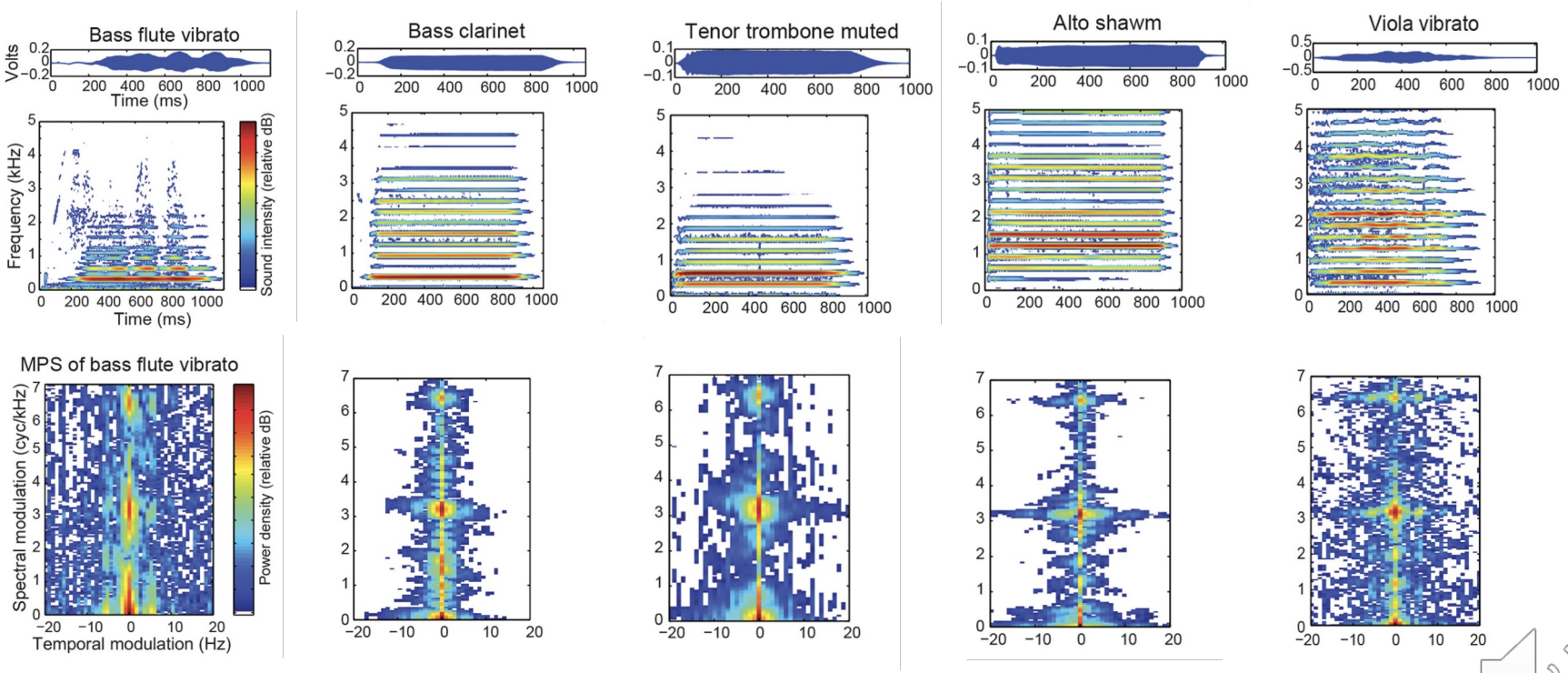
# Use of Speaker Recognition Approaches for Learning and Evaluating Embedding Representations of Musical Instrument Sounds

Xuan Shi[1], Erica Cooper[2], Junichi Yamagishi[2]

[1] University of Southern California
[2] National Institute of Informatics (NII), Japan

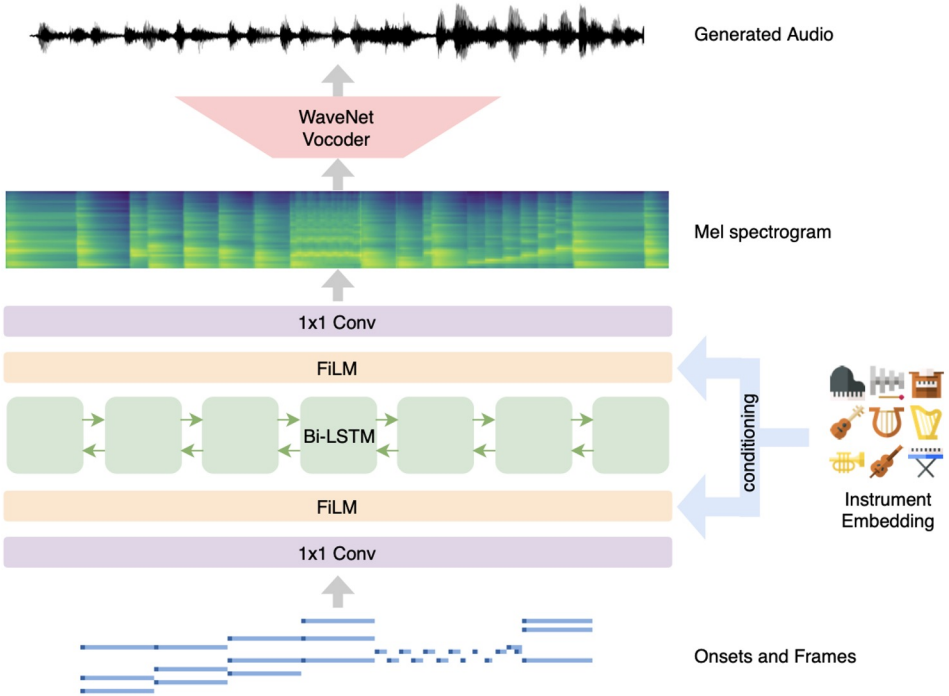
IEEE/ACM Transactions on Audio, Speech and Language Processing
Volume 30, pp 367–377, https://doi.org/10.1109/TASLP.2022.3140549

