Speaker Recognition Adapted for Musical Instruments

Xuan Shi[1], Erica Cooper[2] and Junichi Yamagishi[2]
[1]University of Southern California [2]National Institute of Informatics, Japan

Introduction

Given the similarities between speaker recognition and musical instrument recognition, we adapt speaker recognition algorithms to the task of learning meaningful instrumental timbre representations.

- Introduced a group of trainable filters initialized with Mel and **MIDI filter bank** to address the mismatch between speech and musical instrument sound.
- The modified speaker recognition model was capable of generating discriminative embeddings for instrument and instrument-family, performing well in both **closed-set** and **open-set** scenarios.
- Conducted extensive experiments to characterize the encoded information in learned timbre embeddings.

References

[1] M. Ravanelli et al. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop, SLT*. IEEE, 2018.

[2] K. He et al. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2016.

[3] W. Cai et al. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In *Odyssey 2018: The Speaker and Language Recognition Workshop*. ISCA, 2018.

[4] Z. Huang et al. Angular softmax for short-duration text-independent speaker verification. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*. ISCA, 2018.

[5] N. Zeghidour et al. LEAF: A learnable frontend for audio classification. In *9th International Conference on Learning Representations, ICLR 2021*,. OpenReview.net, 2021.

[6] A. Ramires et al. Data augmentation for instrument classification robust to audio effects. 2019.

[7] D. Raj et al. Probing the information encoded in x-vectors. In *IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE, 2019.

[8] J. Engel et al. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

[9] M. Goto et al. RWC music database: Music genre database and musical instrument sound database. In *ISMIR 2003, 4th International Conference on Music Information Retrieval, Proceedings*, 2003.

