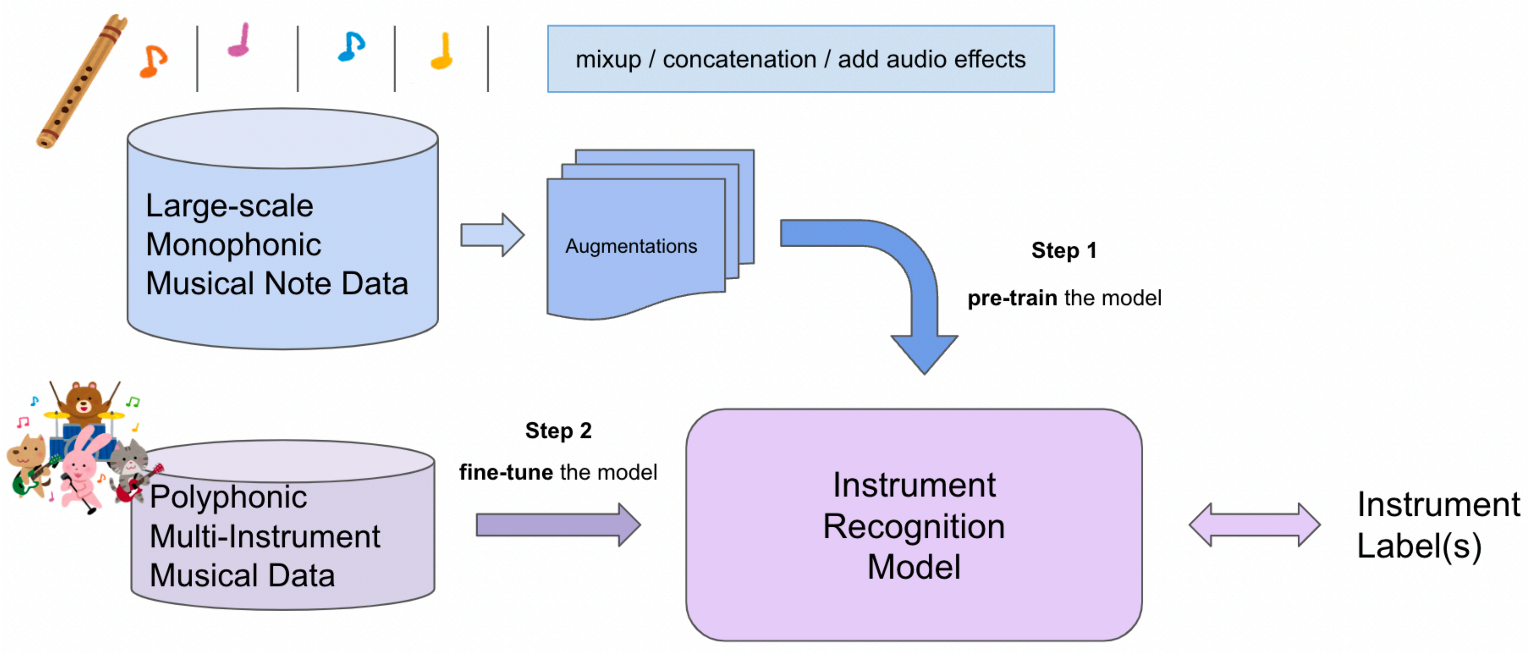


- Domain knowledge is necessary for annotation
- Well-produced music recordings have copyright issues.

Monophonic sounds and **isolated notes** require relatively less effort to collect and label.

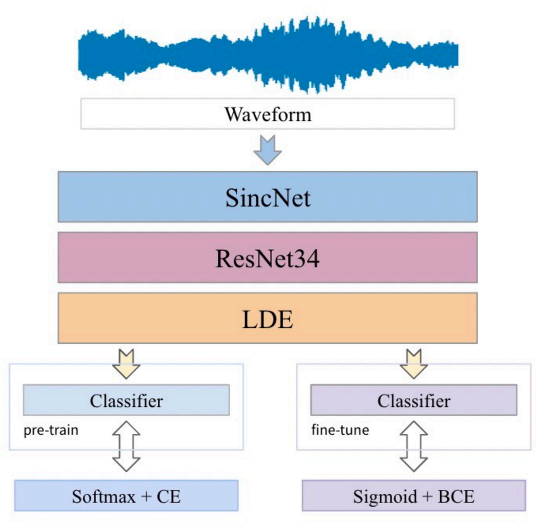
=> Can we use isolated monophonic notes as pre-training data?

