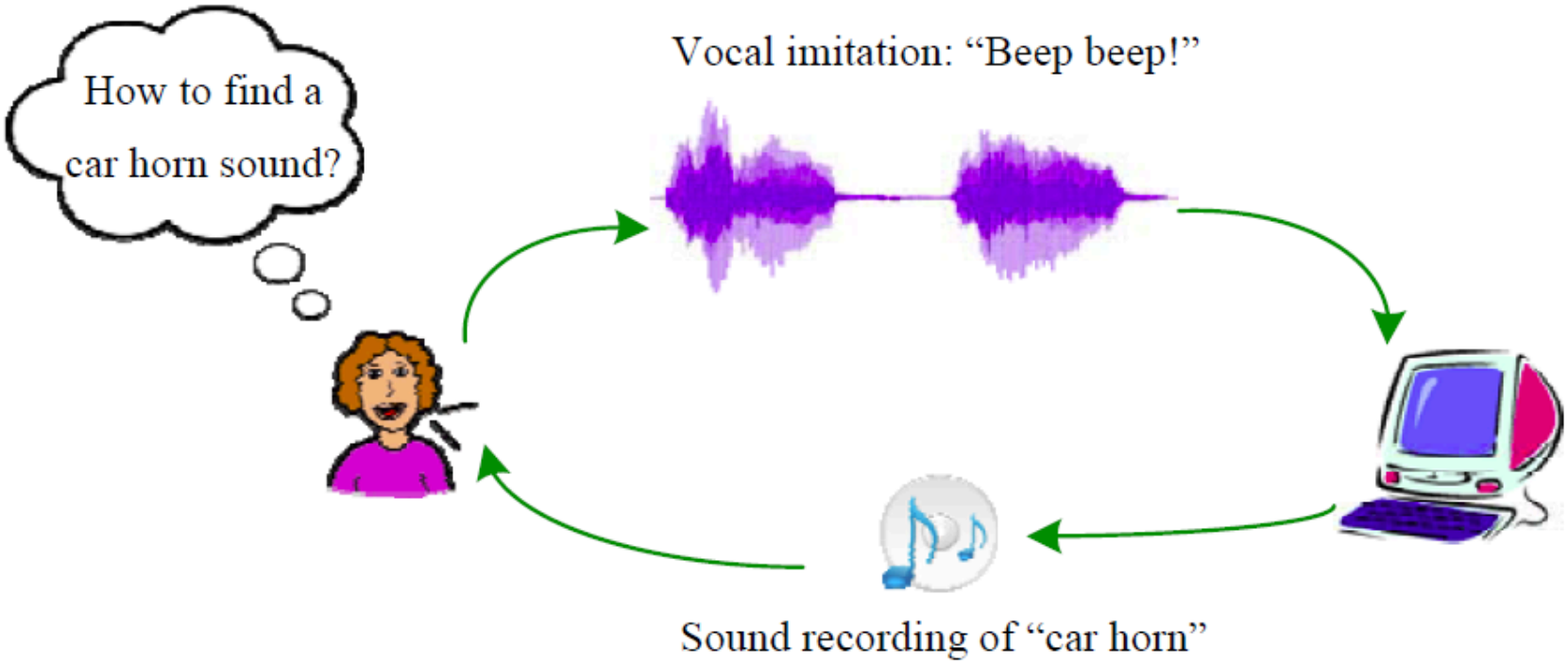
IMISOUND: An Unsupervised System for Sound Query by Vocal Imitation

Yichi Zhang and Zhiyao Duan

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

Query by vocal imitation



Query by vocal imitation

For general sounds:

- Dog barking sound (w/ semantic meaning)

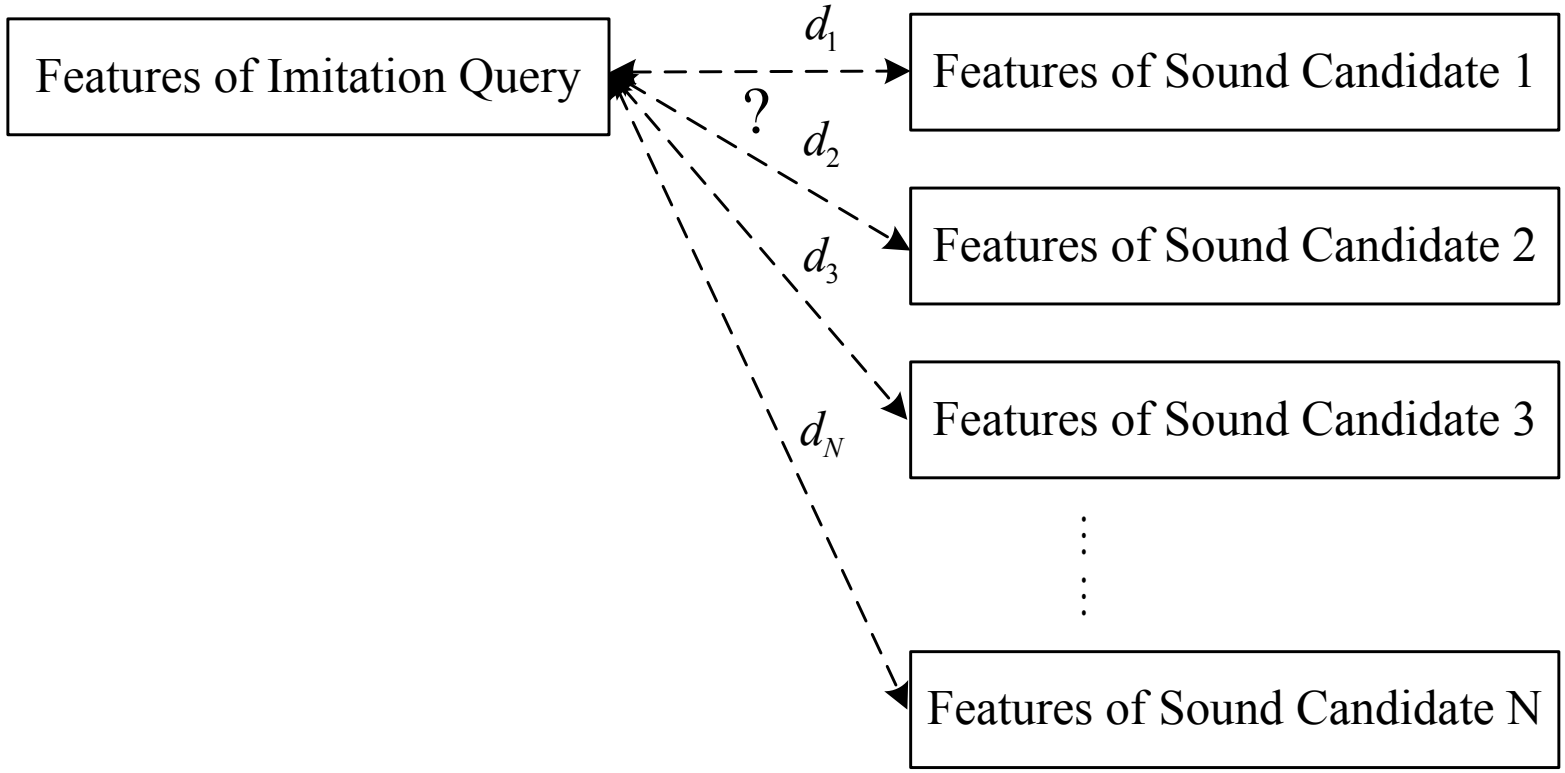
infantile bark or threat bark

Vocal imitation: narrow down the concept

- Synthesized sound (w/o semantic meaning)