Methods

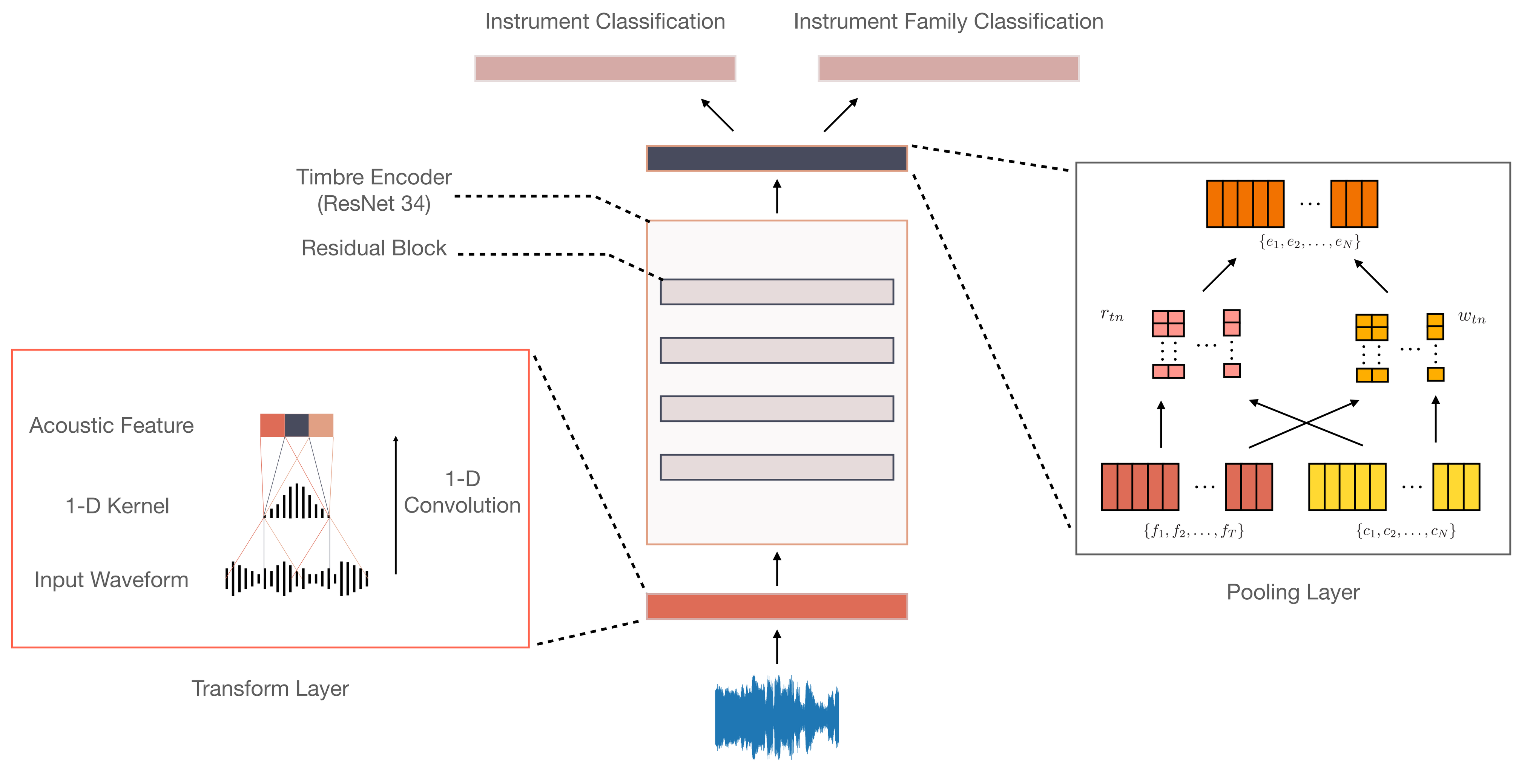

Figure 1: Architecture of proposed musical instrument recognition model inspired by speaker recognition

- Transform Layer based on SincNet[1]
- Encoder based on ResNet [2] and LDE [3]
- Dual outputs based on Angular-Softmax [4]
- MIDI filter bank initialization

Result I: Recognition

- Two Recognition Scenarios: instrument verification, instrument-family identification.
- Database: NSynth Dataset [8] (individual notes from 1,006 instruments)
- Training Strategy: data augmentation, Angular-Softmax.

Table 1: Instrument verification and instrument-family identification results on NSynth database.

Systems	EER	Micro F1
Melspec-aug-asm	**3.186**	77.00
wav-transMel-aug-asm	3.424	77.34
wav-transMIDI-aug-asm	3.737	**77.76**
LEAF [5]		72.0
Baseline in [6]		73.78
Best in [6]		74.73

Result II: Generalization

- Task: generalize the model trained on NSynth to RWC dataset [9] (45 categories).
- Training Strategy: training from the scratch, training based on pre-trained model.
- Results: pre-trained parameters from NSynth help the model to converge faster and achieve higher accuracy on RWC.

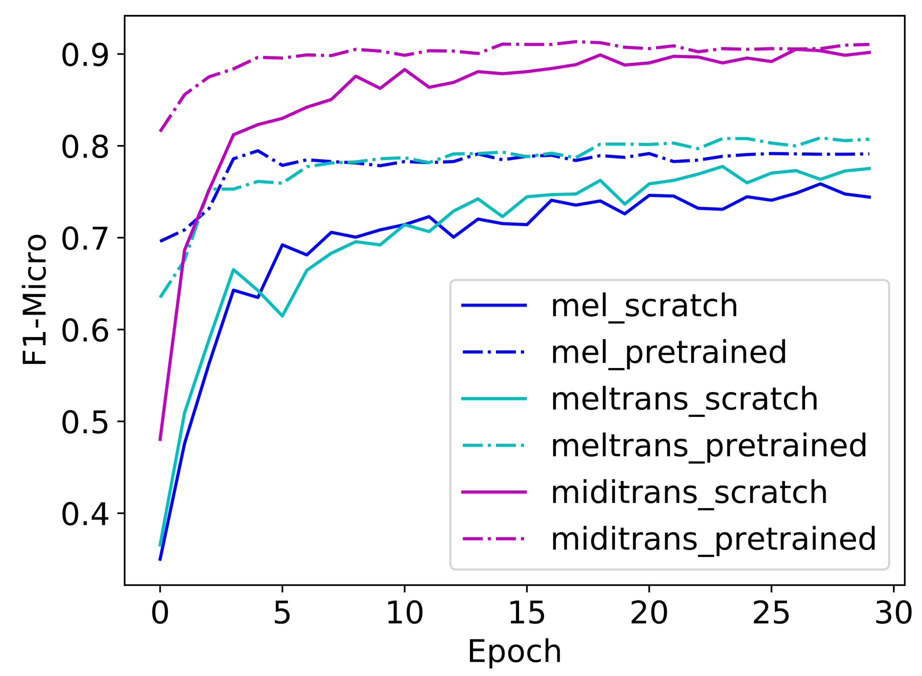

Figure 2: F1-Scores on RWC dataset. Solid lines indicate training from scratch, and dashed lines indicate fine-tuning from pre-trained model.

Result III: Probing the Encoded Information

- Task: probing the encoded information in the timbre embeddings obtained from the proposed model in a similar way to [7].
- Models: a series of shallow classifiers.

