VOCAL IMITATION SET: A DATASET OF VOCALLY IMITATED SOUND EVENTS USING THE AUDIOSET ONTOLOGY

Bongjun Kim[1], *Madhav Ghei*[2], *Bryan Pardo*[3], *Zhiyao Duan*[4]

[1] [2] [3] Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA,
[1]bongjun@u.northwestern.edu, [2]madhavghei2018@u.northwestern.edu, [3]pardo@northwestern.edu
[4] Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA,
zhiyao.duan@rochester.edu

ABSTRACT

Query-By-Vocal Imitation (QBV) search systems enable searching a collection of audio files using a vocal imitation as a query. This can be useful when sounds do not have commonly agreed-upon text-labels, or many sounds share a label. As deep learning approaches have been successfully applied to QBV systems, datasets to build models have become more important. We present Vocal Imitation Set, a new vocal imitation dataset containing $11,242$ crowd-sourced vocal imitations of 302 sound event classes in the AudioSet sound event ontology. It is the largest publicly-available dataset of vocal imitations as well as the first to adopt the widely-used AudioSet ontology for a vocal imitation dataset. Each imitation recording in Vocal Imitation Set was rated by a human listener on how similar the imitation is to the recording it was an imitation of. Vocal Imitation Set also has an average of 10 different original recordings per sound class. Since each sound class has about 19 listener-vetted imitations and 10 original sound files, the data set is suited for training models to do fine-grained vocal imitation-based search within sound classes. We provide an example of using the dataset to measure how well the existing state-of-the-art in QBV search performs on fine-grained search.

***Index Terms*—** Vocal imitation datasets, audio retrieval, query-by-vocal imitation search

1. INTRODUCTION

Imitating sounds with one's voice is a natural and effective way of delivering an audio concept in human-to-human communication. It can be even more effective than describing sound with words, when it is not clear how to describe the sound using words [1, 2]. This communication is possible because a human listener can identify what the imitation represents. If a machine can understand a human's vocal imitation, users can interact with the machine in this natural way for various audio-related tasks, such as sound designing [3, 4, 5, 6], or searching for melody in a music database by humming the desired melody [7].

Vocal imitations have recently gotten attention as a query method for general sound event search [8, 9, 10]. Commercially-deployed sound search and retrieval systems for general audio (e.g., Soundcloud [1], Freesound [2]) rely on text-based search. Text search fails when there is no search-relevant metadata about the audio content in the file. Text search may also be insufficient when one wants to retrieve a sound that does not have a commonly agreed-upon label or has a label unknown to the user (e.g., a new synthesizer sound). Search using a text label also often produces too many examples (e.g., "dog bark" producing over 1000 examples of dogs barking) and does not provide the specificity required (e.g., the particular bark of a frightened beagle). Using a vocal imitation as the search query, known as Query by Vocal Imitation (QBV), the user can provide information about the desired audio in a way complimentary to text querying.

QBV systems compare the vocal imitation to the content of each audio file in a collection. As deep neural networks have become a typical approach to sound classification tasks [11], they also have been successfully applied to QBV [10, 12]. However, while researchers have put significant effort into developing datasets for various sound classification tasks such as the DCASE dataset [13], the Urban Sound dataset [14], and AudioSet [15], developing datasets for QBV systems has had less attention. This is probably because collecting vocal imitation datasets requires much more human effort. Mehrabi et.al [16] created a dataset of 420 vocal imitations of 30 drum samples, which is useful for a musician to search for drum sounds, but not broad enough in coverage to train general-purpose QBV retrieval systems.

Cartwright and Pardo [17] created the VocalSketch dataset which covers more varieties of sound classes. It includes 240 reference recordings in 4 broad groups: Acoustic Instruments (40), Commercial Synthesizers (40), Everyday Sound (120), and Single Synthesizer (40). Each reference recording has about 10 imitations collected through Amazon Mechanical Turk where the participants were asked to listen to reference recordings (e.g., sound of dog barking) and imitate them vocally. Although the dataset has been successfully used to build QBV retrieval systems [10, 18, 12], it has only $2,400$ vocal imitations made in direct response to an audio file. This is much smaller than other environmental sound datasets and might be insufficient for training a deep model, which often contains many more parameters than the number of vocal imitations in this dataset. VocalSketch dataset also contains only one reference audio file per sound class (e.g. just one "dog barking" file). This means that systems trained on VocalSketch dataset can only learn coarse-grained distinctions between broad classes of sound ("dog barking" vs "violin"), as opposed to fine-grained within-class search (the right dog bark from a set of many dog barks).