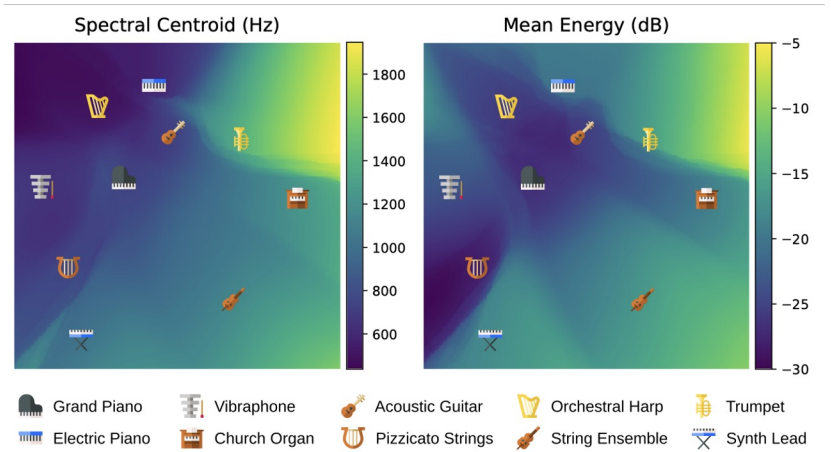
# Motivation – Musical Instrument Audio Perception




[1] Elliott, Taffeta M., Liberty S. Hamilton, and Frédéric E. Theunissen. "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones." *The Journal of the Acoustical Society of America* 133.1 (2013): 389-404.

# Motivation – Musical Instrument Audio Synthesis




[2] Kim, Jong Wook, et al. "Neural music synthesis for flexible timbre control." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

# Questions

- Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?
- Q2: If we answer yes to Q1, considering the **differences** between speech and music, how can we narrow the gap between them?
- Q3: In the embeddings extracted from instrument recognition, do they only encode timbre information?

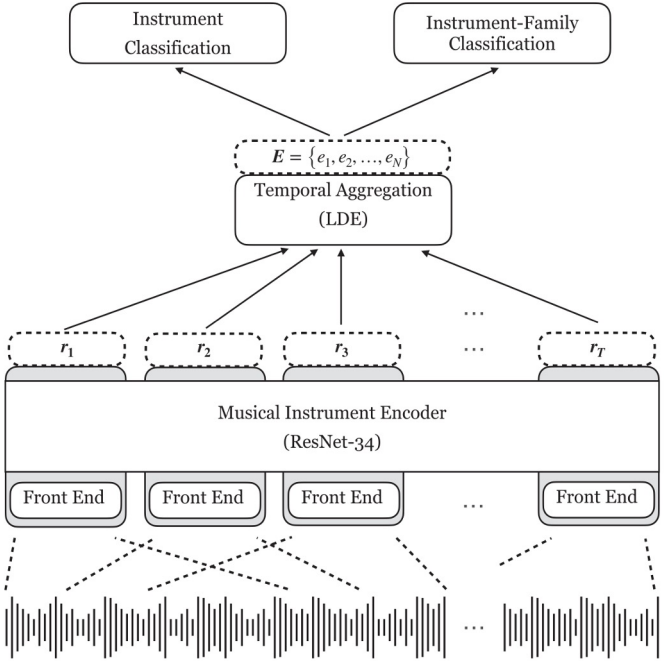
# Questions

- Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?
- Q2: If we answer yes to Q1, considering the **differences** between speech and music, how can we narrow the gap between them?
- Q3: In the embeddings extracted from instrument recognition, do they only encode timbre information?

# Methods

Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?

- Architecture



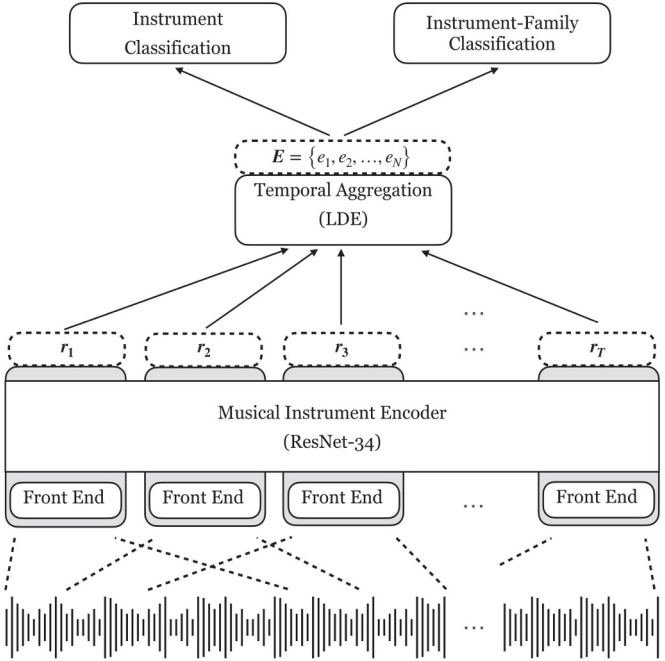

Borrow Ideas from Auto Speaker Verification (ASV)