Vocal imitation: might be the only way to convey the concept

Towards Sound Retrieval



Challenges

- People tend to imitate different aspects for different recordings

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

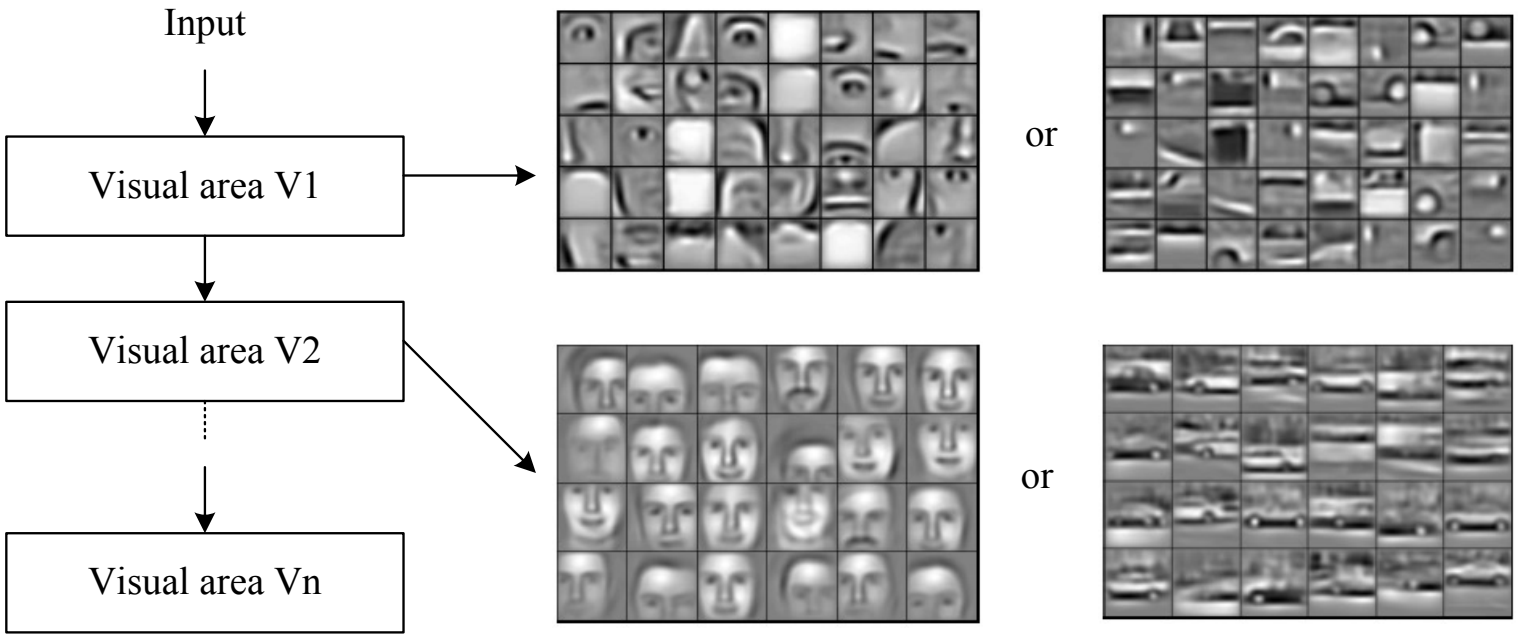
- Even for the same recording, different people may imitate differently

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

- Hand crafted features such as pitch, timbre, loudness, etc. would not work well…

- Solution: Deep Neural Networks (DNN)

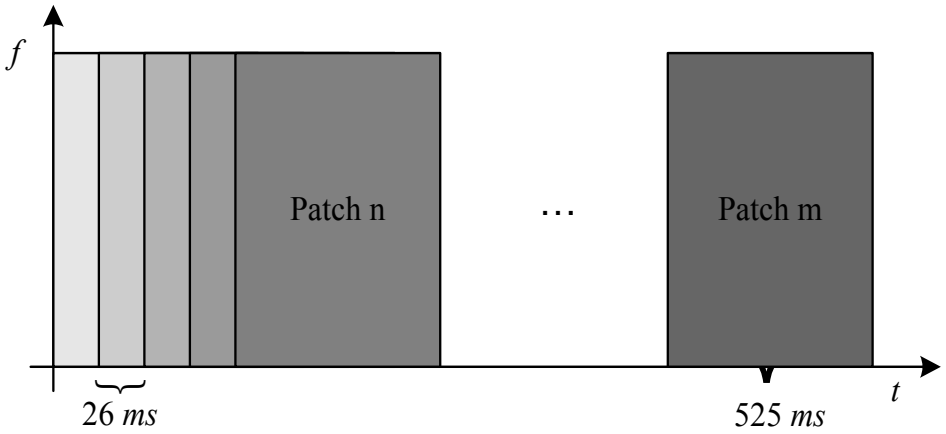
Automatic feature learning



[1] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, Unsupervised learning of Hierarchical representations with convolutional deep belief networks, 2011

Pre-processing

Constant Q Transform (CQT) spectrogram



Parameters:

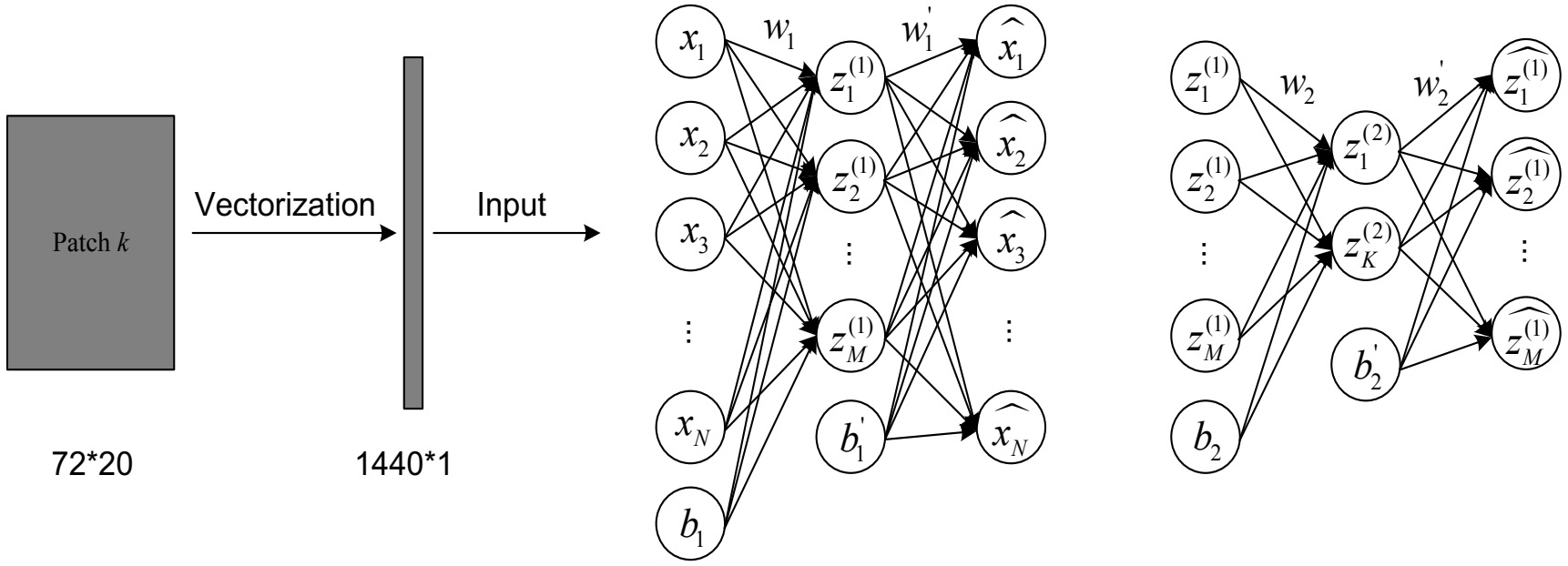
- ✓Patch length: 525 ms (20 frames)
- ✓Freq range: 50-3200 Hz (6 octaves)
- ✓12 bins per octave

Rationale:

- ✓ one syllable in normal English speech: 200 ms
- ✓ 50 Hz to 3200 Hz basically covers telephone frequency range

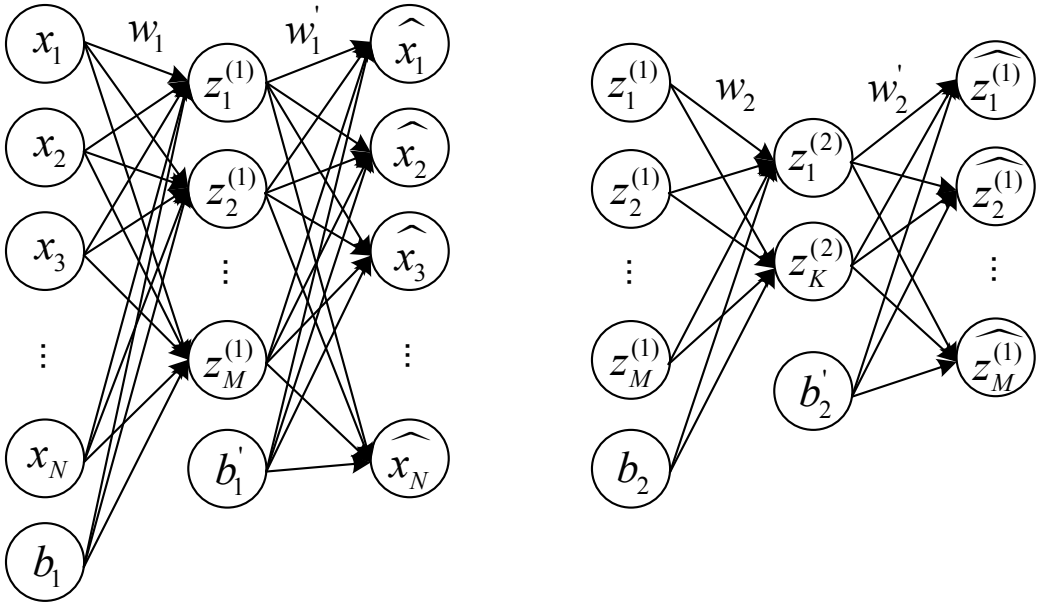
Feature Extraction

- Stacked Auto-encoder (SAE) is chosen as the neural network model



Feature Extraction

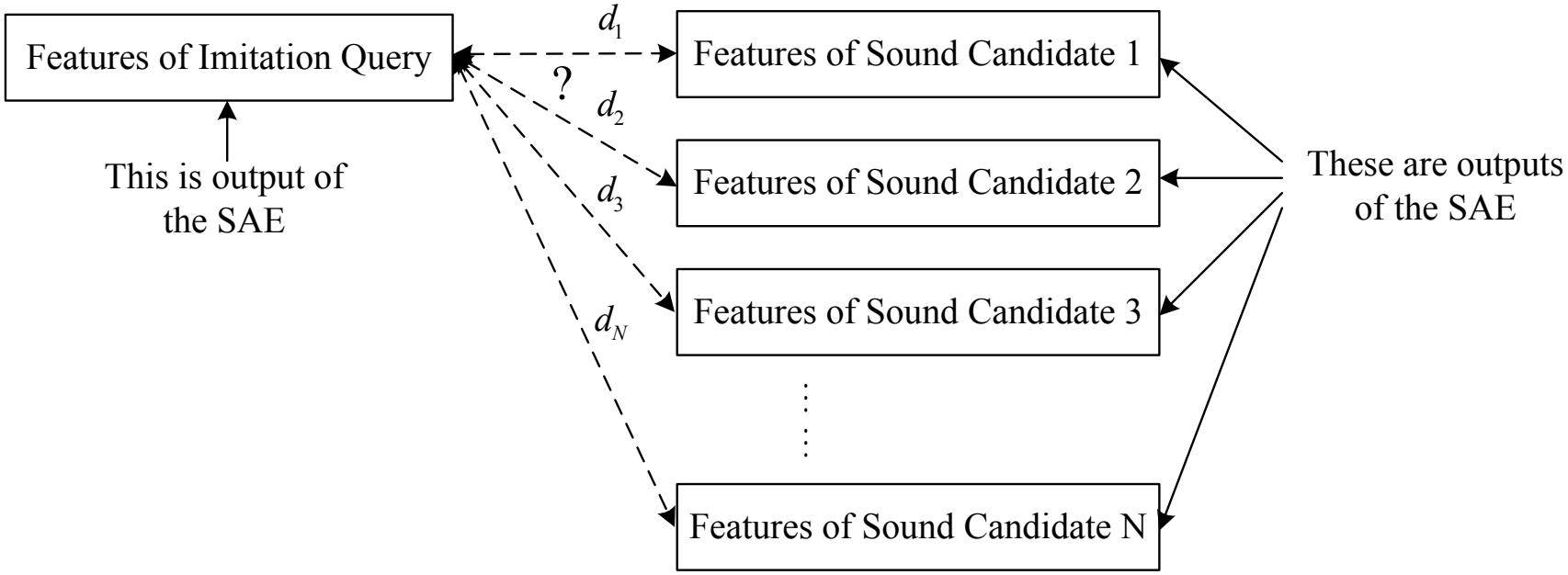
Auto-encoder tries to learn the weights and biases so that the output could approximate the input