In this work, we introduce *Vocal Imitation Set*, a new crowd-sourced vocal imitation dataset. Vocal Imitation Set has more than double the number of imitations available in VocalSketch dataset.

This work was supported by NSF Grants 1617497 and 1617107.

[1] `https://soundcloud.com/`

[2] `https://freesound.org/`

Vocal Imitation Set has $11,242$ recordings consisting of $5,601$ high-quality imitations that passed our inclusion criteria, as well as an additional $5,641$ draft, training, and excluded imitations. Vocal Imitation Set is the first dataset of vocal imitations that uses a widely-used ontology. Its sound classes were selected from the AudioSet ontology [15]. Each high-quality imitation was also rated by a human evaluator for perceptual similarity to the audio file it was an imitation of. Perceptual similarity ratings could be very useful in building and testing QBV retrieval systems. Lastly, Vocal Imitation Set has a mean of 10 different original recordings per sound class (e.g., 10 distinct police siren recordings). This will enable the development of fine-grained vocal imitation-based search algorithms, which are more useful when a database has multiple sound events with the same text tags.

2. DATA COLLECTION

2.1. Reference audio collection

Since our goal is to create a vocal imitation dataset that can be used to build a general-purpose QBV search system, the set of sound classes should cover a wide range of sound events. Therefore, we selected sound classes from the AudioSet ontology [15]. This ontology contains 632 sound classes that are structured hierarchically with a maximum depth of 6 levels. The top-level categories include *Animal sounds*, *Channel/environment/background sounds*, *Human sounds*, *Music*, *Natural sounds*, *Sounds of things*, and *Source-ambiguous sounds*. The sound classes in the AudioSet ontology were manually curated to represent a broad set of audio events one might encounter in real-world recordings and each class is assumed to be distinguishable from other classes based on sound alone without any additional information (e.g., visual cue or details of context). For each sound class, AudioSet provides links to YouTube videos that were tagged with the text label for that class. The audio tracks from these videos typically contains multiple, overlapping sounds. Perhaps for this reason, audio from these YouTube videos has been widely used as a benchmark dataset for sound event detection and scene classification [19, 20]. For more details about AudioSet, refer to [15].

The AudioSet ontology contains many sound classes that cannot be readily imitated vocally, such as *guitar amplifier* and labels related to music genres. After excluding these classes, 302 sound classes from the AudioSet ontology remained. AudioSet's actual audio typically contains scenes with multiple sounds, rather than isolated sounds. Since the goal of our data set is to provide clear pairings of vocal imitations to reference sounds, this makes AudioSet's audio sub-optimal. Therefore, we collected our sounds from a repository where contributors typically provide isolated, single-sound recordings. For each of the 302 selected sound classes, we collected an average of 10 audio recordings from *Freesound* using the class name as the search key. All files were truncated to a maximum of 20 seconds and encoded in the WAV format with a sample rate of either 44.1 kHz or 48 kHz.

A single high-quality recording was selected from the collected recordings for each class as a reference recording to be imitated by crowd-workers. Each reference audio file was confirmed to contain a clean sound event for the selected sound class and no other sound events. The other recordings that were not used for imitation collection are also included in the released dataset. Although they do not have associated vocal imitations, we expect that they will be useful for developing and evaluating fine-grained search algorithms (i.e., searching among sounds within the same class). We will show an example of using the recordings for fine-grained search in Section 4.

2.2. Vocal imitation collection

We collected vocal imitations from crowd-workers through Amazon Mechanical Turk using the VocalSketch interface and protocol presented in [17]. Imitators were asked to listen to a reference recording (i.e., one of the 302 collected reference recordings) and imitate the sound. Once they recorded their imitations, they were required to listen to their imitations to compare them with the reference recording. They were allowed to re-record their vocal imitations unlimited times before submitting the final one. Discarded imitations were saved as *draft recordings* in the released dataset. Finally, each imitator was asked how satisfied they were with their imitations using a 7 level scale. In each session, imitators were given five reference recordings (one recording from each class) to imitate. Imitators were paid \$0.80 per session. The first imitation of each imitator in a new session was saved as a *training recording*.