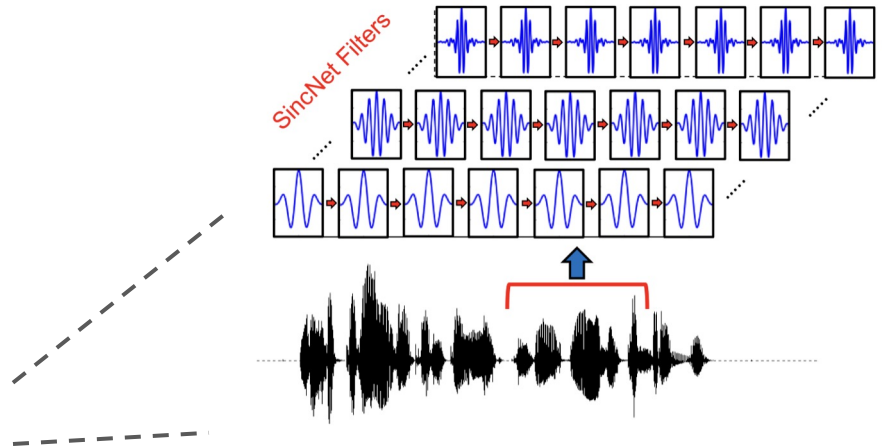
- Front End: SincNet
- Feature Transformation: ResNet34
- Feature Aggregation: Learnable Dictionary Encoding (LDE)
- Classification: Angular Softmax

# Methods

Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?

- SincNet



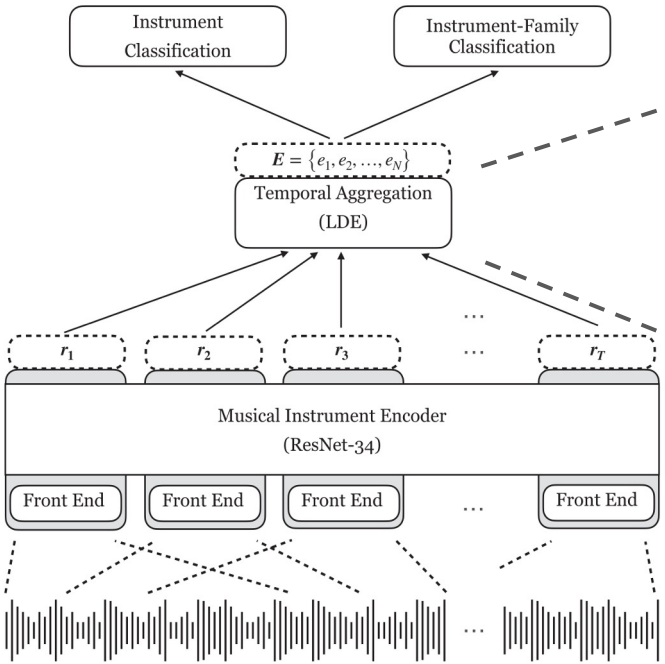



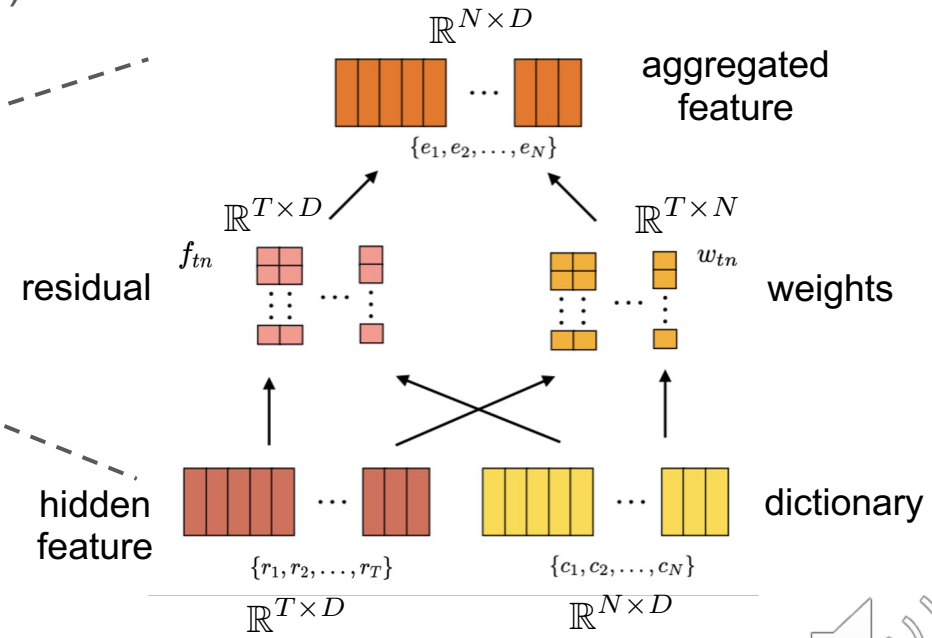

[4] Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.

# Methods

Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?

