Visualization and Interpretation of Siamese Style Convolutional Neural Networks for Sound Search by Vocal Imitation

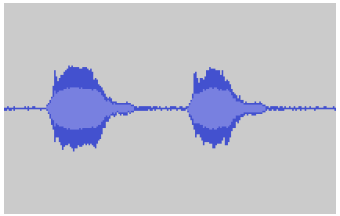
Yichi Zhang and Zhiyao Duan

Audio Information Research (AIR) Lab
Department of Electrical and Computer Engineering
University of Rochester

Session: MLSP-L8: Deep Learning III

Query by Vocal Imitation

Vocal imitation of this sound



How to find an audio file from sound effect libraries?

Sound recording of “Metaloid”

Vocal Imitations in Daily Life

Bad Turbo: boooOOOOOOooo

Boiling Coolant: bllgh blllgggh bllllllgggghh

Clutch Screech: screek, screek, screek

Engine Knock: tuckaTHUCKtuckaTHUCKtucka

Vocal Imitation Challenges

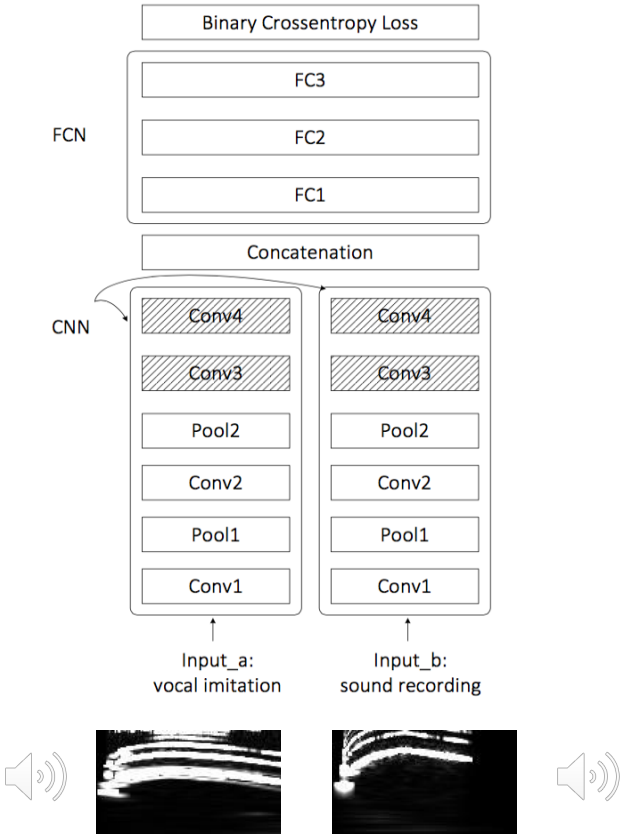
Question 1: What feature representations and similarity measures are effective? How to learn them together?

Solution: Siamese-style Convolutional Neural Network (SCNN)

Question 2: How does SCNN work? What feature representations are learned at different layers of the network?