We collected a total of $11,242$ recordings from 455 unique people. There were $6,115$ final-submission vocal imitations, $4,444$ draft recordings and 683 training recordings. The $6,115$ final submission vocal limitations resulted in an average of roughly 20 imitations for each of the 302 reference recordings. We focused on this set of final submissions in our quality assessments.

3. QUALITY ASSESSMENT

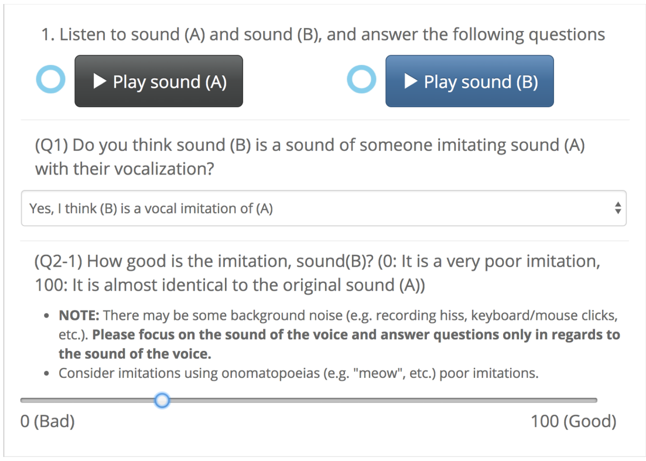
Crowd-sourced data collection suffers from noisy data in many cases. Therefore, we conducted an internal quality assessment of the $6,115$ final submissions, where experts evaluated the quality of all the final collected imitations. Training and draft vocal recordings were not evaluated. The purposes of the quality assessment are the following: 1) removing non-identifiable vocal imitations from the data set, and 2) measuring perceptual similarity between a reference recording and its imitations. The people who performed quality assessment were experts in audio processing: students and researchers from the Interactive Audio Lab [3] at Northwestern University and the Audio Information Research Lab [4] at the University of Rochester. There were, in total 15 evaluators, who listened to $6,115$ vocal imitations on a web interface designed for this particular listening task.

Figure 1 shows the web interface for our quality assessment. A single session consists of listening to a pair of recordings: one reference and one vocal imitation (Sound A and Sound B in Figure 1). An evaluator was first asked if the imitation was a vocal imitation of the reference recording. If the answer was "YES", then the evaluator was asked to assess the quality of the imitation on a scale from 0 to 100 (0: a very poor imitation; 100: almost identical to the reference sound). If the answer was "NO", then the recording was not evaluated for quality and it was placed in the *excluded* directory of the released dataset. The evaluator was then asked if the recording was a vocal imitation at all and this answer was saved.

Due to the size of the dataset, each imitation was evaluated by a single person. To measure consistency and reliability of each evaluator, we designed the task in following ways. First, an average of 2 out of every 30 pairs evaluated by an individual were incorrect pairs, where we paired an imitation with a reference recording that

[3] http://music.cs.northwestern.edu/

[4] http://www.ece.rochester.edu/projects/air/



1. Listen to sound (A) and sound (B), and answer the following questions

(Q1) Do you think sound (B) is a sound of someone imitating sound (A) with their vocalization?

Yes, I think (B) is a vocal imitation of (A)

(Q2-1) How good is the imitation, sound(B)? (0: It is a very poor imitation, 100: It is almost identical to the original sound (A))

- **NOTE:** There may be some background noise (e.g. recording hiss, keyboard/mouse clicks, etc.). **Please focus on the sound of the voice and answer questions only in regards to the sound of the voice.**
- Consider imitations using onomatopoeias (e.g. "meow", etc.) poor imitations.

0 (Bad) 100 (Good)

Figure 1: A screenshot of the interface for the internal quality assessment

it was not an imitation of. This let us measure how reliably evaluators were able to detect incorrect pairs. Second, an average of 4 out of every 30 pairs presented to an evaluator were repeated pairs, previously presented within the current batch (30 pairs). This let us measure the evaluation consistency for each evaluator.