- Learnable Dictionary Encoding (LDE)



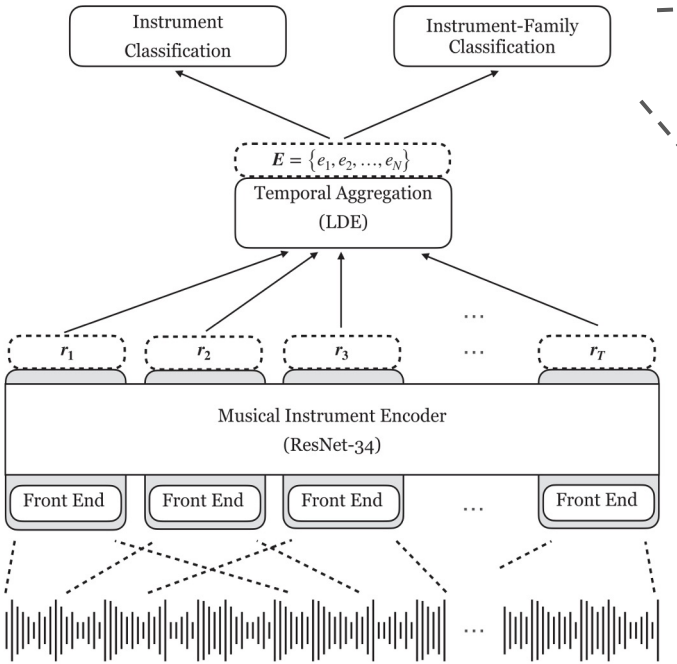



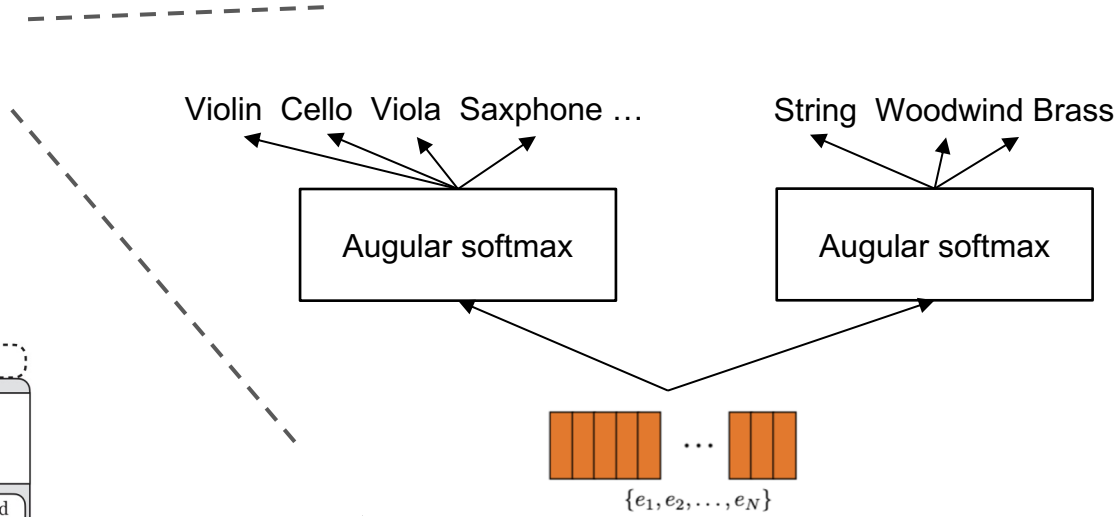

[5] Cai, Weicheng, et al. "A novel learnable dictionary encoding layer for end-to-end language identification." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

# Methods

Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?

- Classification: Angular Softmax



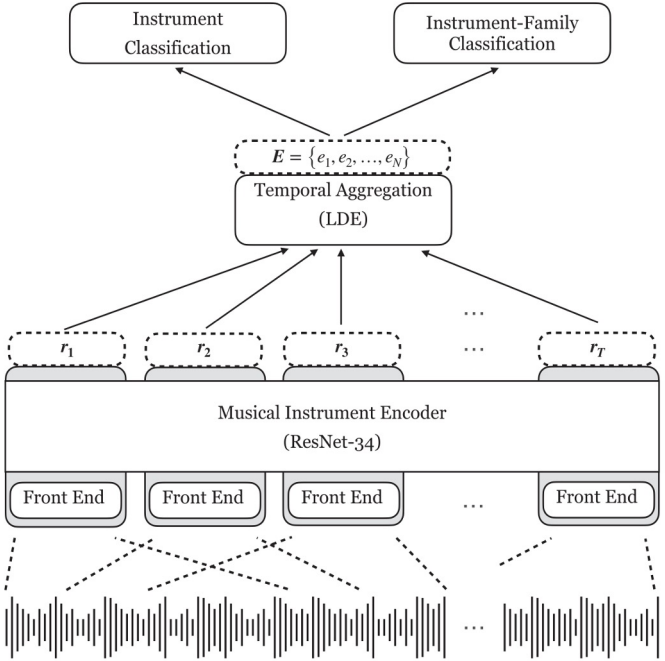




[6] Li, Yutian, et al. "Angular softmax loss for end-to-end speaker verificatio 018 11th International Symposium on Chinese Spoken Language Processing (ISC LP). IEEE, 2018.

# Methods

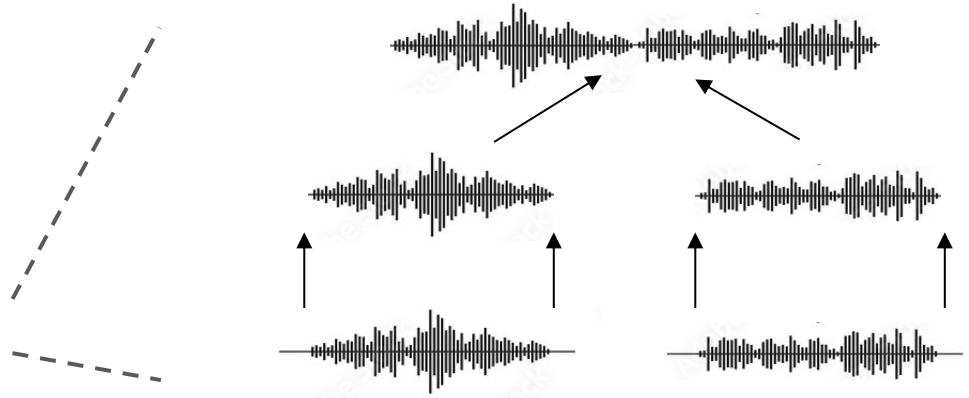
Q1: Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?