1st hidden layer neurons = 500 2nd hidden layer neurons = 100

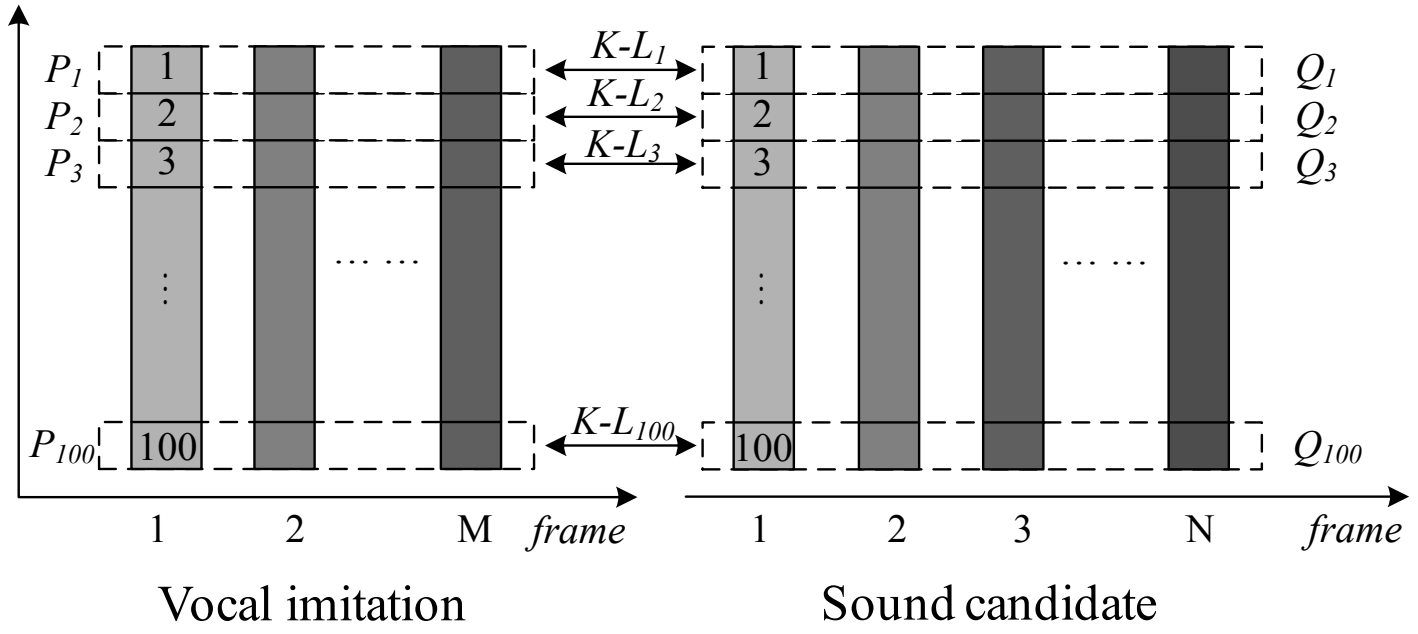
- Weights are trained by half of all the vocal imitations

Distance Calculation

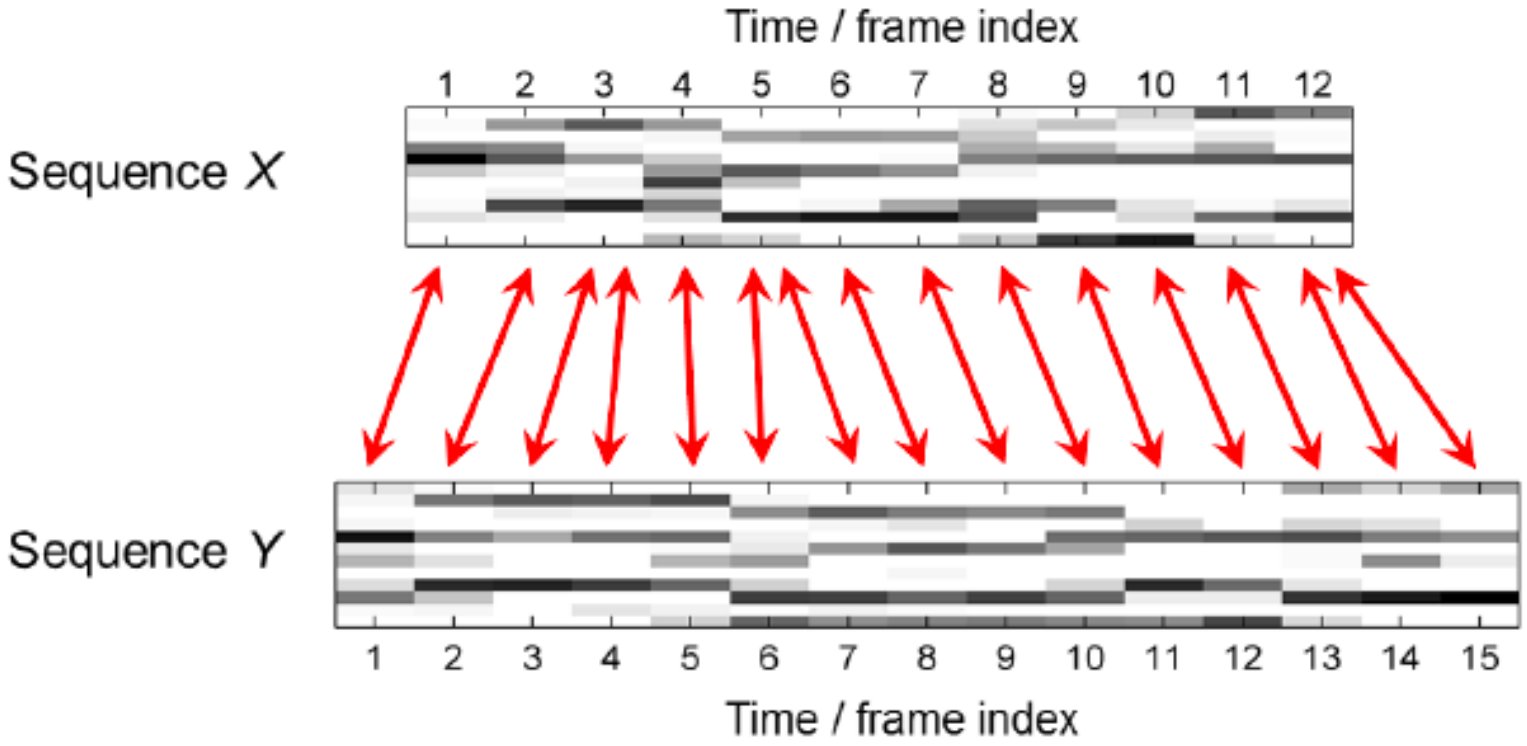


K-L Divergence

$$D_{kl\_sym}(P \| Q) = \frac{1}{2}(D_{kl}(P \| Q) + D_{kl}(Q \| P)) = \frac{1}{2}(\sum_i P(i) \ln \frac{P(i)}{Q(i)} + \sum_i Q(i) \ln \frac{Q(i)}{P(i)})$$

