VOCALSOUND: A DATASET FOR IMPROVING HUMAN VOCAL SOUNDS RECOGNITION

Yuan Gong[1], Jin Yu[2], James Glass[1]

[1]MIT CSAIL, Cambridge, MA 02139, USA   [2]Signify Research, Cambridge, MA 02142, USA
{yuangong,glass}@mit.edu, jin.yu@signify.com

ABSTRACT

Recognizing human non-speech vocalizations is an important task and has broad applications such as automatic sound transcription and health condition monitoring. However, existing datasets have a relatively small number of vocal sound samples or noisy labels. As a consequence, state-of-the-art audio event classification models may not perform well in detecting human vocal sounds. To support research on building robust and accurate vocal sound recognition, we have created a *VocalSound* dataset consisting of over 21,000 crowdsourced recordings of laughter, sighs, coughs, throat clearing, sneezes, and sniffs from 3,365 unique subjects. Experiments show that the vocal sound recognition performance of a model can be significantly improved by 41.9% by adding VocalSound dataset to an existing dataset as training material. In addition, different from previous datasets, the VocalSound dataset contains meta information such as speaker age, gender, native language, country, and health condition.

Index Terms— vocal sounds, audio classification, corpus

1. INTRODUCTION

Automatic human vocal sound recognition is an important task and has a wide range of applications, e.g., it can help the automatic speech recognition system transcribe both speech and non-speech vocalizations. Recognizing health-related sounds like cough and sneeze could also provide insights into the general well-being of occupants in the office, at home, or other public or private spaces, e.g., the detection of coughing and sneezing and their density, intensity, and other features could be used as an indicator of group health [1, 2, 3].

To build an accurate and robust non-speech vocal sounds recognizer, a dataset with reasonable volume and variety, and accurate annotation is crucial. However, to our knowledge, currently, there is no such large-scale publicly available vocal sound dataset. Moreover, it has been found that a generic audio event classification model trained with existing datasets such as AudioSet [4] does not perform well in classifying human vocal sounds, e.g., the average precision of state-of-the-art models on cough and sneeze classes are only around 0.5 on the AudioSet evaluation set [5, 6]. Potential reasons include the fact that corpora such as ESC-50 [7],

FSD50K [8], and AudioSet [4] have a relatively small number of human vocal sound samples (summarized in Table 1) and the AudioSet annotation quality for these sounds may be lacking due to the challenge of annotating with a large sound vocabulary [6, 9, 10]. To address this limitation, in this paper we introduce the *VocalSound* dataset consisting of over 21,000 crowdsourced recordings of laughter, sighs, coughs, throat clearing, sneezes, and sniffs, collected via Amazon Mechanical Turk. The VocalSound dataset is class-balanced, collected from 3,365 speakers from 60 countries, with their ages ranging from 18 to 80. To the best of our knowledge, the VocalSound dataset has the largest number of human vocal sound samples. While one potential limitation of VocalSound dataset is the audio samples are not produced spontaneously but acted by the subjects, our experiments show that the model vocal sound recognition performance on an evaluation set consisting of real vocal sounds can be significantly improved by over 41.9% by adding VocalSound dataset to existing dataset as training material, demonstrating the usefulness of the VocalSound dataset. In addition, in contrast to previous datatsets [7, 8, 4], the VocalSound dataset contains meta information such as speaker age, gender, native language, country, and health condition to support research.

2. RELATED WORK

There are a few existing datasets for generic audio event classification that also contain human vocal sound samples such as AudioSet [4], FSD50K [8], ESC-50 [7], and DCASE [11]. Specifically, AudioSet is currently the largest publicly available dataset for generic audio classification consisting of over 2 million audio clips excised from YouTube and labeled with a set of 527 labels. The FSD50K dataset consists of 51,197 audio clips unequally distributed in 200 sound classes. While AudioSet and FSD50K consist of a large number of audio samples, they are class imbalanced and have a relatively small number of vocal sound samples (summarized in Table 1). Also, limited by the data acquisition scheme, they do not provide speaker information such as age, gender, native language, etc. Due to the difficulty of annotating YouTube videos with a large sound vocabulary, the noisy label issue has been found in AudioSet [6, 9, 10], which could also impact the performance of the model trained on it. Recently, there are a few efforts to collect cough samples for building COVID-19 classification models [12, 13, 14, 15, 16, 17, 18].