- Training with data augmentation




Data augmentation by concatenating notes

- Remove silent segments
- Concatenate notes from the same instrument



# Experiments

- Datasets

| | NSynth | RWC instrument |
|---|---|---|
| Categories in database | 1,006 | 45 |
| Categories in training set | 953 | 45 |
| Categories in valid. set | 53 | 30 |
| Categories in test set | 53 | 30 |
| Samples in database | 305,979 | 1556 |
| Samples in training set | 289,205 | 1144 |
| Samples in valid. set | 12,678 | 209 |
| Samples in test set | 4,096 | 203 |
| Total playback time (with silence) | 340.0 hours | 71.7 hours |

[6] Engel, Jesse, et al. "Neural audio synthesis of musical notes with WaveNet autoencoders." *ICML*. PMLR, 2017.
[7] Goto, Masataka, et al. "RWC Music Database: Popular, Classical and Jazz Music Databases." *ISMIR*. Vol. 2. 2002.

# Experiments

- Metric: equal error rate (EER)

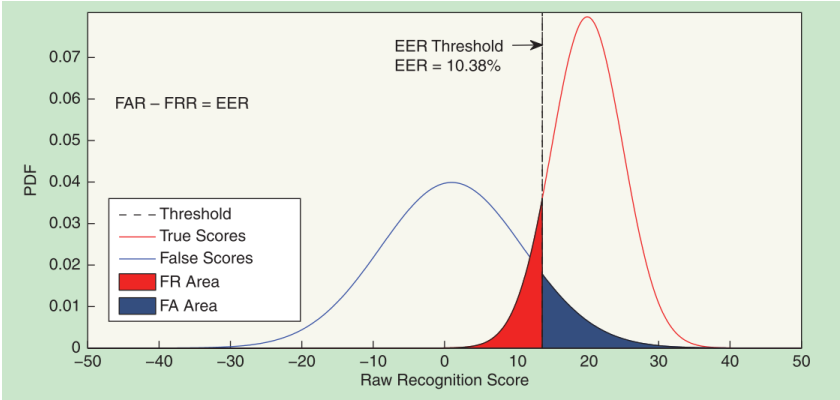



Fig.10 of [7]

- Results of ablation study

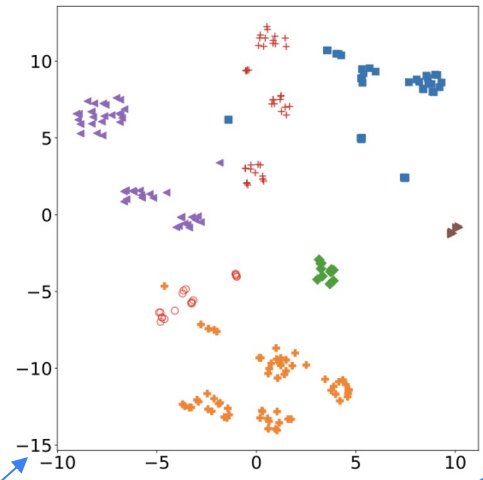
| System | NSynth EER (%) ↓ | RWC EER (%) ↓ | Average EER (%) ↓ |
|---|---|---|---|
| Base | 3.74 ± 0.37 | 1.03 ± 0.11 | 2.39 |
| Base w/o data augmentation | 7.58 ± 2.34 | 2.22 ± 0.29 | 4.90 |
| Base w/o instrument family | 3.14 ± 0.34 | 1.12 ± 0.20 | 2.13 |
| Base w/o A-softmax | 4.60 ± 0.23 | 2.44 ± 0.08 | 3.52 |

[8] John H L Hansen, et.al. “Speaker Recognition by Machines and Humans: A Tutorial Review”. *IEEE Signal Processing Magazine* 2015.

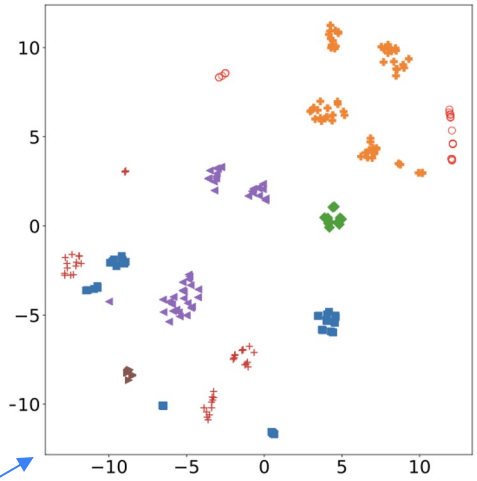
# Experiments

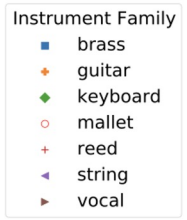
- Results of ablation study

T-SNE visualization of the Base system




T-SNE visualization of the Base w/o instrument family system




| System | NSynth EER (%) ↓ | RWC EER (%) ↓ | Average EER (%) ↓ |
|---|---|---|---|
| Base | 3.74 ± 0.37 | 1.03 ± 0.11 | 2.39 |
| Base w/o data augmentation | 7.58 ± 2.34 | 2.22 ± 0.29 | 4.90 |
| Base w/o instrument family | 3.14 ± 0.34 | 1.12 ± 0.20 | 2.13 |
| Base w/o A-softmax | 4.60 ± 0.23 | 2.44 ± 0.08 | 3.52 |

Musical Instrument Recog

/Users/shixuan/REPO/musical_instrument_embeddings_visualization/index.html

# Musical Instrument Recognition Model [Paper]

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

## musical instrument embeddings visualization

Color scheme:

Instrument Type | Pitch

Dataset:

Instrument:

Click a dot to play the soundfont sample:

Toggle rotation

This page is modified from the Template developed by Music and Audio Research Laboratory (MARL) from New York University (NYU).