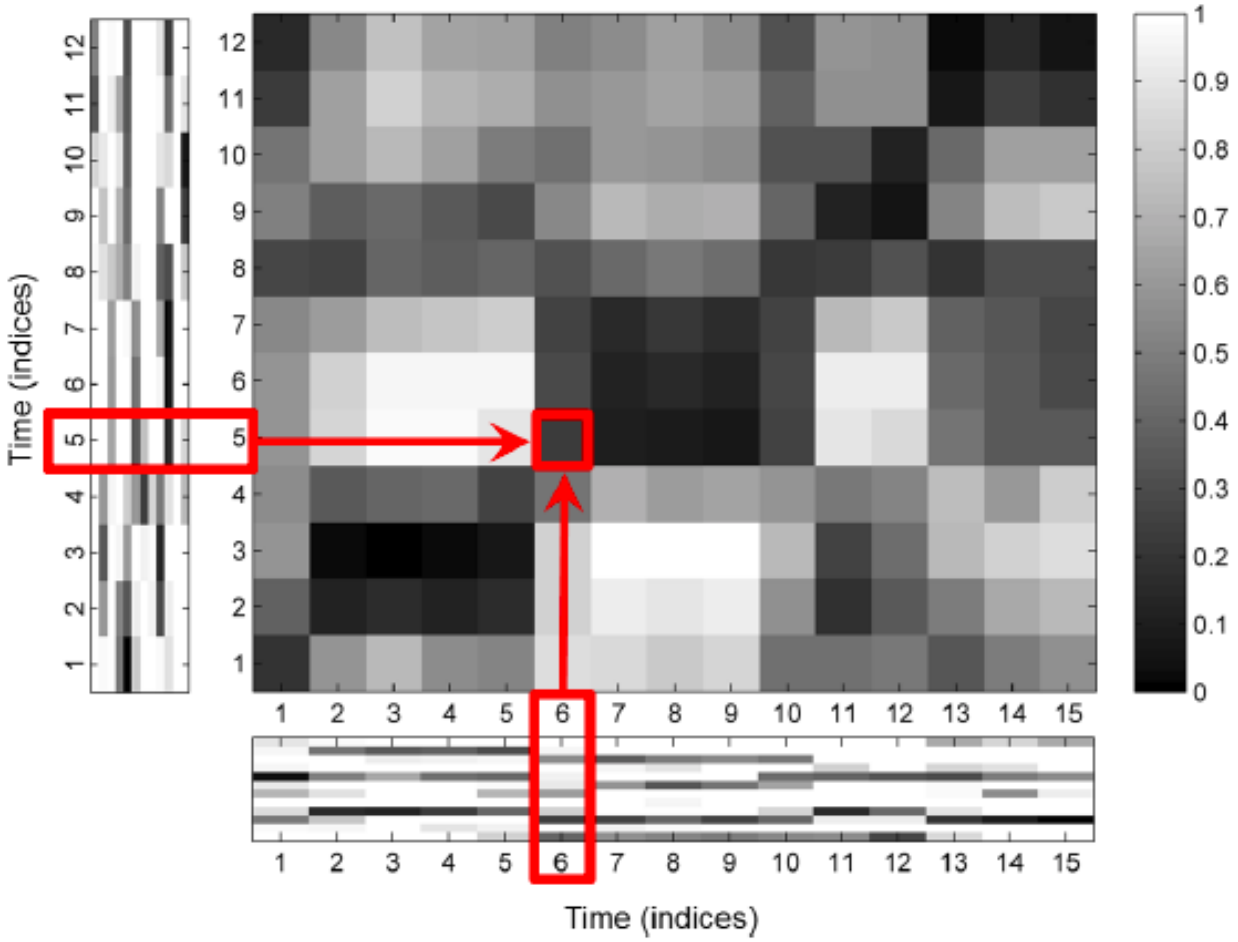


DTW Distance



[1] A. Mueller, First course on music processing, Preliminary version, 2014

DTW Distance

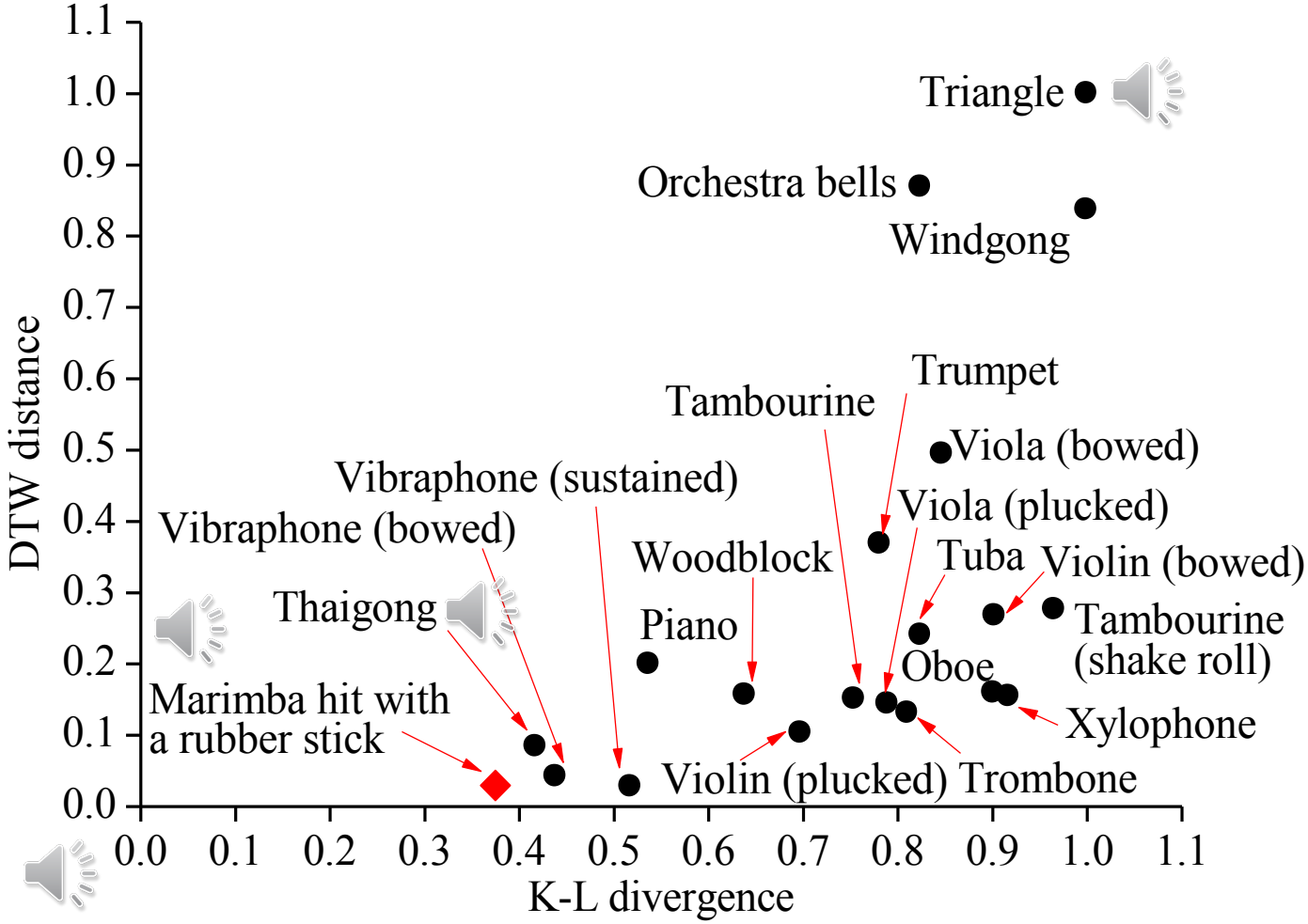


Distance Calculation

- K-L divergence: dissimilarity in probability distribution
- DTW distance: dissimilarity in temporal domain

$$D = \frac{D_{KL}}{\max(D_{KL})} + \frac{D_{DTW}}{\max(D_{DTW})}$$

Sound Retrieval Example



Experimental Setup

1) Use vocal imitations of half of all the sound concepts to train the SAE
    - # hidden layers = 2
    - # neurons in the 1st hidden layer = 500
    - # neurons in the 2nd hidden layer = 100

2) Use the other half for sound retrieval experiment within each category
    - # sound concepts in Acoustic Instruments = 20
