VocalSound: A Dataset for Improving Human Vocal Sound Recognition

4584

Yuan Gong[1], Jin Yu[2], James Glass[1] *([1]MIT CSAIL & [2]Signify Research)*

Dataset and baseline code at: groups.csail.mit.edu/sls/downloads/vocalsound

Introduction

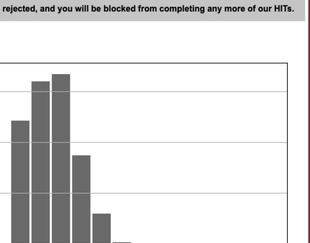
Recognizing human ***non-speech vocalizations*** is an important task and has broad applications such as automatic sound transcription and health condition monitoring.

Problems:

A) Existing datasets have a relatively small number of vocal sound samples.

	ESC-50	FSD50K	AudioSet	VocalSound
Laughter	40	1,186	5,696	3,504
Sigh	-	136	301	3,504
Cough	40	385	871	3,504
Throat Clearing	-	-	355	3,504
Sneeze	40	125	1,200	3,504
Sniff	-	-	205	3,504
Others	1,880	49.4K	2M	0
Vocal Sound Total	120	1,832	8,628	21,024

B) Models trained with existing datasets do not perform well on vocal sound detection.



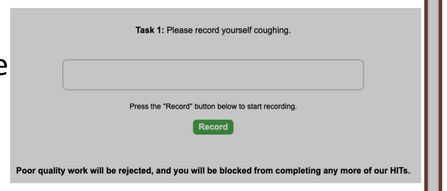
Class-wise performance of the PSLA audio tagging model trained on AudioSet. The model does not perform well on the cough/sneeze classes.

Proposed: ***VocalSound*** dataset consisting of over 21,000 crowdsourced recordings of laughter, sighs, coughs, throat clearing, sneezes, and sniffs from 3,365 unique subjects.

Dataset Collection

- Crowdsource the recordings via Amazon Mechanical Turk.
- Six vocal sounds *laughing, sighing, coughing, throat clearing, sneezing*, and *sniffing* along with meta information *gender, age, country, native language,* and *health information* are collected from each subject.

- We collected *3,504* HITs completed by *3,365* unique subjects.
- Only a small number of subjects recorded > 1 time
- Roughly gender balanced: 55% Male, 45% Female.
- Reasonable age distribution, range from 18-80.
- Data are collected globally, the majority group is US subjects.

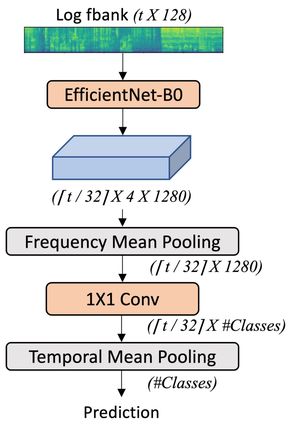


Dataset Quality

- We approve HITs according to the following three criteria*:
  1) Audio length is longer than 2 seconds.
  2) Audios should not be transcribed as words by Google Speech.
     - We manually check samples that are transcribed as words.
  3) Audios should match the corresponding class.
     - We use an automatic model with a relatively low threshold.
- We manually verified 600 samples from the dataset, with about **96%** judged to be high quality recordings.
- **Clean Evaluation Set**: We pay special attention to the evaluation set and manually checked one sample from each speaker and removed low-quality recordings.

*A small number of samples collected in the early stage are below the criteria.

Six-class Vocal Sound Classification



Test Set	Accuracy (%)
VocalSound Validation Set	90.1±0.2
VocalSound Evaluation Set	90.5±0.2
Different Age Group	
Age 18-25	91.5±0.3
Age 26-48	90.1±0.2
Age 49-80	90.9±1.6
Different Gender Group	
Male	89.2±0.5
Female	91.9±0.1

- We train an EfficientNet[1]-based model to classify the 6 vocal sound classes.
  - Overall, our baseline model achieves ~90% accuracy.
  - The vocal sound recognition accuracy differs for different age and gender groups.
    - The performance does not solely depend on the number of training samples.
    - The VocalSound dataset could also facilitate future research on reducing the machine bias among different subgroups of people.

[1]Tan et al., Rethinking Model Scaling for Convolutional Neural Networks, 2019

Combine VocalSound with Existing Data

Training Set	Laughter		Sigh		Cough		Sneeze		Background	
	F1	AP	F1	AP	F1	AP	F1	AP	F1	AP
FSD50K Only	0.45 ±0.04	0.46 ±0.05	0.31 ±0.01	0.28 ±0.02	0.41 ±0.04	0.35 ±0.02	0.61 ±0.02	0.57 ±0.07	0.97 ±0.00	0.99 ±0.00
FSD50K+ VocalSound	0.59 ±0.01	0.54 ±0.02	0.41 ±0.03	0.37 ±0.05	0.65 ±0.01	0.67 ±0.01	0.71 ±0.07	0.77 ±0.01	0.98 ±0.00	0.99 ±0.00
Improvement	29.7%	18.1%	30.5%	32.2%	58.6%	93.9%	16.0%	34.3%	1.5%	0.0%

- We compare models trained with
  1) FSD50K training set only
  2) FSD50K training set + VocalSound

  then evaluate the models on the FSD50K evaluation set for the vocal sound detection task.
- In addition to vocal sound classes, we include background sounds (from FSD50k)
  - A more realistic but also more challenging task because there are some background sounds that are similar to vocal sounds.
- Adding our VocalSound dataset to existing dataset can significantly improve the performance.

Data Usage and Availability

VocalSound dataset is ideal for the following usages:

- Build vocal sound classifier.
  Support research on removing model bias on various speaker groups.
- Evaluate pretrained model's performance on vocal sound classification.
- Combine with existing large-scale general audio datasets to further improve model performance.
- VocalSound dataset is **freely available** at: groups.csail.mit.edu/sls/downloads/vocalsound
- Six-class vocal sound classification baseline code is available at: github.com/yuangongnd/vocalsound
- Google Colab Script (No GPU needed) Open in Colab

Acknowledgement

This work is partly supported by Signify.