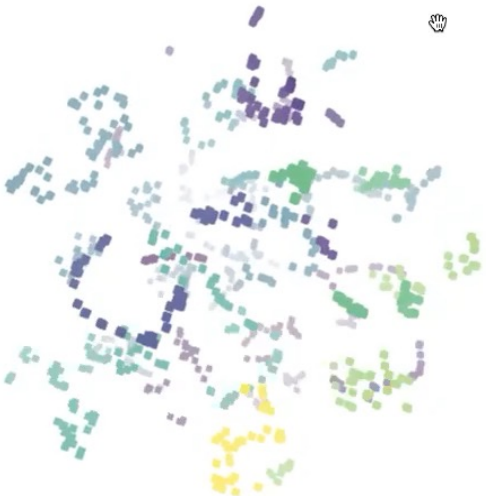
Musical Instrument Recognitio

/Users/shixuan/REPO/musical_instrument_embeddings_visualization/index.html

# Musical Instrument Recognition Model [Paper]

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

## musical instrument embeddings visualization



Color scheme:

Instrument Type | Pitch

Dataset:

Instrument:

Click a dot to play the soundfont sample:

Toggle rotation

This page is modified from the Template developed by Music and Audio Research Laboratory (MARL) from New York University (NYU).

Musical Instrument Recognitio

/Users/shixuan/REPO/musical_instrument_embeddings_visualization/index.html

# Musical Instrument Recognition Model [Paper]

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

## musical instrument embeddings visualization

Color scheme:

Instrument Type | Pitch

Dataset:

Instrument:

Click a dot to play the soundfont sample:

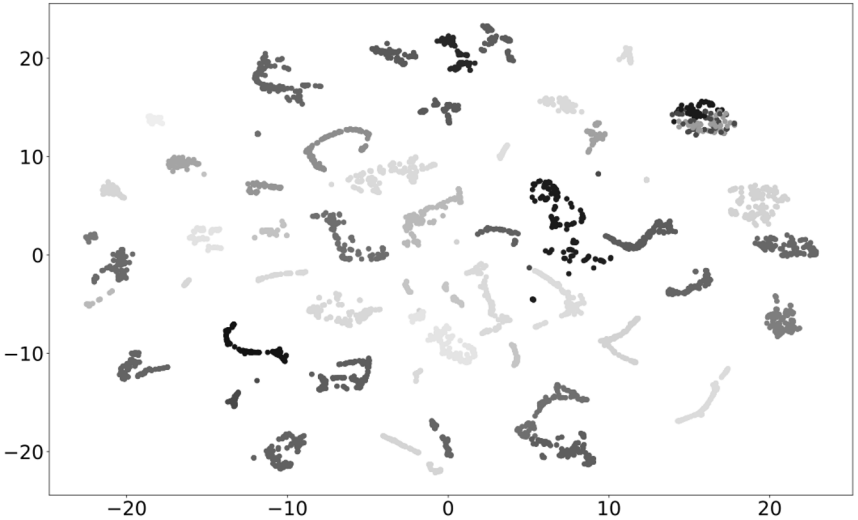
Toggle rotation

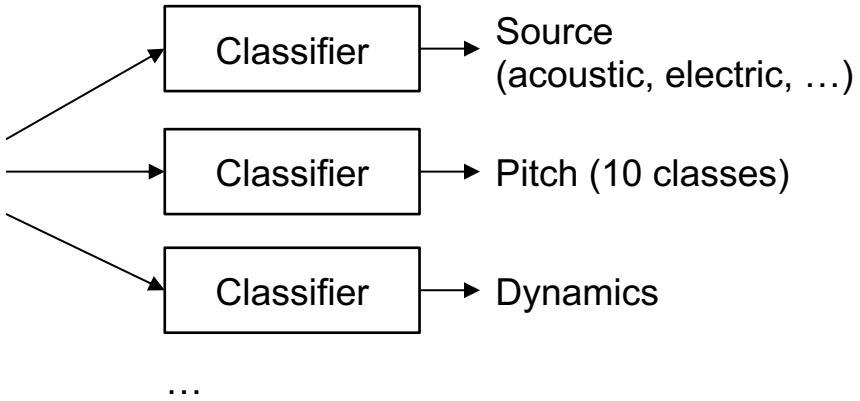
This page is modified from the Template developed by Music and Audio Research Laboratory (MARL) from New York University (NYU).

# Experiments

Q3: In the embeddings extracted from instrument recognition, do they only encode timbre information?