    - # sound concepts in Commercial Synthesizers = 20
    - # sound concepts in Everyday = 60
    - # sound concepts in Single Synthesizer = 20

Dataset

Table 1. VocalSketch Data Set v1.0.4 [1]

Category	Sound Concepts (#)	Examples
Acoustic Instruments	Orchestral instruments playing a C note (40)	Orchestra bells Triangle
Everyday	Acoustic events in everyday life (120)	Knocking Sheep
Commercial Synthesizers	Apple's Logic Pro (40)	Metaloid Shimmer
Single Synthesizer	A single 15-parameter subtractive synthesizer playing C note (40)	Subsynth_2217 Subsynth_8828

- Each class has 10 vocal imitations on average

[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

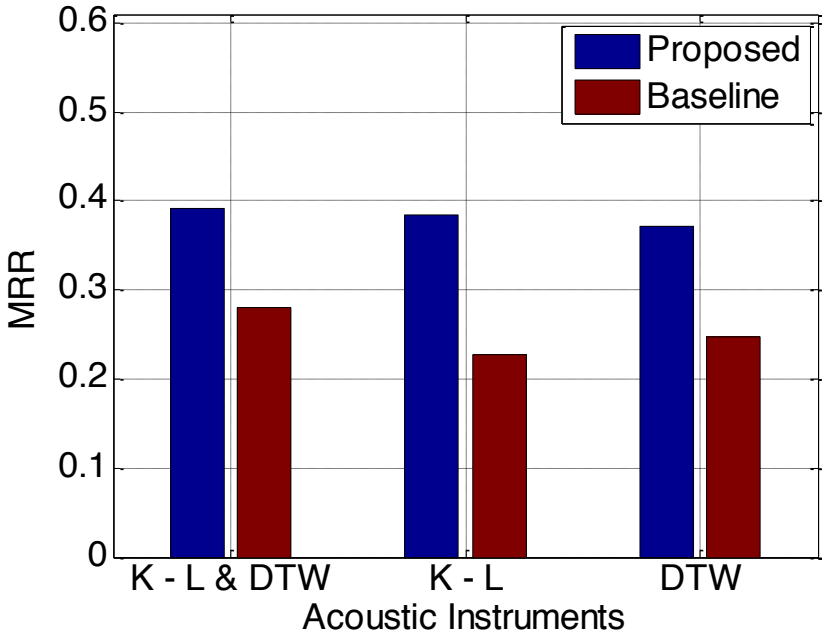
- ✓ 0 <= MRR <= 1
- ✓ The higher the better