In total, 452 incorrect pairs were presented to evaluators and 80% of them (363 pairs) were successfully identified as incorrect pairs. The remaining 20% (89 pairs) were incorrectly called correct pairs and they were given an average quality rating of 31.4 out of 100. The mean quality rating across all imitations is 60.3. This indicates that most evaluators correctly identified wrong pairs or gave them low scores if they called them a correct pair. Figure 2 shows how consistently evaluators rated repeated pairs. In total, 978 unique pairs of reference and imitation recordings were repeated. We computed the maximum difference of the multiple ratings to each of the 978 repeated pairs. For example, if a pair of reference and imitation recording was repeatedly rated three times by an evaluator and the ratings were 50, 60 and 70, then the maximum difference is 20 (70-50). As shown in Figure 2, the maximum differences of a majority of repeated pairs is very low (Mean: 7.63, SD: 10.96), which indicates that our evaluators rated vocal imitations with high consistency.

When collecting imitations, imitators were asked how satisfied they were with their own imitation using a 7 level scale. (1 - *completely dissatisfied,* 2 - *mostly dissatisfied,* 3 - *somewhat dissatisfied,* 4 - *neither satisfied or dissatisfied,* 5 - *somewhat satisfied,* 6 - *mostly satisfied,* 7 - *completely satisfied*). Figure 3 shows how the evaluator's ratings change with different self-satisfaction levels from imitators. There is a positive correlation between the imitators' self-satisfaction levels and evaluators' quality assessment scores. Yet, there are some imitations where the imitator's self-satisfaction disagrees with the quality reported by an evaluator. It would be interesting future work to learn the reason for the dichotomy.

Evaluators reported that 514 vocal imitations were not vocal imitation of the reference sound played to the imitator who made the imitation. These recordings were placed in the *excluded* directory of the released dataset. This left 5,601 recordings that have quality ratings, which are saved in the *included* directory of the dataset. We included all the quality rating on these 5,601 recordings in the released dataset.

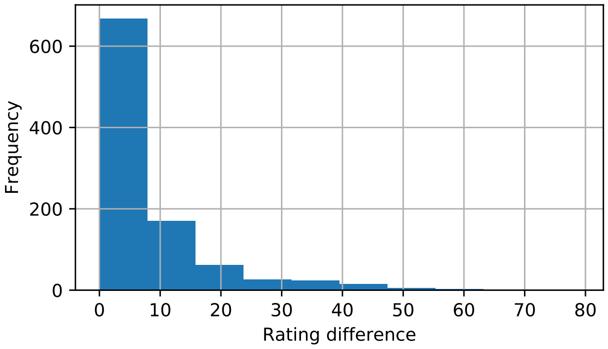


Figure 2: Histogram of maximum differences of quality ratings on a 100 point scale between two presentations of the same pairing of reference and imitation recording (Mean: 7.63, SD: 10.96)

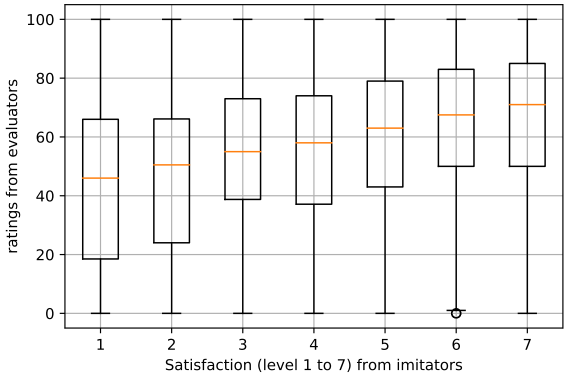


Figure 3: Relationship between self-satisfaction scores by imitators and quality assessment by evaluators.

4. BASELINE FINE-GRAINED SEARCH RESULTS

One expected use of Vocal Imitation Set is to train and test systems on fine-grained search. To this end, we provide an example of using this data set to measure how well the existing state-of-the-art in QBV search performs on fine-grained search. We used vocal imitations that were vetted by listeners (5,601 imitations of 302 classes). Each class contains one reference recording that was imitated and an average of 9 sound recordings that were not imitated. Each reference recording has an average of 18.6 imitations. Each time, we took one vocal imitation as the *query* to search for its reference recording (*target*) within all sound recordings of its class. An output of a search engine is an ordering of these sound recordings within each class, from most similar to the query to least similar.