Solution: Visualizing input patterns that activate a certain neuron the most

Previous IMINET Model

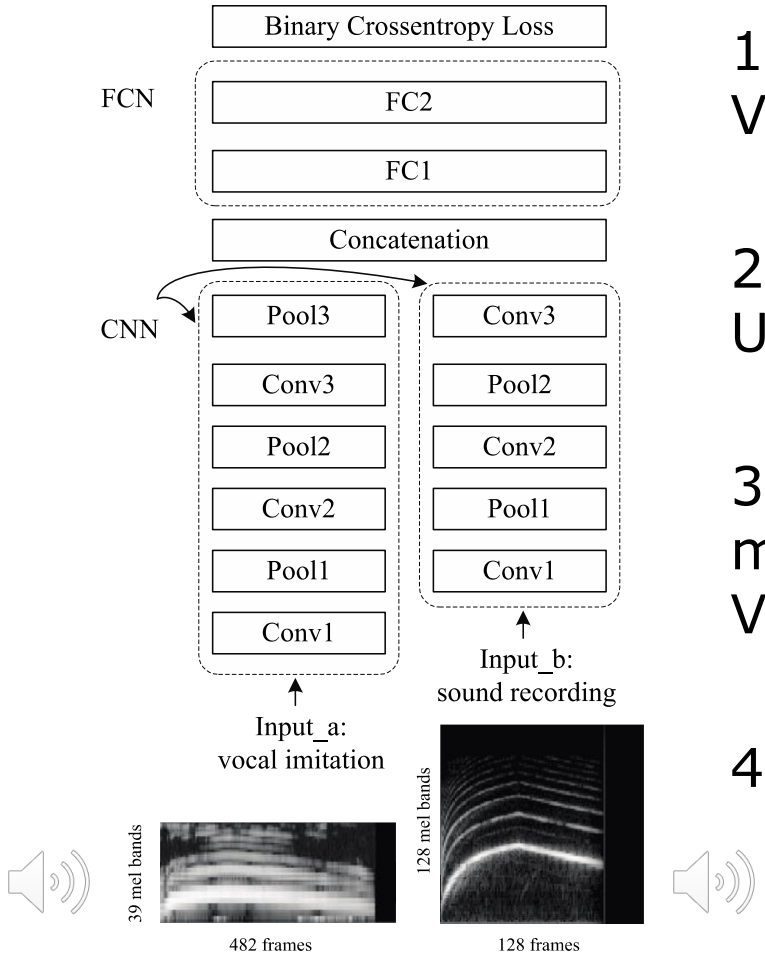


1. Pre-processing: Constant-Q Transform spectrograms
2. Feature Extraction: Convolutional Neural Networks
3. Metric Learning: Fully Connected Networks
4. Sound Retrieval: Ranking output probabilities

Note: The work of this ICASSP paper is derived from our previous IMINET model in [1]

[1] Y. Zhang and Z. Duan, IMINET: Convolutional Semi-Siamese Networks for Sound Search by Vocal Imitation, WASPAA 2017

Proposed TL-IMINET Model



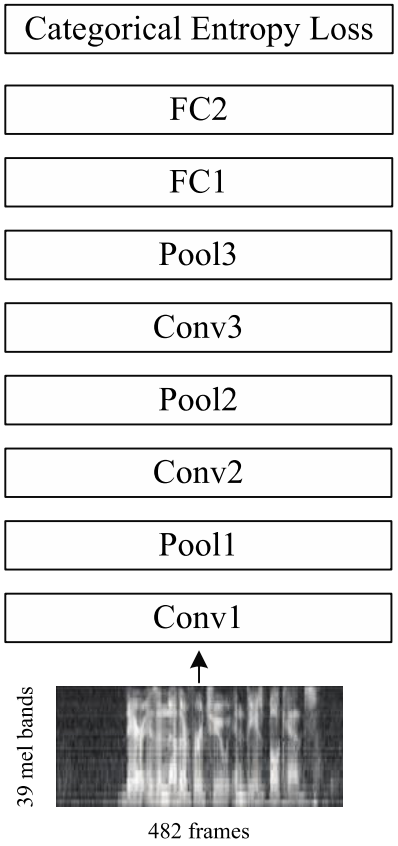
1. Pre-train imitation tower using VoxForge data set

2. Pre-train recording tower using UrbanSound8K

3. Fine-tune two towers with metric learning module using VocalSketch Data Set

4. Sound retrieval

Imitation Tower Pre-training



Dataset: VoxForge spoken language classification dataset

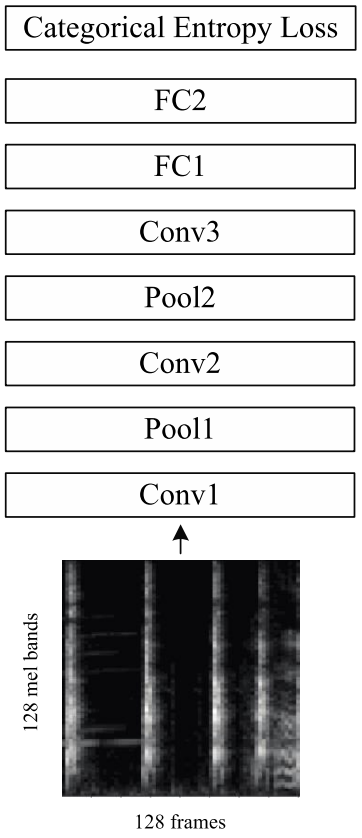
7-class spoken language recognition: Dutch, English, French, German, Italian, Russian, and Spanish

Input: 39-band log-mel spectrogram, 8.33 ms window/hop size, freq. range: 0 - 5kHz

Classification acc: 69.8%

[1] G. Montavon, Deep Learning for Spoken Language Identification, NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

Recording Tower Pre-training



Dataset: UrbanSound8K dataset

10-class environmental sound classification: Air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music.

