Improving Content-based Audio Retrieval by Vocal Imitation Feedback

*Bongjun Kim and Bryan Pardo
Computer Science, Northwestern University, USA

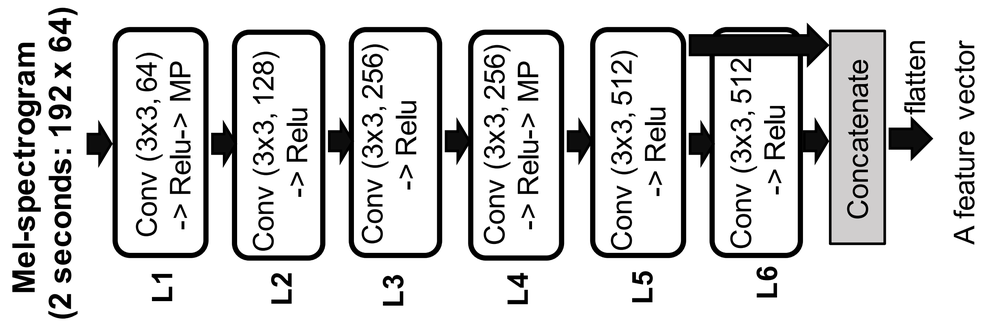
*bongjunkim.com
bongjun@u.northwestern.edu

Content-based audio retrieval

- Finding the desired audio by providing an audio example (e.g. find recordings that sound similar to this dog bark).
- Useful when there is no search-relevant metadata about the audio contents, so text-based search fails.
- We present a way to improve search results using **vocal imitation feedback** in two search scenarios: **Query-by-vocalization and Query-by-example**

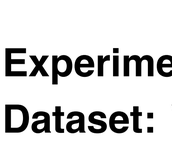
Audio features for vocal imitation feedback

Fully CNN-based feature extractor using conv layers from VGGish pretrained model [4]. The similarity between vocal imitations and original recordings are measured in the feature space.



Scenerio-1: Query-by-Vocalization

❑ Problem



Query (Imitation of a dog bark)

- What if the top ranked item is not the target sound?
- Before listening to other items sequentially, **can a user improve the search results quickly?**

❑ Solution: Provide a second vocal imitation of either the desired target or the top ranked **irrelevant** sound item (negative)

The updated similarity:

$$S(q,x) = \frac{1}{N_{vp}} \sum_{i=1}^{N_{vp}} C(v_p^i, x) - \frac{1}{N_{vn}} \sum_{i=1}^{N_{vn}} C(v_n^i, t) \cdot C(v_n^i, x)$$

- q: the original query (mixed)
- X: a recording in the datasbase
- v^i_p: i-th positive vocal imitation (N_{vp} in total)
- v^i_n: i-th negative vocal imitation (N_{vn} in total)
- t: the top-ranked erroneous search item
- C: cosine similarity function

❑ Experiment setting

- **Dataset:** VocalSketch dataset [1] containing 240 recordings of 4 categories: Acoustic Instrument (AI), Comercial Synthesizer (CS), Everyday sound (ED), Signle Synthesizer (SS). Each recording has 10 associated vocal imitations.
- **Performance Metric:** Mean Reciprocal Rank
- User feedback simulation

If the top-ranked item is not the target sound, the initial search ranking is updated using vocal imitation feedback

❑ Result

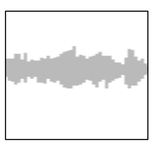
Model	AI	CS	ED	SS
TL-IMINET [2]	0.503	0.439	0.259	0.452
Our model	0.545	0.488	0.499	0.518
Our model-Δ	+0.087	+0.087	+0.104	+0.083
Our model-Δ (P)	+0.051	+0.066	+0.091	+0.063
Our model-Δ (N)	+0.058	+0.031	+0.020	+0.023

- Δ: The performance gain by vocal imitation feedback
- Our model-Δ (P): only using **positive** vocal imitation feedback
- Our model-Δ (N): only using **negative** vocal imitation feedback

This work was supported by NSF Grant 1617497

Scenerio-2: Query-by-Example

❑ Problem: Query example contains multiple sounds

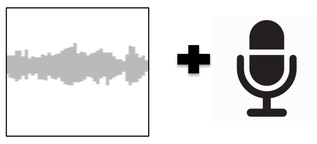


Query
(**a dog bark** + car engine)

The system does not know **which sound in the query is your target**.

Therefore, it matches the wrong sound.

❑ Solution: Provide a vocal imitation of the target or non-target sound



Query
(**a dog bark** + car engine)

Imitating the target sound:
"High-ranked items should sound like this!"

Imitating non-target sounds:
"High-ranked items should NOT sound like this!"

The updated similarity:

$$S(q,x) = C(q,x) - \frac{1}{N_{vn}} \sum_{i=1}^{N_{vn}} C(v_n^i, x) + \frac{1}{N_{vp}} \sum_{i=1}^{N_{vp}} C(v_p^i, x)$$

- q: the original query (mixed)
- X: a recording in the datasbase
- C(q, x): cosine similarity between q and x
- v^i_n: i-th vocal imitation of an irrelevant sound in the query (N_{vn} in total)
- v^i_p: i-th vocal imitation of the target sound in the query (N_{vp} in total)

❑ Experiment setting

- **Dataset:** subset of Vocal Imitation Set [3] with **2,683 isolated audio recordings**
- **301 *clean queries* (CQ) containing just one sound event**
- **301 *mixed queries* (MQ)** by mixing randomly selected pairs of clean queries.
- **Performance Metric:** mean recall @ k (k=10, 20, and 30)

❑ Result

Query type	MR@10	MR@20	MR@30
Clean Query (CQ)	0.256	0.356	0.412
Mixed Query (MQ)	0.128	0.172	0.211
MQ + random imitation	0.119	0.152	0.177
MQ + positive imitation	0.124	0.152	0.179
MQ + negative imitation	**0.147**	0.187	0.216
MQ + pos imit. + neg imit.	0.144	**0.192**	**0.225**

[1] M. Cartwright and B. Pardo, 2015 "Vocalsketch: Vocally imitating audio concepts,"
[2] Yichi Zhang and Zhiyao Duan. 2018. "Visualization and interpretation of Siamese style convolutional neural networks for sound search by vocal imitation"
[3] Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan, 2018. "Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology"
[4] Shawn Hershey. Et. al 2017. "Cnn architectures for large-scale audio classification"