# IMINET: Convolutional Semi-Siamese Networks for Sound Search by Vocal Imitation

Yichi Zhang & Zhiyao Duan

Department of Electrical and Computer Engineering, University of Rochester

{yichi.zhang, zhiyao.duan}@rochester.edu

## Introduction

Q: How to search for a sound that matches the concept in your head?

A: Current ways: through its name or other semantic labels.

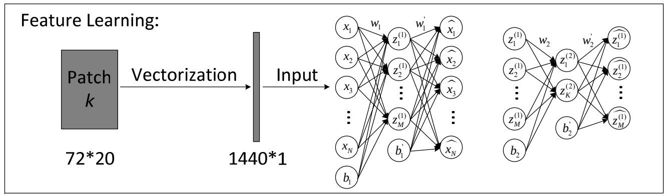
Q: What if you don't remember its name, or what you are looking for simply doesn't have a semantic meaning?

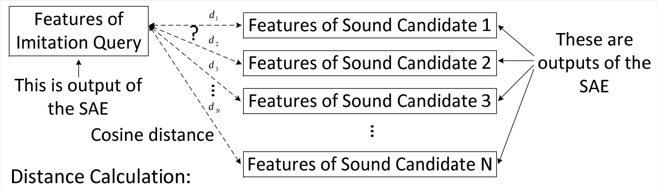
**A: Imitate the concept with your voice!**

- Dog barking sound: infantile bark threat bark
- Synthesized sound:

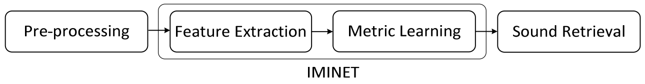
## Prior Work

IMISOUND: Feature learning through SAE on vocal imitations + Predefined distance calculation between imitation query and sound candidates [1]


Feature Learning:
Patch k
Vectorization
Input
72*20
1440*1
$x_1$ $x_2$ $x_3$ $x_N$ $b_1$
$w_1$ $w_1'$
$z_1^{(1)}$ $z_2^{(1)}$ $z_M^{(1)}$ $b_2$
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$ $\hat{x}_N$
$w_2$ $w_2'$
$z_1^{(2)}$ $z_2^{(2)}$ $z_K^{(2)}$ $b_2$
$\hat{z}_1^{(1)}$ $\hat{z}_2^{(1)}$ $\hat{z}_M^{(1)}$


Features of Imitation Query
This is output of the SAE
Cosine distance
Distance Calculation:
?
$d_1$ $d_2$ $d_3$ $d_N$
Features of Sound Candidate 1
Features of Sound Candidate 2
Features of Sound Candidate 3
Features of Sound Candidate N
These are outputs of the SAE

## Proposed System



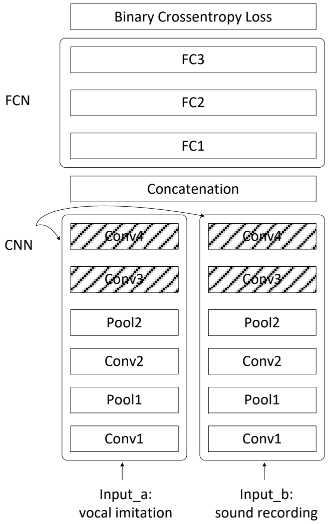
- Data-driven: learns both features and similarities from data
- Supervised training: requires pos/neg pairs of imitations and sounds
- Unsupervised retrieval: no need to train on imitation-sound pairs of a certain sound concept for the retrieval of that sound concept

References:

[1] Y. Zhang and Z. Duan, "IMISOUND: AN unsupervised system for sound query by vocal imitation," ICASSP 2016.

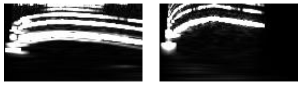
[2] M. Cartwright and B. Pardo, "VocalSketch: Vocally imitating audio concepts," CHI 2015.