Input: 128-band log-mel spectrogram, 23 ms window/hop size, freq. range: 0 – 22,050 Hz

Classification acc: 70.2%

[1] J. Salamon and J. P. Bello, Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification, IEEE Signal Processing Letters, 2017.

Dataset – VocalSketch [1]

Category (#concepts)	# train concepts	# test concepts	Examples
Acoustic Instruments (40)	20	20	Triangle
Commercial Synthesizers (40)	20	20	Metaloid
Everyday (120)	60	60	Knocking
Single Synthesizer (40)	20	20	Subsynth_2217

- Each concept has 10 imitations (~3 sec), ~2 hours in total
- Training: 120*7=840 positive / negative pairs
- Validation: 120*3=360 positive / negative pairs

[1] M. Cartwright and B. Pardo, VocalSketch: Vocally imitating audio concepts, in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015

Evaluation Measure

- Mean Reciprocal Rank (MRR)

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

Number of queries in experiment

Rank of the target sound in the returned sound list for the i-th query

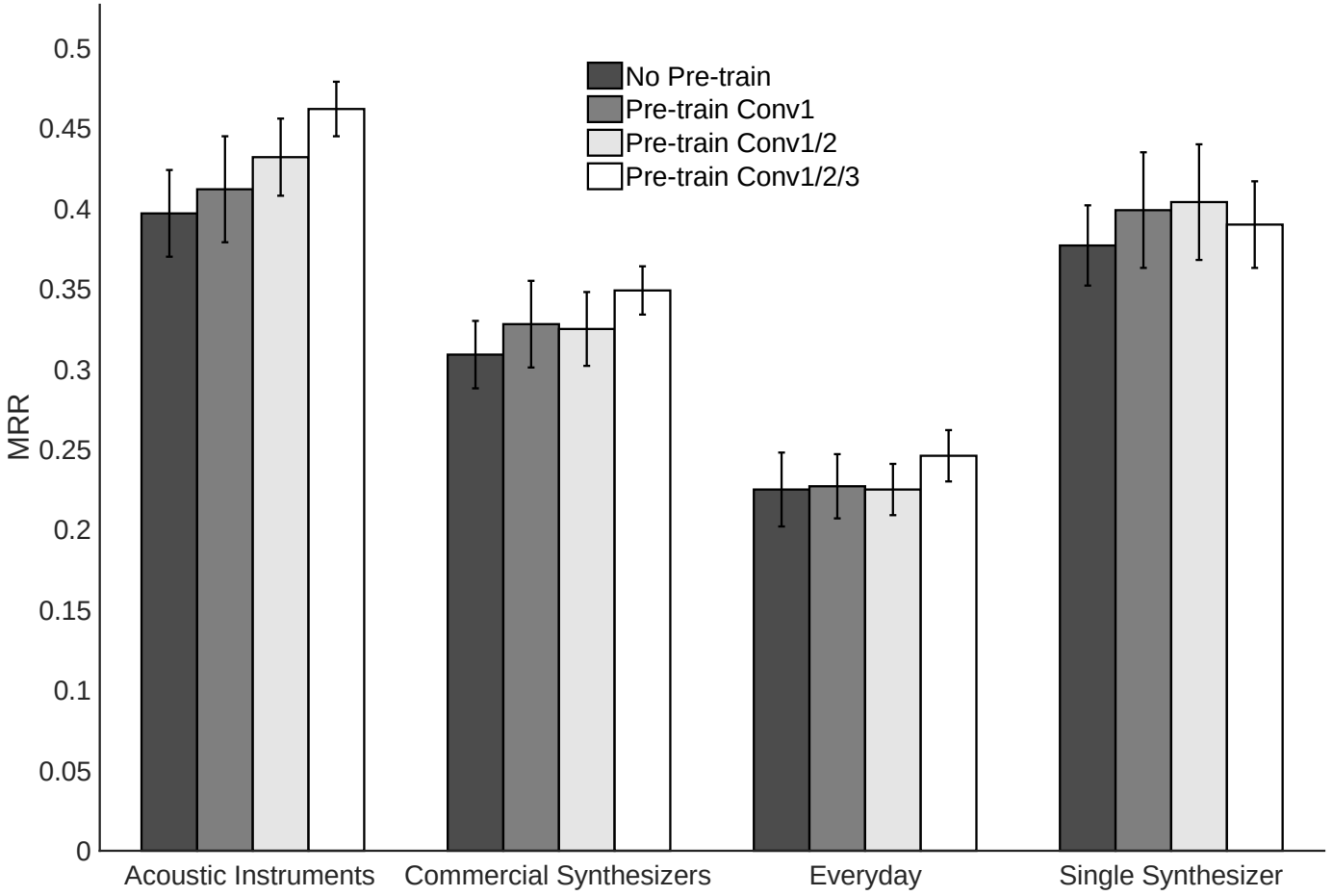
- ✓ 0 <= MRR <= 1
- ✓ The higher the better