- Probing Logits




Classifier:

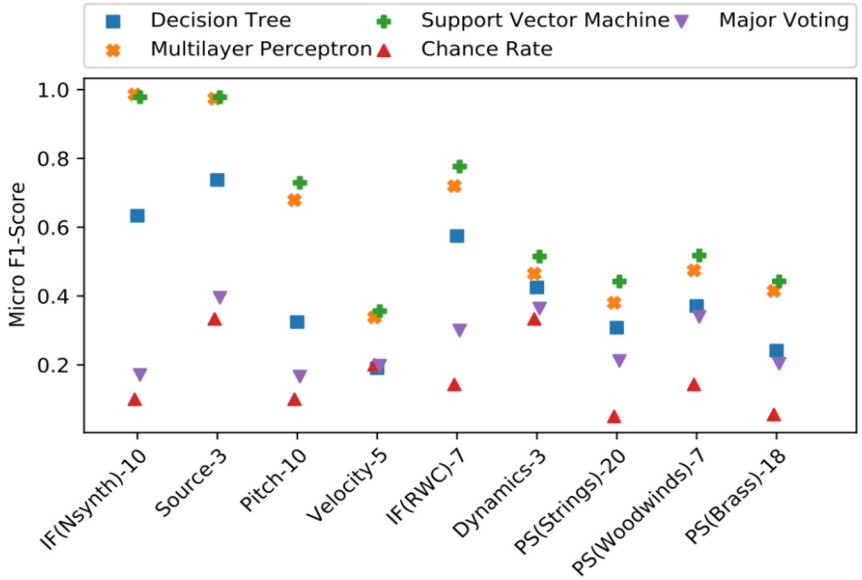
- Machine learning: Decision Tree, Support Vector Machine, Multilayer Perceptron
- Baselines: Majority Voting, Chance Rate

# Experiments

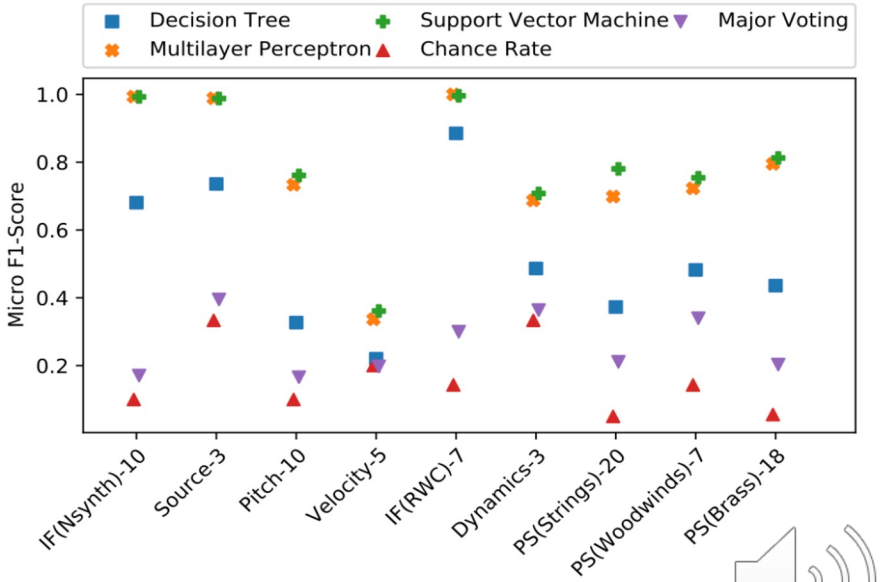
Q3: In the embeddings extracted from instrument recognition, do they only encode timbre information?