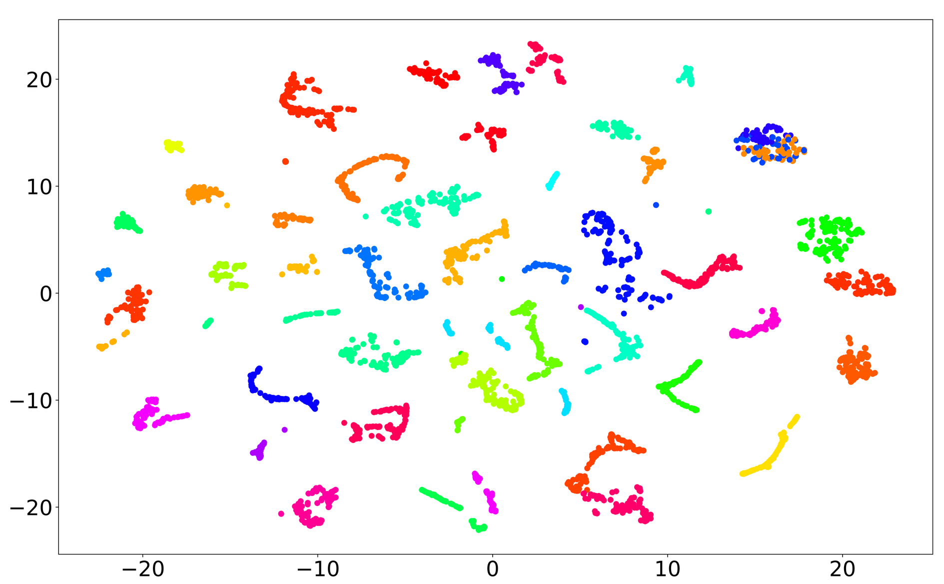

Figure 3: T-SNE visualization of embeddings extracted from wav-transMel-aug-asm.

- Results: some meta information is encoded in embeddings, such as pitch, source.

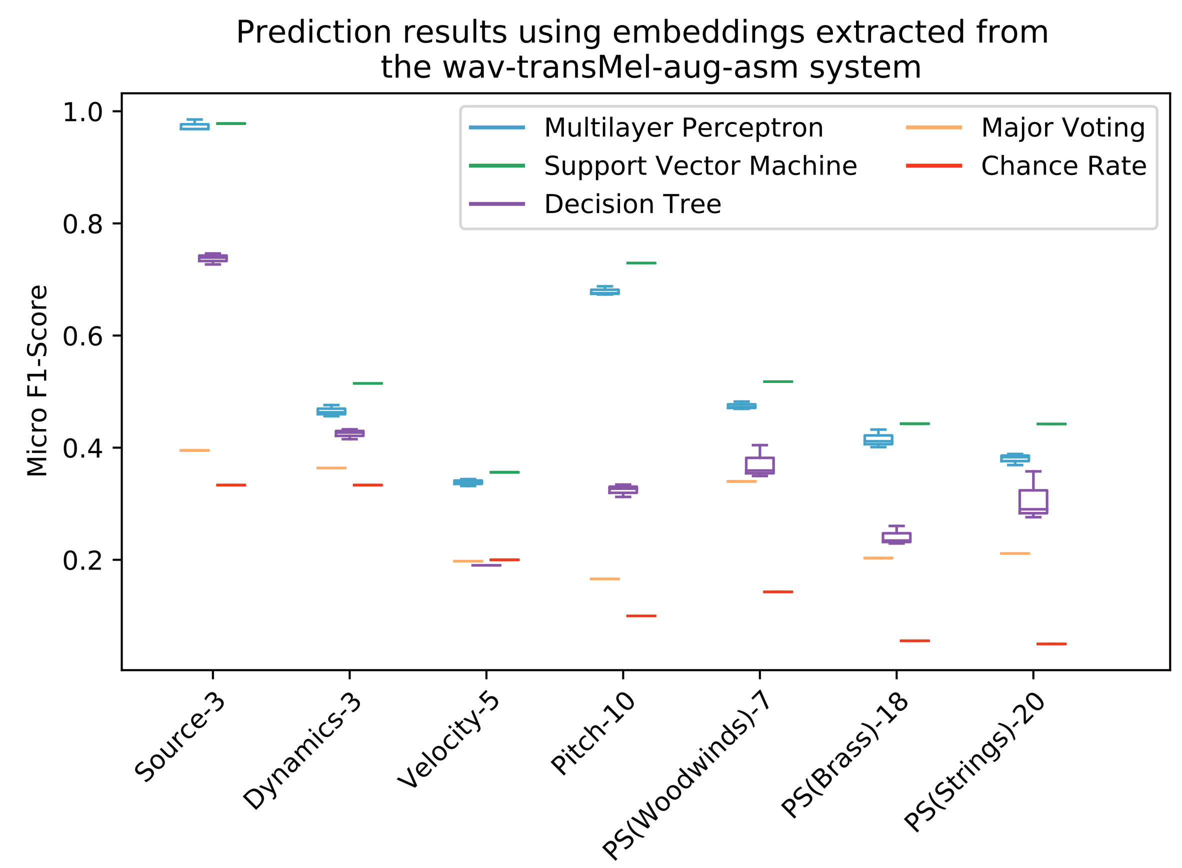

Figure 4: Prediction results.

Future Research

- Construct the instrument timbre space for polyphonic musical instrument sound input
- Apply the timbre representation in multi-instrument sound synthesis