Experimental Results

Table 2. MRR (mean ± std) comparisons.

Config.	Acoustic Instr.	Commercial Synthesizers	Everyday	Single Synthesizer
IMINET	0.40±0.03	0.33±0.02	0.16±0.01	0.38±0.02
TL-IMINET (w/o pretrain)	0.40±0.03	0.31±0.02	0.23±0.02	0.38±0.03
TL-IMINET (w/ pretrain)	**0.46**±0.02	**0.35**±0.02	**0.25**±0.02	**0.40**±0.03

Experimental Results



Visualization Using Activation Maximization

Motivation: understand how TL-IMINET works and what features are learned

Activation Maximization (AM): Neuron activation as objective function, using gradient ascent to update input pixels while keeping weights unchanged

Visualization Using Activation Maximization

CNN neurons: visualizing learned features from imitation/recording input

$$\underset{x}{\operatorname{argmax}}(A_{c_{ij}} - \lambda \|x\|^2)$$

Neuron activation in convolutional layers

Input of Imi/Rec tower

FC neurons: visualizing learned similarity between imitation/ and recording

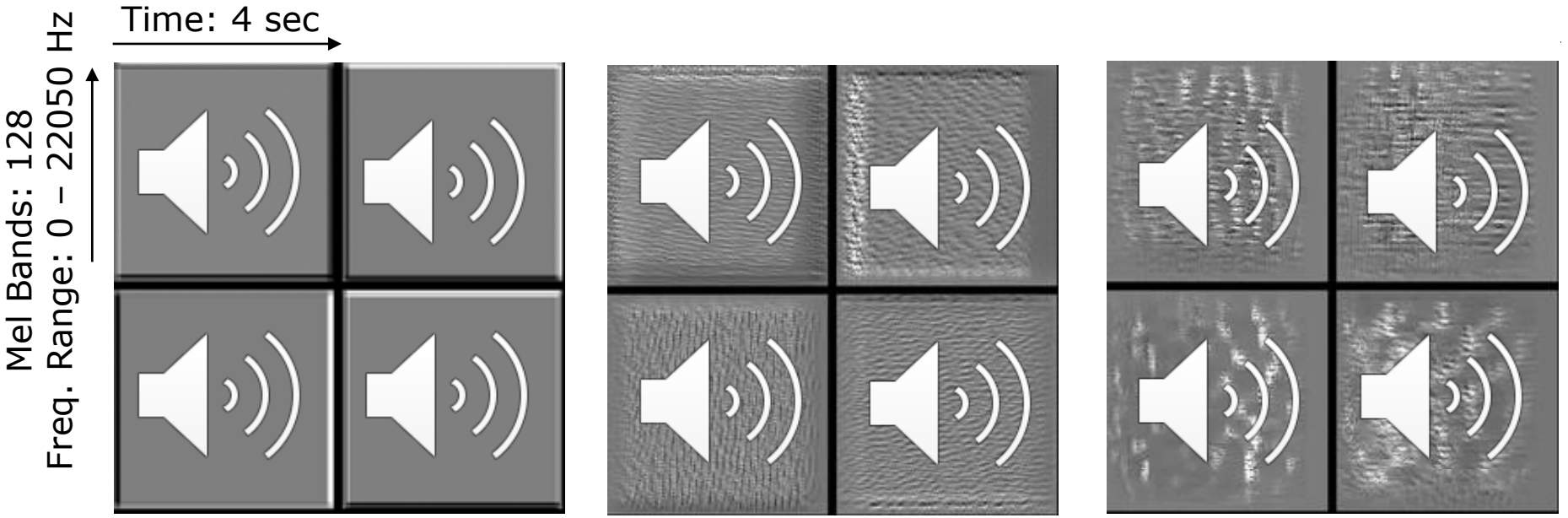
$$\underset{x_{imi},x_{rec}}{\operatorname{argmax}} [A_{f_{ij}} - \lambda(\|x_{imi}\|^2 + \|x_{rec}\|^2)]$$