Methodology



Overview of our proposed method

- **Augment** the monophonic musical note data by mix-up [Zhang+, 2017] [Tokozume+, 2017], concatenating, and adding effects, to alleviate the domain gap
- **Pre-train** the model with augmented monophonic musical note data.
- **Fine-tune** the pre-trained model using polyphonic, multi-instrument musical recordings



Instrument Recognition Model Architecture

Dataset

Pre-training: NSynth [J. Engel+, 2017]

- Samples of instruments sustaining a note for 3s and letting it decay for 1s

TABLE I
SUMMARY OF THE NSYNTH DATASET AND THE IRMAS DATASET

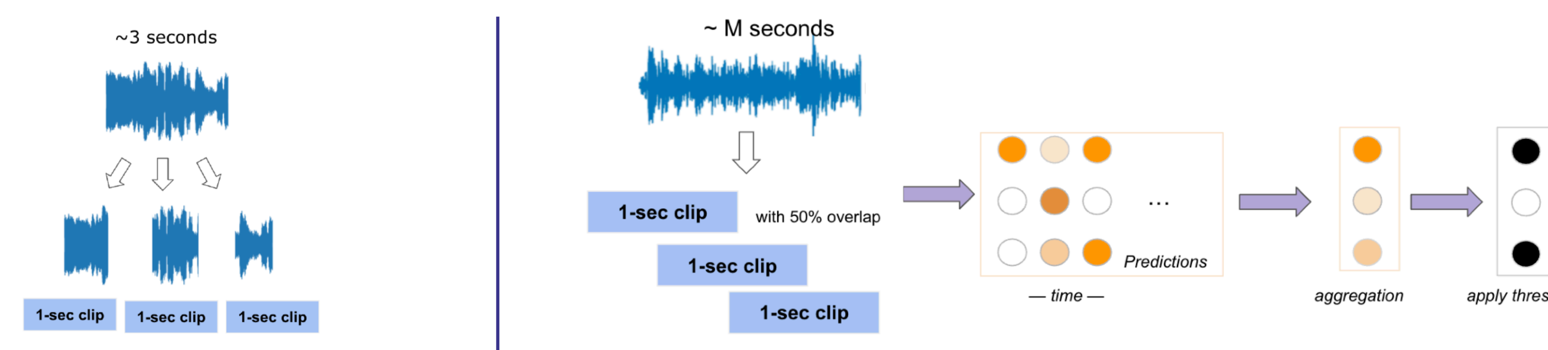
Dataset	NSynth	IRMAS - train	IRMAS - test
# Instruments	1,006	11	11
# Samples	305,979	6,705	2,874
Duration per sample	4 seconds	3 seconds	5 - 20 seconds
Total duration	340.0 hours	5.6 hours	13.5 hours

Fine-tuning: IRMAS [Bosch+, 2012]

- Professionally produced western music recordings of various genres, with excerpt-wise predominant instrument labels of 11 classes: cello (cel), clarinet (cla), flute (flu), acoustic guitar (gac), electric guitar (gel), organ (org), piano (pia), saxophone (sax), trumpet (tru), violin (vio), and human singing voice (voi)

Experimental Settings

https://github.com/nii-yamagishilab/predominant-instrument-recognition



Training

- Divide input audio into 1-second clips

Testing

- Average the clip-wise predictions to get the segment-wise predictions

Evaluation Metrics

1. F1-score

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\mathrm{P}_{macro} = \frac{1}{L}\sum_{l=1}^{L}\frac{\mathrm{TP}_l}{\mathrm{TP}_l + \mathrm{FP}_l}, \quad \mathrm{P}_{micro} = \frac{\sum_{l=1}^{L}\mathrm{TP}_l}{\sum_{l=1}^{L}(\mathrm{TP}_l + \mathrm{FP}_l)},$$

$$\mathrm{R}_{macro} = \frac{1}{L}\sum_{l=1}^{L}\frac{\mathrm{TP}_l}{\mathrm{TP}_l + \mathrm{FN}_l}, \quad \mathrm{R}_{micro} = \frac{\sum_{l=1}^{L}\mathrm{TP}_l}{\sum_{l=1}^{L}(\mathrm{TP}_l + \mathrm{FN}_l)},$$

2. LRAP (label ranking average precision)

Ground Truth $y \in \{0,1\}^{n_{\text{samples}} \times n_{\text{labels}}}$

Predictions $\hat{f} \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{labels}}}$

Need no threshold!

$$LRAP(y, \hat{f}) = \frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}\frac{1}{||y_i||_0}\sum_{j:y_{ij}=1}\frac{|\mathcal{L}_{ij}|}{\text{rank}_{ij}}$$

where $\mathcal{L}_{ij} = \{k : y_{ik} = 1, \hat{f}_{ik} \geq \hat{f}_{ij}\}$ and $\text{rank}_{ij} = |\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\}|$. $|\cdot|$ computes number of elements of the set and $||\cdot||_0$ computes the number of nonzero elements in a vector.