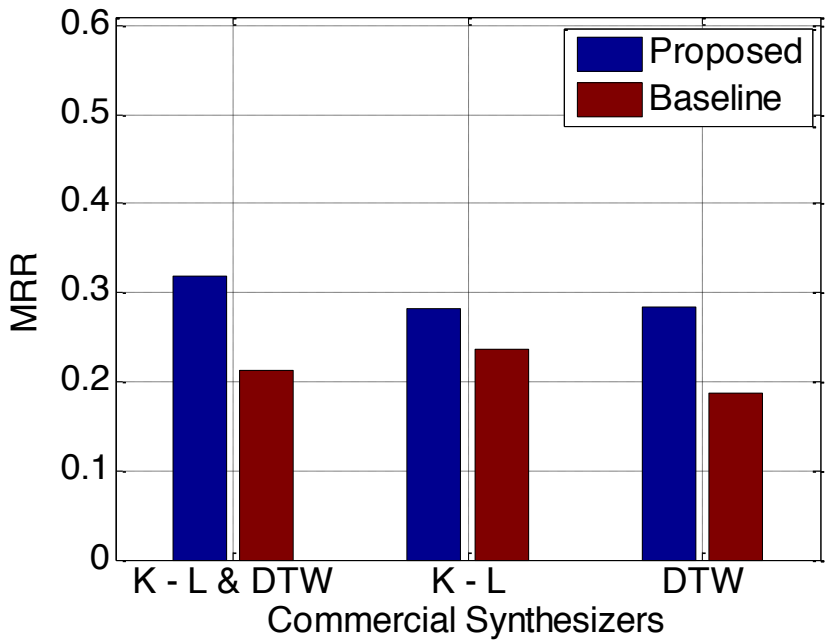
Comparison Method

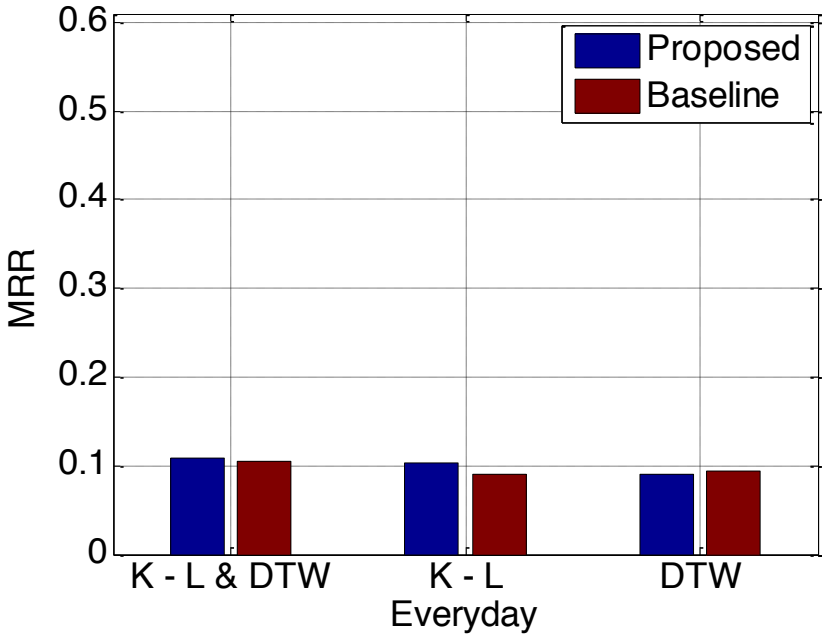
Hand-crafted features:
Mel-frequency cepstral coefficients (MFCC)

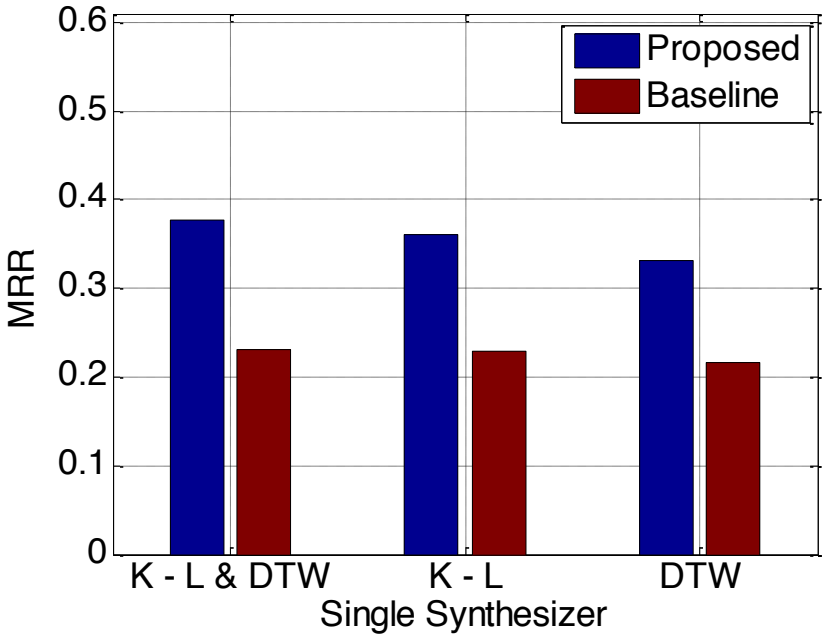
39-dimensional MFCC vectors, including

- 13 MFCC coefficients
- 13 first-order derivatives
- 13 second-order derivatives