- Probing Logits

Embedding vectors contain information of instrument family




(a) Initialization using the Mel filterbank



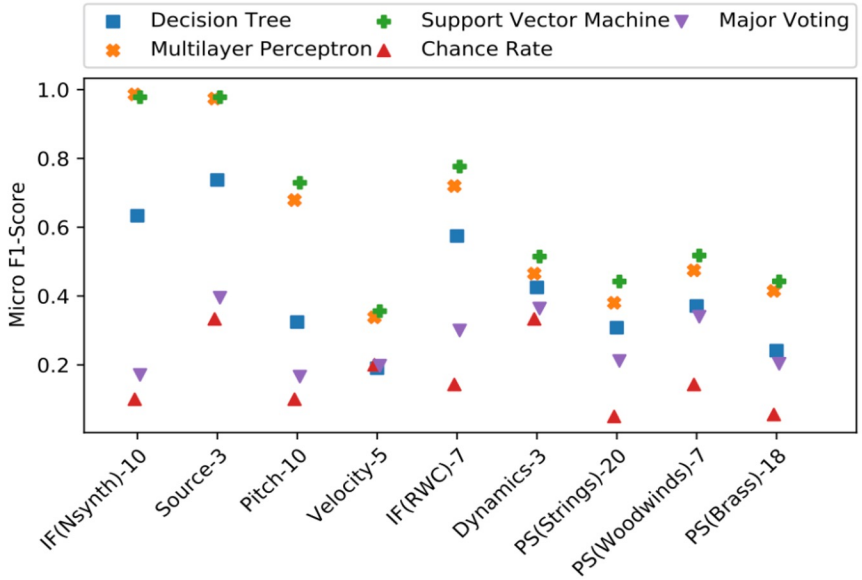

(b) Initialization using the CQT filtebank

# Experiments

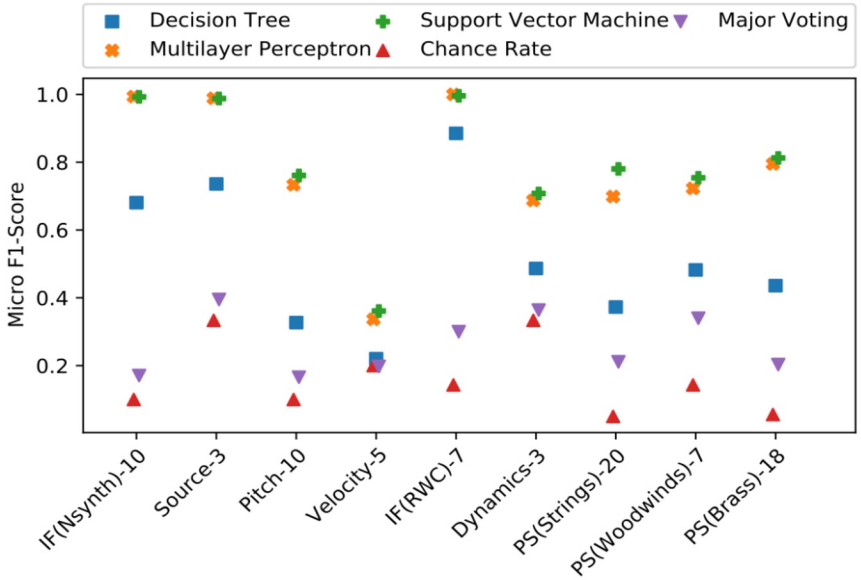
Q3: In the embeddings extracted from instrument recognition, do they only encode timbre information?

- Probing Logits

Embedding vectors contain source and pitch information




(a) Initialization using the Mel filterbank



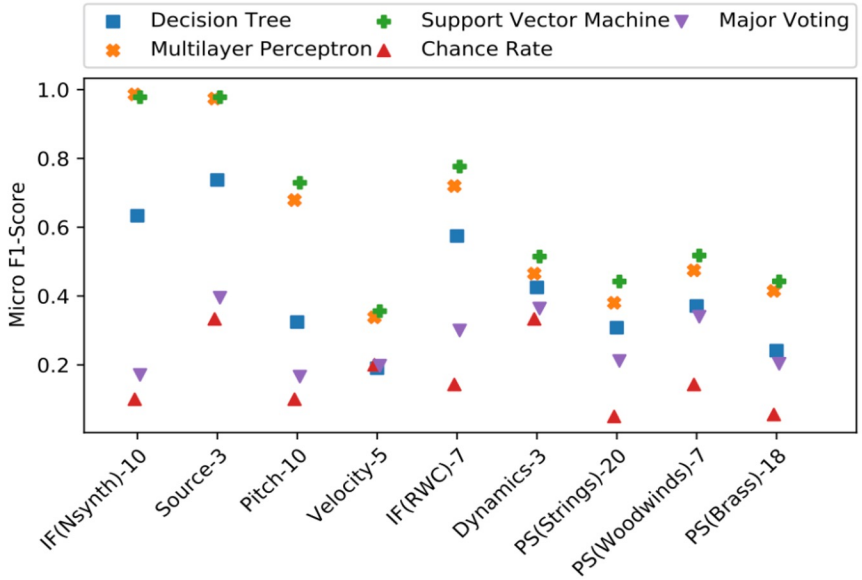

(b) Initialization using the CQT filtebank

# Experiments

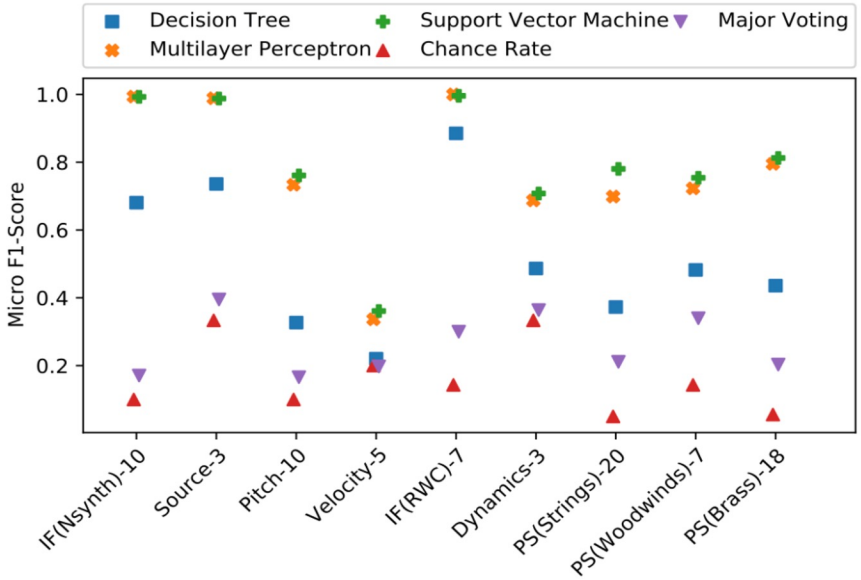
Q3: In the embeddings extracted from instrument recognition, do they only encode timbre information?

- Probing Logits

Embedding vectors contain less information on velocity




(a) Initialization using the Mel filterbank




(b) Initialization using the CQT filtebank

# Conclusion

- Is it possible to use End-to-End **speaker recognition** methods to train End-to-End **musical instrument recognition**?
  - **Yes!** ASV techniques (SincNet, LDE, …) -> construct meaningful embeddings for unseen instruments
- In the embeddings extracted from instrument recognition, do they only encode timbre information?
  - **No!** Meta information is encoded with different scale, such as source, pitch, and playing style.

USC University of Southern California

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

# Thank you

Xuan Shi, Erica Cooper, Junichi Yamagishi

Use of Speaker Recognition Approaches for Learning and Evaluating Embedding Representations of Musical Instrument Sounds

IEEE/ACM Transactions on Audio, Speech and Language Processing
Volume 30, pp 367–377, https://doi.org/10.1109/TASLP.2022.3140549

Code: https://github.com/Alexuan/musical_instrument_embedding

MIDI2Audio Project Link: https://nii-yamagishilab.github.io/sample-midi-to-audio/