Experimental Results

We report the results on the *IRMAS testing data*.

- NSynth pre-training strongly **improves** performance

TABLE II
TRAINING WITH RANDOM INITIALIZATION *vs.* WITH NSYNTH PRE-TRAINING

Initialization	F1-micro	F1-macro	LRAP
Random	0.634 ± 0.0075	0.536 ± 0.0127	0.780 ± 0.0057
NSynth	0.674 ± 0.0068	0.584 ± 0.0068	0.814 ± 0.0020

- All augmentation techniques help, and **mixing two samples with soft labels** has the most impact

TABLE III
ABLATIONS OF PRE-TRAINING AUGMENTATION METHODS

Augmentations	F1-micro	F1-macro	LRAP
All	0.674 ± 0.0068	0.584 ± 0.0068	0.814 ± 0.0020
- mixup	0.657 ± 0.0029	0.560 ± 0.0045	0.804 ± 0.0040
- audio effect	0.671 ± 0.0031	0.576 ± 0.0055	0.812 ± 0.0030
- both[a]	0.642 ± 0.0050	0.535 ± 0.0031	0.791 ± 0.0037
- concatenation	0.670 ± 0.0012	0.576 ± 0.0015	0.813 ± 0.0013

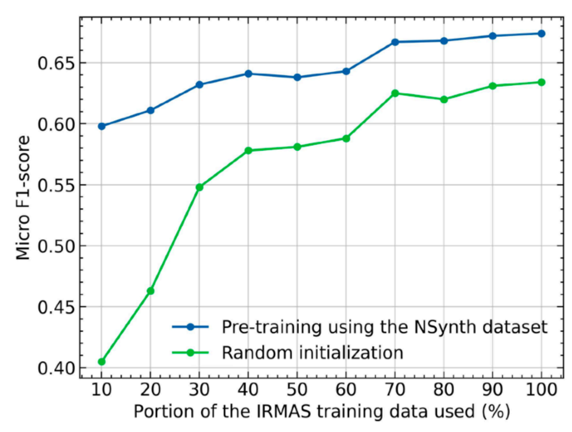
[a] Without mixup and audio effects

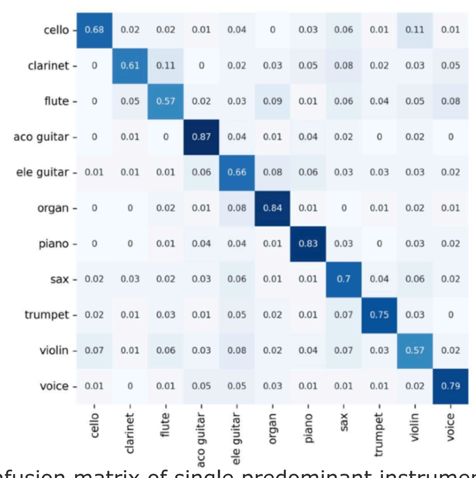
- Outperforms previous end-to-end system by 0.066 in micro F1-score (**10.9%** relative improvement)
- Better performance than most previous methods that use time-frequency representations as inputs, except for [19], whose model has 25.5M parameters, while our model has **1.3M**

TABLE IV
COMPARISON OF EVALUATION RESULTS ON THE IRMAS TESTING DATA

Methods	Features	F1-micro	F1-macro	LRAP
This work	Waveform	0.674	0.584	0.814
Avramidis *et al.* [18]	Waveform	0.608	0.543	0.747
Kratimenos *et al.* [4]	CQT	0.647	0.546	0.805
Zhong *et al.* [19][a]	Mel	0.680	0.600	0.818
Reghunath & Rajan [17]	Mel[b]	0.66	0.62	-
Yu *et al.* [16]	Mel	0.661	0.569	-
Pons *et al.* [15]	Mel	0.589	0.516	-
Han *et al.* [14][c]	Mel	0.619	0.513	-

- NSynth pre-training helps regardless of the volume of fine-tuning data. However, with pre-trained weights and 10% of IRMAS training data, we can train a reasonable model.





Confusion matrix of single predominant instrument identification. The columns are predictions and the rows are ground truth labels.

Conclusion

- A pre-training and fine-tuning approach using monophonic isolated musical note data proves effective in predominant instrument recognition.
- Data augmentation techniques during pre-training contributes to the robustness of our model.
- Our best model achieves a micro F1-score of 0.674 and an LRAP of 0.814, marking a significant improvement of 10.9% and 8.9% relative to the previous end-to-end approach.