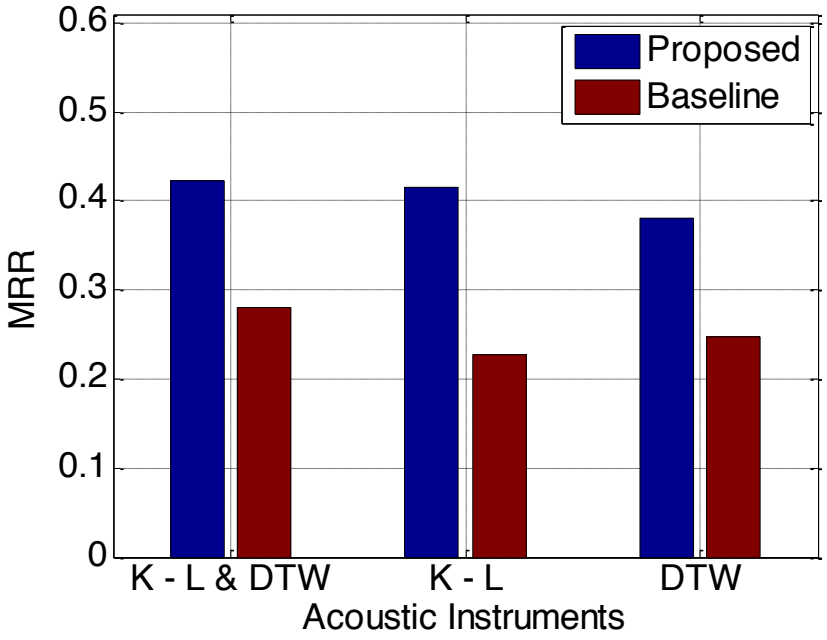


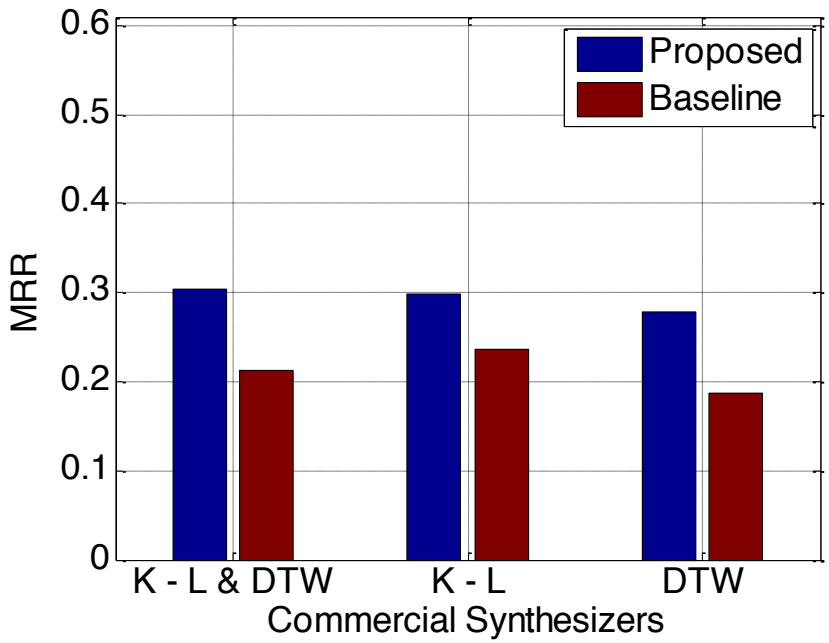


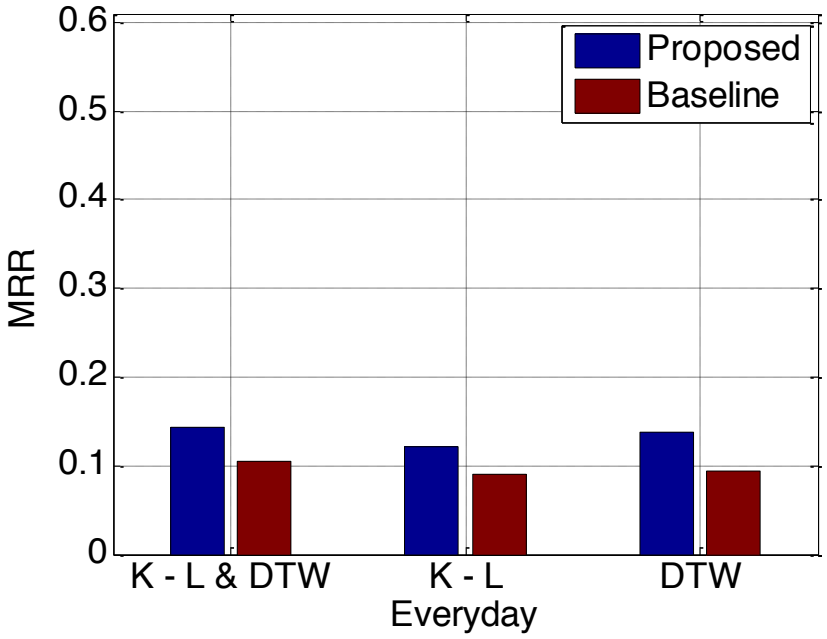


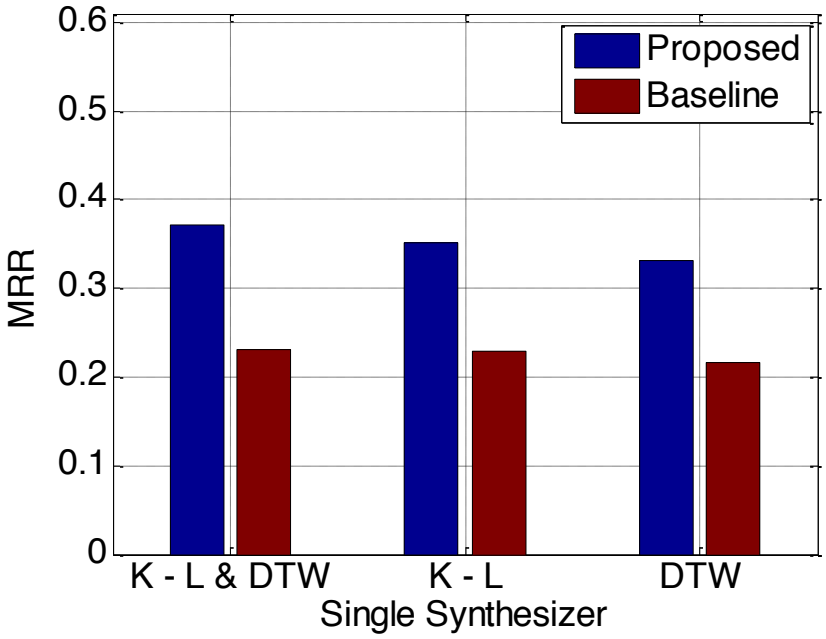


 # neurons in the 1st hidden layer: 500 **# neurons in the 2nd hidden layer: 100**









 # neurons in the 1st hidden layer: 1000 # neurons in the 2nd hidden layer: 600

Conclusions & future work

Conclusions

- Proposed the first unsupervised sound query-by-vocal-imitation system which is evaluated in a large dataset
- Achieved significantly better results by automatic feature learning than hand-crafted features

Future work

- Experiments on CNN and RNN

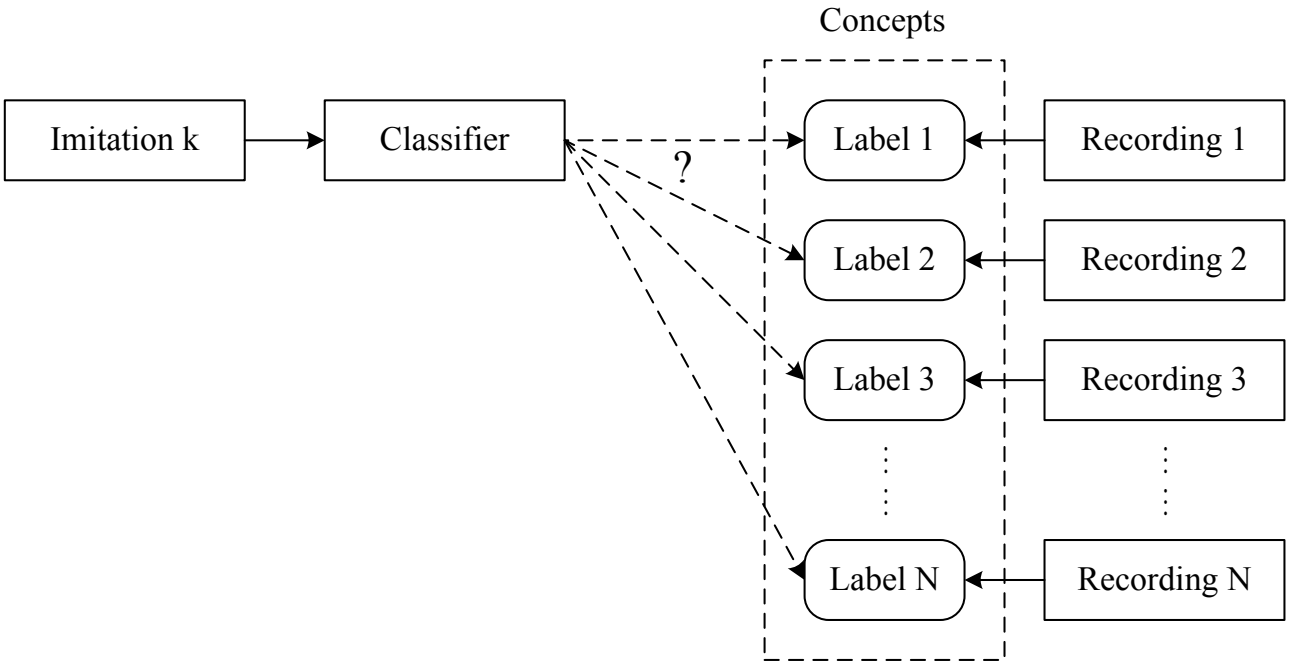
Vision

- Sound query by vocal imitation will be the trend

The End

Thank you for your attention !

Supervised Query-by-Vocal-Imitation System



Assumptions:

- Closed set scenario
- Training data exist for each concept