Poster for MLSP 2015

Retrieving Sounds by Vocal Imitation Recognition

Yichi Zhang & Zhiyao Duan

AIR Lab, Department of Electrical and Computer Engineering, University of Rochester

Introduction

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

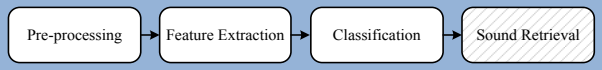
A: Current ways: through its name or other semantic labels.

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!

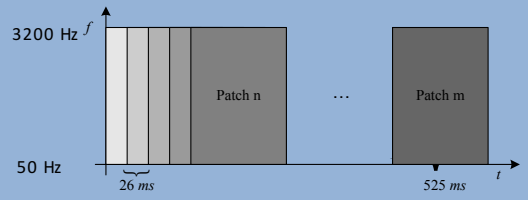
- Dog barking sound: infantile bark threat bark
- Synthesized sound:

Proposed System



Pre-processing:

Convert imitation audio into spectrogram by Constant-Q Transform (CQT), then segment it into overlapping patches.

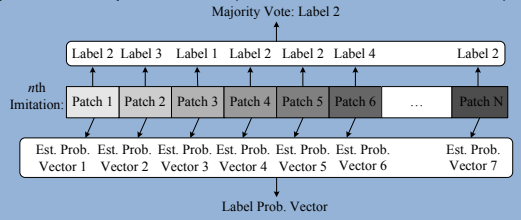


Feature Extraction:

Use Stacked Auto-encoder (SAE) to learn features from training patches automatically.

Classification & Retrieval:

Use multi-class Support Vector Machine (SVM) to generate probability output for concept retrieval. (Limitation: close-set scenario)



Challenges

A big challenge in vocal imitation recognition is feature extraction.

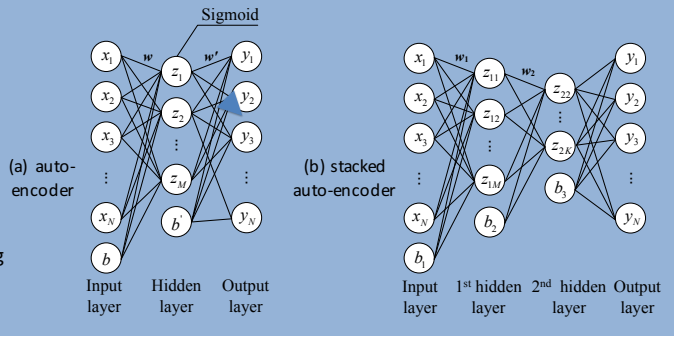
- People tend to imitate different aspects for different recordings:

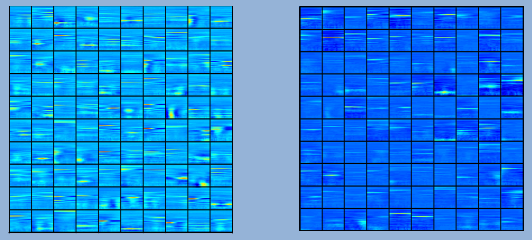
car horn: [ ] cat: [ ] guitar note: [ ]

- Even for the same recording, different people imitate differently:

car horn 1: car horn 2: car horn 3:

Automatic Feature Learning





(a) the 1st hidden layer (b) the 2nd hidden layer

Experimental Results

Table 1. Description of the VocalSketch v1.0.4 dataset

Category	# classes	Sound Concepts
Acoustic instruments	40	Orchestral instruments playing a single note with the pitch C (in an appropriate octave chosen for each instrument)
Commercial synthesizers	40	Various recordings from Apple's Logic Pro music production suite
Everyday	120	A wide variety of acoustic events in everyday life
Single synthesizer	40	Recordings from a single 15-parameter subtractive synthesizer playing a note with the pitch C (octave varies depending on the parameter settings)

Table 2. Recording-level 10-fold cross validation results.

Category	# classes	Proposed		MFCC	
		Accuracy	MRR	Accuracy	MRR
Acoustic instruments	17	**23.61%**	**0.4259**	21.94%	0.3789
Commercial synthesizers	13	**20.00%**	**0.3577**	12.69%	0.2960
Everyday	48	**10.71%**	**0.2666**	10.00%	0.2368
Single synthesizer	40	**12.00%**	**0.2732**	6.25%	0.2188

Acknowledgements

We thank Mark Cartwright and Bryan Pardo for generously providing us with the VocalSketch Data Set v1.0.4.