To measure search quality, we computed Reciprocal Rank (RR), which is calculated as $1/r$ where r is the rank of the target. For instance, if the target ranks the third, then the RR is $1/3$. We measured *Mean Reciprocal Rank (MRR)*, which is the mean of RRs across all the queries (i.e., vocal imitations). We also computed a *Mean Recall@k* metric which indicates the proportion of queries that successfully retrieved the target within top k items in search results. For example, if only 50% of queries retrieved the target recording within top 2 items, then *Mean Recall*@2 is 0.5.

We used TL-IMINET [12], which is the best system we are aware of for coarse-grained QBV retrieval. TL-IMINET is a Siamese-style neural network with two Convolutional Neural Networks (CNN) towers: one tower for vocal imitations and the other for reference recordings. The imitation tower was pre-trained on a language classification task using sound clips from 7 different languages gathered from Voxforge [5]. The reference tower was pre-trained on a environmental sound classification task using sound clips from the 10 different classes in UrbanSound8K [14]. Then, TL-IMINET was trained on positive and negative pairs of reference sounds and imitations from the VocalSketch dataset [17]. In this training, negative pairs were always from an entirely different sound class (e.g., an imitation of a dog bark with a reference recording of a door slamming), so these pairs can be considered coarse-grained.

We performed fine-grained search with the trained model as follows. The trained model outputs the similarity between two input recordings. An imitation as a query is compared to each audio file within the sound class of that imitation's canonical reference recording (i.e., target). Since TL-IMINET takes only four seconds of audio as an input, each audio file is segmented into windows of length four seconds, with 50% overlap between each window, which gives us segment-level cross-similarities between the two recordings. To obtain the recording-level similarity between the reference and query file, we took the maximum similarity between any two segments in the two recordings. Based on the similarities, the rank of the target within the class is determined. By running this search using every vetted vocal imitation $(5,601$ in total) as a query, we compute MRR as well as Mean Recall@k covering all classes and the variety of queries within each class.

TL-IMINET gave a MRR of 0.356 for within-class search. Mean Recall@1 was 0.151 and Mean Recall@2 was 0.278. The class with the best MRR was *"Water stream"* with MRR of 1.0 and the worst MRR was for *"Bird's chirp, tweet"* with a MRR of 0.105. Since the mean number of recordings per class is roughly 10, chance ranking of the target is 5.5 which leads to chance MRR of $(1/5.5) = 0.18$. The results show that the state-of-the-art system which were designed for coarse-grained search performs much better than chance. However, the score is still similar or lower than scores from coarse-grained search performed in [12]. They achieved a MRR of about 0.4 in searches through 20 recordings and a MRR of 0.246 in searches with 60 recordings. This comparison shows challenges of fine-grained search. We believe that Vocal Imitation Set will enable researchers to build and test new models for fine-grained QBV search.

5. VOCAL IMITATION SET

Vocal Imitation Set is now publicly available[6]. It includes $2,985$ original recordings of 302 classes (an average of 9.89 per class) and $11,242$ vocal imitations of 302 reference recordings selected from the set of original recordings (1 reference recording per class). The set of vocal imitations consists of $5,601$ imitations that passed the quality assessment as well as $5,642$ recordings of draft, training recordings, and imitations excluded during the quality assessment. Table 1 shows the number of classes, listener-vetted imitations (i.e., imitations that have quality ratings), and original recordings for each top-level classes of AudioSet ontology. Figure 4 shows a histogram of quality assessment ratings of the $5,601$ listener-vetted imitations. The collected ratings give researchers another opportunity to build more robust vocal imitation-based interaction systems by using human quality assessments as a training signal.