Dataset available at groups.csail.mit.edu/sls/downloads/vocalsound

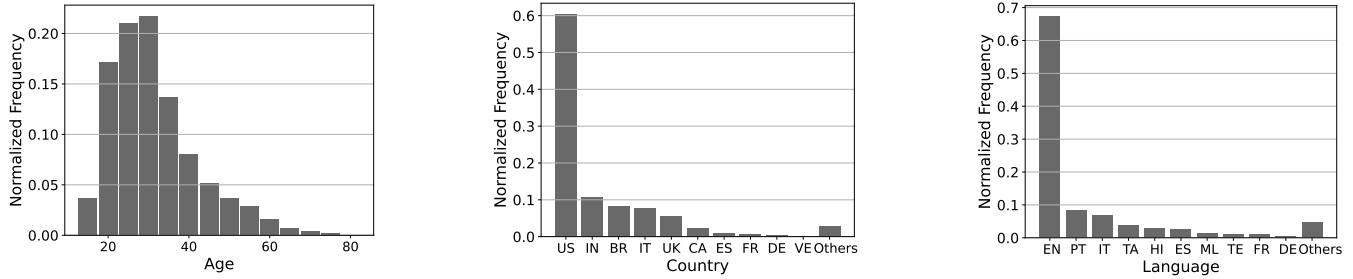**Fig. 1**. The speaker age (left), country (center), and native language (right) distribution.

	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

Table 1. Data volume comparison of VocalSound and existing datasets.

In comparison with the proposed dataset, existing imitated vocal sound datasets [19, 20] are much smaller in size.

The proposed VocalSound dataset differs from previous efforts in that 1) the VocalSound dataset is class-balanced and has more vocal sound samples collected from a large number of speakers with reasonable gender and age distributions. Due to the data acquisition scheme, the labels are also reliable. 2) the VocalSound dataset has rich meta information, including speaker gender, age, native language, country, and health condition, which broadens the application of the dataset, e.g., the metadata can be used to study the impact of gender, age, language on the performance of vocal sound classification models; the health label can potentially be used to build speech-based health classification system; the anonymous speaker label can potentially be used to build vocal sound-based speaker re-identification system, etc.
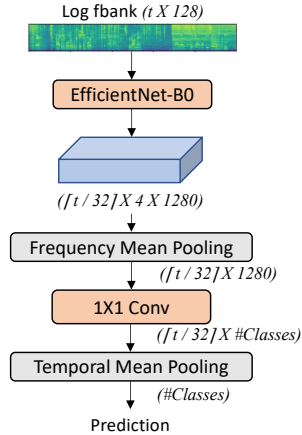
3. VOCALSOUND DATA COLLECTION

We crowdsource the VocalSound recordings via Amazon Mechanical Turk (AMT). Subjects volunteer to complete our Human Intelligence Tasks (HITs) on AMT and get compensation after the HITs are reviewed and approved by us. Our HIT consists of seven subtasks. First, we ask the gender, age, country, native language, and health information of the subject. For the health condition, we ask the question "do you have a cold, allergy, or other health-related symptoms that might affect your speech today?". Then in subtasks 2-7, we ask the subject to record themselves laughing, sighing, coughing, clearing their throat, sneezing, and sniffing. We do not collect personally identifiable information from the subject or the recording de-

vice, and the data collection is anonymous. We approve HITs according to the following three criteria: 1) the audio length is longer than 2 seconds; 2) we use Google Speech API to transcribe the audio to make sure no speech is contained, audios that can be transcribed as words (e.g., haha) are manually verified; 3) We use the model in [6] to verify if the audio matches the corresponding class. As the prediction of the model might not be accurate, we only use a low threshold to exclude obvious unrelated samples. We apply these criteria during the data collection process, to provide immediate feedback to the Turker and to improve the overall quality of the recordings. We manually verified 600 samples from the dataset, with about 96% judged to be high quality recordings.