Neuron activation in fully connected layers

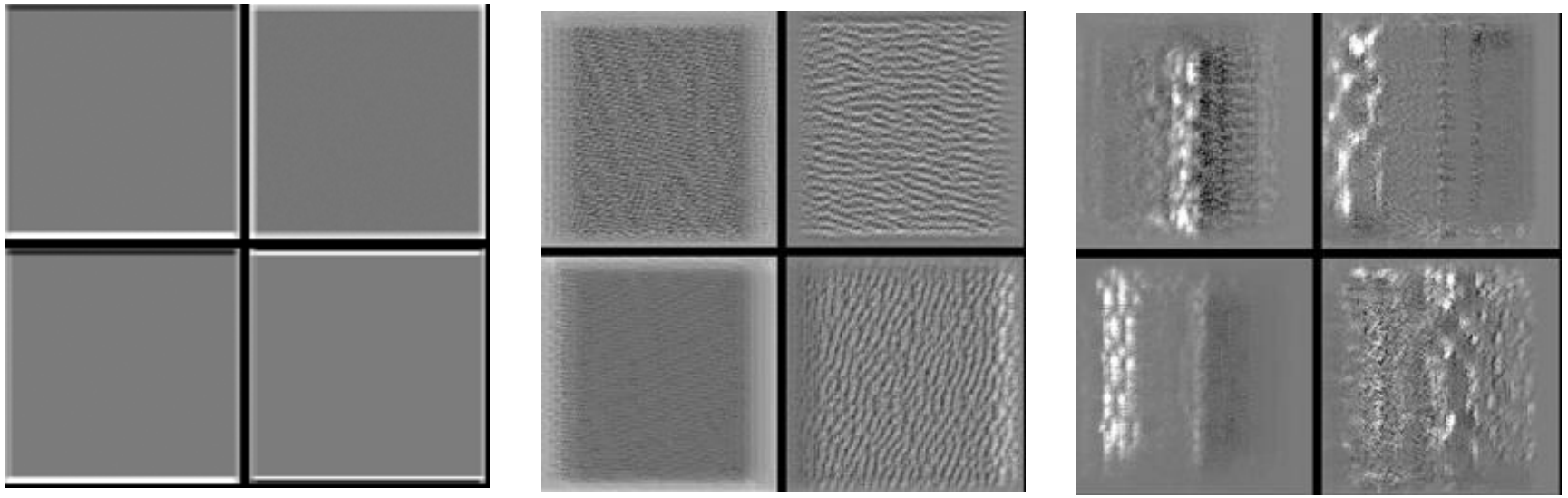
Input of Imi tower

Input of Rec tower

Recording Tower Visualization

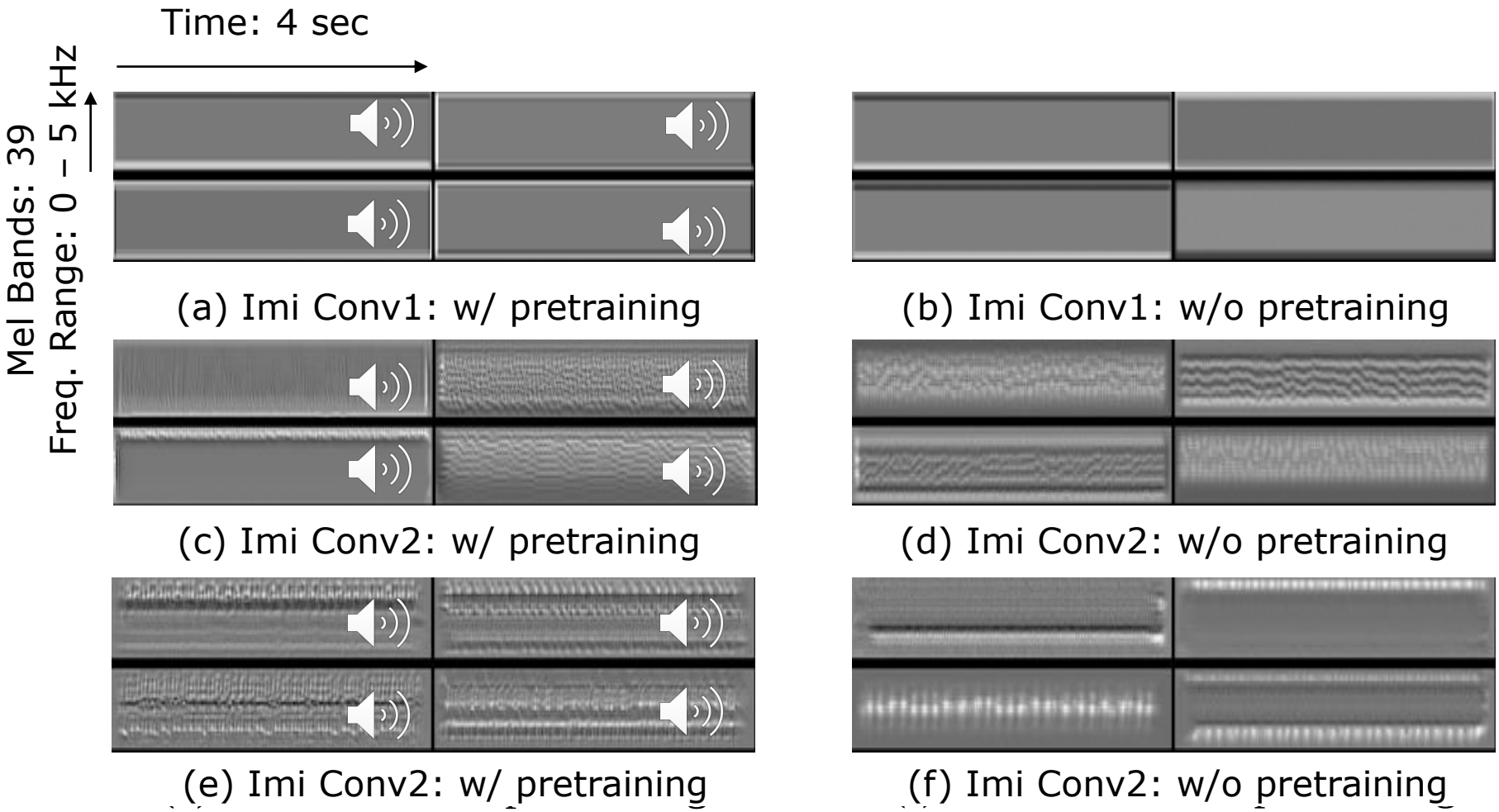


(a) Rec Conv1: w/ pretraining (b) Rec Conv2: w/pretraining (c) Rec Conv3: w/pretraining



(d) Rec Conv1: w/o pretraining (e) Rec Conv2: w/o pretraining (f) Rec Conv3: w/o pretraining