## The IMINET Model



FC3: 1 unit, sigmoid activation
FC2: 32 units, ReLU activation
FC1: 432 units, ReLU activation

Conv4: 6 filters, receptive field: 3 * 3
Conv3: 6 filters, receptive field: 3 * 3
Pool2: maximum, stride: 2 * 4
Conv2: 12 filters, receptive field: 3 * 3
Pool1: maximum, stride: 2 * 4
Conv1: 12 filters, receptive field: 3 * 3



Example of a positive pair:

Left: CQT spectrogram of an imitation of a police siren

Right: CQT spectrogram of a recording of a police siren

**Step 1: Pre-processing**

Constant-Q Transform (CQT) with 2 seconds raw input
Spectrogram size: 72 * 129

**Step 2: Feature Extraction**

Use Convolutional Neural Networks (CNN) to learn features in two towers with tied, partially tied (layers in shade are weight-shared), and untied weights

Note: configs. in blue are Semi-Siamese

**Step 3: Metric Learning**

Use Fully Connected Networks (FCN) to learn the pair-wise similarity and generate a single value output in [0, 1]

**Step 4: Sound Retrieval**

Pair the imitation query with each recording in the library to calculate its likelihood of being a positive pair. Likelihood scores are ranked in descending order

## Late Fusion

- Fuse the retrieval results (similarity likelihoods) of tied, partially tied, and untied weights of IMINET:

$$L_{fusion}(i) = L_{tied}(i) * L_{untied}(i) * L_{partial}(i)$$

- Fuse the retrieval results of IMINET with IMISOUND:

$$L_{sae}(i) = \frac{e^{-D(i)}}{\sum_{n=1}^{N} e^{-D(n)}} \qquad L_{fusion}(i) = L_{csn}(i) * L_{sae}(i)$$

## Dataset & Evaluation Measure

Table 1. VocalSketch Data Set V1.0.4 [2]

| Category | # classes | # samples |
|---|---|---|
| Acoustic instr. | 40 | 400 |
| Comm. Synthesizers | 40 | 404 |
| Everyday | 120 | 1209 |
| Single synthesizer | 40 | 405 |

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

## Experimental Results

Table 2. MRR (mean ± std) comparisons of various IMINET configurations

| Category | Config. | Acoustic Instr. | Comm. Synthesizers | Everyday | Single Synthesizer |
|---|---|---|---|---|---|
| Baseline | IMISOUND | 0.450 | 0.308 | 0.126 | 0.380 |
| Proposed | Untied | 0.377 ± 0.019 | 0.318 ± 0.020 | 0.154 ± 0.014 | 0.325 ± 0.020 |
| Proposed | Partial | 0.384 ± 0.027 | 0.304 ± 0.015 | 0.154 ± 0.015 | 0.340 ± 0.031 |
| Proposed | Tied | 0.401 ± 0.028 | 0.327 ± 0.019 | 0.158 ± 0.012 | 0.380 ± 0.018 |
| Proposed | Untied + Partial + Tied | 0.438 ± 0.015 | 0.343 ± 0.020 | 0.175 ± 0.012 | 0.382 ± 0.013 |
| Proposed | Untied + IMISOUND | 0.470 ± 0.025 | 0.356 ± 0.011 | 0.168 ± 0.010 | 0.402 ± 0.022 |
| Proposed | Partial + IMISOUND | 0.496 ± 0.018 | 0.346 ± 0.025 | 0.173 ± 0.014 | 0.417 ± 0.025 |
| Proposed | Tied + IMISOUND | 0.504 ± 0.014 | 0.355 ± 0.016 | 0.171 ± 0.009 | 0.452 ± 0.020 |
| Proposed | Untied + Partial + Tied + IMISOUND | 0.520 ± 0.020 | 0.371 ± 0.013 | 0.188 ± 0.007 | 0.447 ± 0.012 |

## Conclusions

- Proposed IMINET, a convolutional semi-Siameses network that learns both features and similarities, to search sounds by vocal imitation
- Proposed three IMINET configurations by choosing different weight sharing strategies between the two towers
- Proposed late fusion of the retrieval results of different IMINET configurations and those of IMISOUND to improve retrieval performance