4. DATA DISTRIBUTION

We collected 3,504 HITs completed by 3,365 unique subjects. Only a small number of subjects completed the HIT more than one time. Our goal was to encourage as much diversity across speakers as possible. Among the subjects, 45% are female, 55% are male. Therefore, the VocalSound dataset is roughly gender-balanced. We show the subject age, country, and native language distribution in Figure 1. The age of the subjects ranges from 18 to 80 while most subjects' ages fall between 20 to 40. Nevertheless, there are still 321 subjects that are older than 50, which are adequate for evaluating the model performance on the senior group. The subjects are from 60 countries, where the United States (60.3%), India (10.8%), and Brazil (8.3%) are the majority countries. English (67.2%), Portuguese (8.7%), and Italian (6.8%) are the corresponding dominant native languages of the subjects. 4.0% of the subjects report that they have health-related symptoms that might affect their speech during the data collection. The mean, median, and standard deviation of the audio length is 4.18s, 3.72s, and 1.81s, respectively. The audios are recorded at 44.1kHz in .wav format.

The data is split into training, validation, and evaluation sets with 15570 (74%), 1860 (9%), and 3594 (17%) samples, respectively. The three sets are speaker-independent. We pay special attention to the evaluation set and manually checked every sample of each speaker and removed low-quality recordings. This *clean* evaluation set makes the model evaluation fairer and more effective.

152

Log fbank *(t X 128)*

EfficientNet-B0

([t / 32] X 4 X 1280)

Frequency Mean Pooling

([t / 32] X 1280)

1X1 Conv

([t / 32] X #Classes)

Temporal Mean Pooling

(#Classes)

Prediction

Fig. 2. The model architecture used in our experiments. Each audio waveform is first converted to a sequence of 128-dimensional log Mel filterbank (fbank) features computed with a 25ms Hanning window every 10ms. The $t \times 128$ fbank feature vector is input to an EfficientNet-B0 model [21]. The EfficientNet-B0 model effectively downsamples the time and frequency dimensions by a factor of 32, and the feature dimension is 1280. Thus, the penultimate output of the model is a $\lceil t/32 \rceil \times 4 \times 1280$ tensor. We apply mean pooling over the 4 frequency dimensions to produce a 33×1408 representation that is fed to a set of 1×1 convolutional filters with a sigmoid activation function, where #class is the number of prediction classes. A temporal mean pooling is then performed to produce a final #class dimensional output for each class label.

5. BASELINE EXPERIMENTS

We conduct two baseline experiments using the proposed VocalSound dataset. First, in Section 5.1, we conduct a six-class (laughter, sigh, cough, throat clearing, sneeze, and sniff) classification experiment on the VocalSound dataset to show the model trained with VocalSound dataset can perform well on vocal sound classification. Second, in Section 5.2, we show the VocalSound dataset can help improve the vocal sound recognition from a wide variety of background sounds by combining it with the existing FSD50K dataset.

For both experiments, we use an EfficientNet-B0 [21] based audio classifier (illustrated in Figure 2), which has a similar architecture with the state-of-the-art audio classification model in [6], but uses EfficientNet-B0 and mean temporal pooling instead of EfficientNet-B2 and attention pooling. As discussed in [6], such simplification can greatly improve the computational efficiency while only marginally reducing performance. For both experiments, we train the model using an Adam optimizer [22], an initial learning rate of 1e-4, a batch size of 100, and cross-entropy loss for 50 epochs and select the best model using the development set and evaluate the model on the evaluation set. SpecAugment [23] is used during training. We repeat each experiment 3 times and report the mean and standard deviation of the results.

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

Table 2. Six-class Vocal Sound Classification Results.

5.1. Six-class Vocal Sound Classification

In this experiment, we train a six-class (laughter, sigh, cough, throat clearing, sneeze, and sniff) classifier using the Vocal-Sound dataset. We train, validate, and evaluate the model using the training, validation, and evaluation sets mentioned in Section 4. We downsample the sampling rate to 16kHz, and truncate or pad all audio samples to 5 seconds.

As shown in Table 2, the accuracy on the evaluation set is 90.5±0.2% (on the validation set: 90.1±0.2%), demonstrating the proposed VocalSound dataset can be used as training material to effectively train a vocal sound classifier. Interestingly, we find the classification accuracy varies with the speaker groups. As shown in Table 2, the model achieves an accuracy of 91.5±0.3%, 90.1±0.2%, 90.9±1.6% on the age group of 18-25, 26-48, 49-80, respectively; and 89.2±0.5% and 91.9±0.1% on male and female subjects, respectively. The performance does not solely depend on the number of training samples of each group as the age group of 26-48 and the male group have the largest number of samples but do not have the highest accuracy. Since the VocalSound dataset contains speaker meta information, it can be used to support future research on removing such model bias.