Imitation Tower Visualization



(a) Imi Conv1: w/ pretraining

(b) Imi Conv1: w/o pretraining

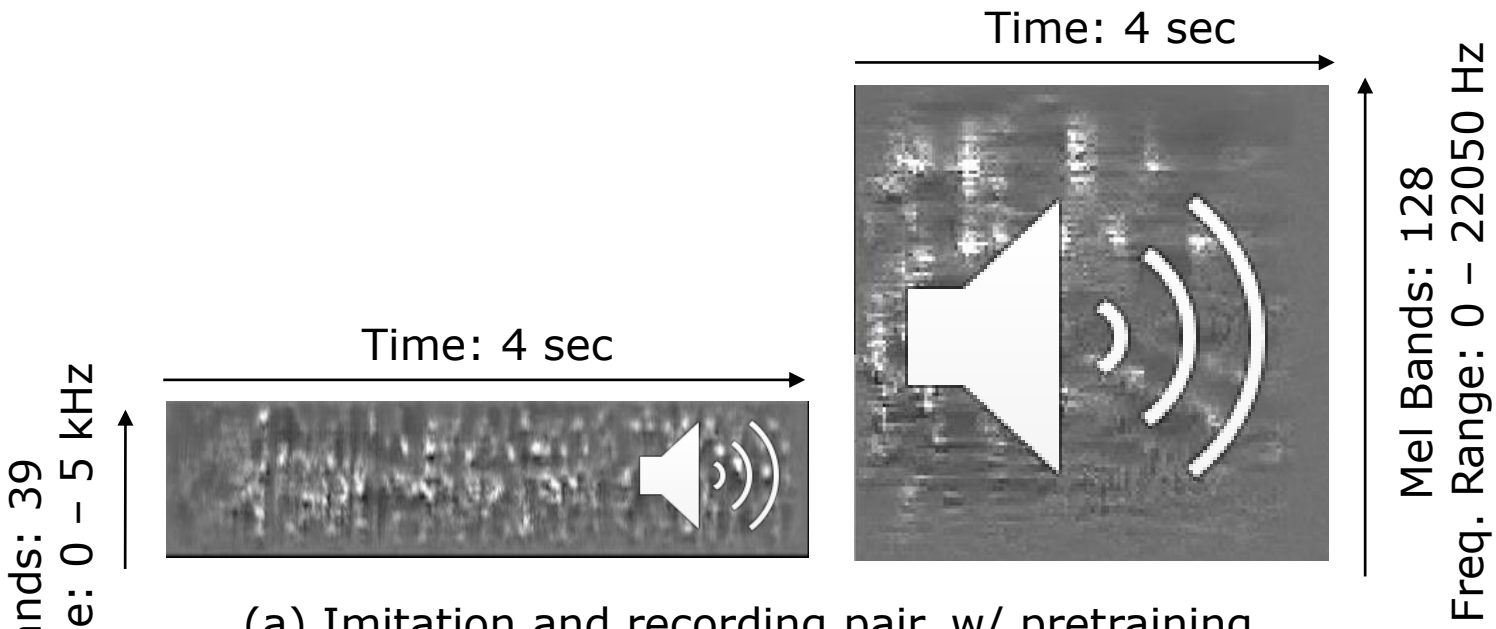
(c) Imi Conv2: w/ pretraining

(d) Imi Conv2: w/o pretraining

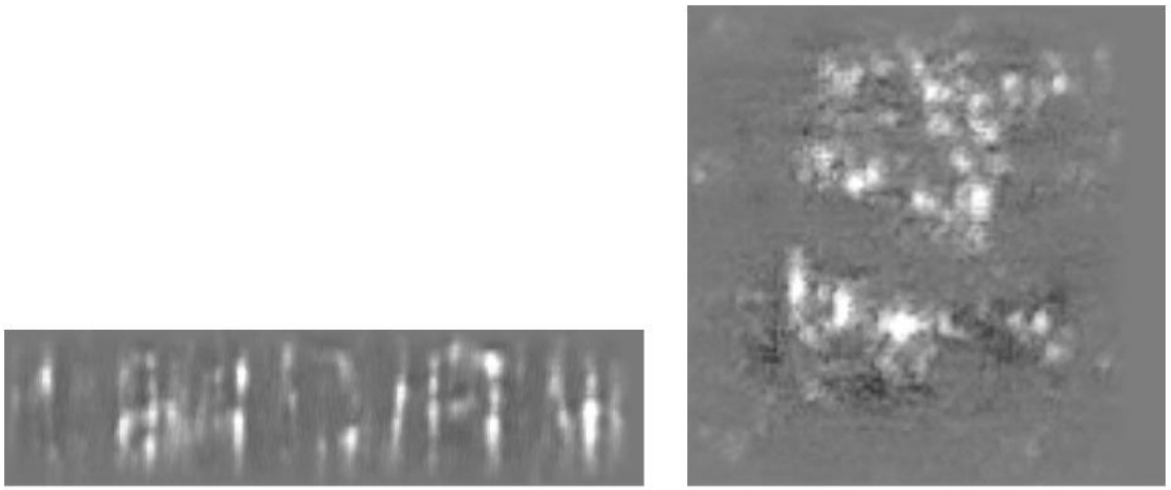
(e) Imi Conv2: w/ pretraining

(f) Imi Conv2: w/o pretraining

FC Layer Visualization



(a) Imitation and recording pair, w/ pretraining



(b) Imitation and recording pair, w/o pretraining

Conclusions & Future Work

Conclusions

- Proposed transfer learning based Siamese style network: TL-IMINET
- Interpreted how TL-IMINET works by visualizing input patterns that maximally activate neurons

Future work

- Conduct subjective studies to use TL-IMINET

Vision

- Sound query by vocal imitation will be widely available

The End

Thank you for your attention !