Table 1: The number of classes, listener-vetted imitations, and original recordings (including reference recordings) for each of the first-level categories in Vocal Imitation Set

Categories	Classes	Imitations	Original Rec.
Animal	31	587	308
Channel, environment and background	4	74	40
Human sounds	38	714	375
Music	65	1247	646
Natural sounds	10	177	100
Sounds of things	134	2448	1316
Source-ambiguous sounds	20	354	200
Total	302	5601	2985

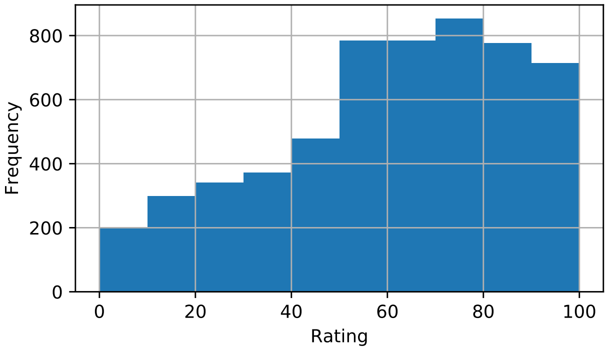


Figure 4: Histogram of quality assessment ratings to $5,601$ vocal imitations that were vetted by evaluators (Mean: 60.3, SD: 25.3)

6. CONCLUSIONS

We introduced *Vocal Imitation Set*, a new dataset of vocal imitations. It contains $11,242$ vocal imitations of 302 sound event classes which were curated based on AudioSet ontology. Sound recordings of the 302 classes were collected from Freesound and their imitations were collected by crowd-sourcing methods. We performed an internal quality assessment to filter out noisy data as well as to measure the perceptual similarity between an imitation and its reference recording. We also showed an example of using the dataset for fine-grained QBV search. We expect that this dataset will help the research community obtain a better understanding of human vocal imitations and build systems that can understand imitations as humans do.

7. REFERENCES

[1] G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of*

[5] http://www.voxforge.org

[6] http://doi.org/10.5281/zenodo.1340763

the Acoustical Society of America, vol. 135, no. 2, pp. 862–873, 2014.

[2] G. Lemaitre, O. Houix, F. Voisin, N. Misdariis, and P. Susini, "Vocal imitations of non-vocal sounds," *PloS one*, vol. 11, no. 12, p. e0168167, 2016.

[3] D. Rocchesso, D. A. Mauro, and C. Drioli, "Organizing a sonic space through vocal imitations," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 474–483, 2016.

[4] O. Houix, S. D. Monache, H. Lachambre, F. Bevilacqua, D. Rocchesso, and G. Lemaitre, "Innovative tools for sound sketching combining vocalizations and gestures," in *Proceedings of the Audio Mostly 2016*. ACM, 2016, pp. 12–19.

[5] M. Cartwright and B. Pardo, "Synthassist: an audio synthesizer programmed with vocal imitation," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 741–742.

[6] D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, and P. Boussard, "Sketching sound with voice and gesture," *interactions*, vol. 22, no. 1, pp. 38–41, 2015.

[7] A. Huq, M. Cartwright, and B. Pardo, "Crowdsourcing a real-world on-line query by humming system," in *Proceedings of the Sixth Sound and Music Computing Conference (SMC 2010)*, 2010.

[8] D. S. Blancas and J. Janer, "Sound retrieval from voice imitation queries in collaborative databases," in *AES 53rd Conference on Semantic Audio*, Audio Engineering Society. London, UK: Audio Engineering Society, 27/01/2014 2014.

[9] G. Roma and X. Serra, "Querying freesound with a microphone," in *Proceedings of the First Web Audio Conference (Ircam, Paris, France), submission*, vol. 39, 2015.

[10] Y. Zhang and Z. Duan, "Supervised and unsupervised sound retrieval by vocal imitation," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 533–543, 2016.

[11] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[12] Y. Zhang and Z. Duan, "Visualization and interpretation of siamese style convolutional neural networks for sound search by vocal imitation," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[13] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.

[14] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[15] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[16] A. Mehrabi, S. Dixon, M. Sandler, *et al.*, "Towards a comprehensive dataset of vocal imitations of drum sounds," in *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, 2016.

[17] M. Cartwright and B. Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 43–46.

[18] Y. Zhang and Z. Duan, "IMINET: Convolutional semi-siamese networks for sound search by vocal imitation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 304–308.

[19] A. Kumar, M. Khadkevich, and C. Fugen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[20] Q. Kong, Y. Xu, W. Wang, and M. Plumbley, "Audio set classification with attention model: a probabilistic perspective," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.