5.2. Vocal Sound Recognition from Background Sounds

While the model trained with just the VocalSound dataset achieves good accuracy on the 6-class vocal sound classification task, recognizing vocal sounds from a wide variety of background natural sounds is a more important and challenging task. Even the state-of-the-art audio classification models in [6, 5] cannot achieve satisfactory results for the vocal sound classes, e.g., the average precision on cough and sneeze classes are only around 0.5 on the AudioSet evaluation set. In this experiment, we show how the proposed VocalSound dataset can help improve the performance for this task. Specifically, we show that combining the VocalSound dataset with the existing FSD50K dataset as training material can noticeably improve vocal sounds recognition from background sounds compared with only using FSD50K as training material. The reason why we use FSD50K rather than

Training Set	Laughter		Sigh		Cough		Sneeze		Background		Vocal Classes Overall	
	F1	AP	F1	AP	F1	AP	F1	AP	F1	AP	Avg. F1	mAP
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	0.45 ±0.02	0.41 ±0.01
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	0.59 ±0.02	0.59 ±0.01
Improvement	29.7%	18.1%	30.5%	32.2%	58.6%	93.9%	16.0%	34.3%	1.5%	0.0%	31.8%	41.9%

Table 3. Vocal Sound Recognition Results on FSD50K Evaluation Set.

Fig. 3. Comparison of the confusion matrixes of the model trained with only FSD50K (left) and with FSD50K + Vocal-Sound (right), evaluated on FSD50K evaluation set. Results averaged from three runs and rounded to integer. Adding Vocal-Sound in the training set can improve the precision of the vocal sound classes (highlighted in bold numbers).

AudioSet as the base dataset is because FSD50K, especially its evaluation set, has more accurate labels [8] while labels of AudioSet are relatively noisy [6, 9, 10]. The FSD50K consists of 51K audio clips distributed in 200 sound classes so a wide variety of background sounds are included. Since FSD50K only contains 4 vocal sound classes, we consider a 4+1-class (laughter, sigh, cough, and sneeze + background class) classification problem. For the FSD50K dataset, we relabel all samples that are not labeled as laughter, sigh, cough, and sneeze as a new "background" class. FSD50K is a multi-label dataset but there are only 5, 1, and 13 samples having more than one vocal sound label in the training, validation, and evaluation set, respectively. We randomly select one label for these samples, making the task a single-class classification problem.

We compare two training set settings: 1) FSD50K only, we use the official 37k training split of FSD50K as the training set, among the 37k samples, only 1,241 samples are vocal sound samples and other samples are background sounds; 2) FSD50K + VocalSound, VocalSound dataset samples are combined with the FSD50K training set to form a new training set. It is worth mentioning that both datasets are severely class-imbalanced. The background class has 10× more samples than each vocal sound class even after VocalSound sam-

ples are added. Therefore, we use a balanced sampling strategy [6] to make the model see roughly the same number of samples of each class during training, specifically, we use the `torch.utils.data.WeightedRandomSampler` function. This also makes the comparison between the models trained with these two training sets fairer. In addition to balanced sampling, we also apply SpecAugment [23] and random time shift to alleviate the class-imbalance issue.

We train the EfficientNet models with the aforementioned two training sets with the same setting, validate the models using the FSD50K validation set, and evaluate the models using the FSD50K evaluation set. Note that we intentionally evaluate on FSD50K (real sounds, independent from the VocalSound dataset) rather than the VocalSound dataset itself to more fairly show the advantage of adding VocalSound for training. Since the evaluation set is also class-imbalanced, we report average precision (AP) and f1-score (F1) rather than accuracy. As shown in Table 3, training with FSD50K + VocalSound can significantly boost the vocal sound recognition performance by a relative f1-score improvement of 31.8% and an average precision improvement of 41.9%. In Figure 3, we compare the confusion matrix of the two models. We find that adding VocalSound in the training set can greatly improve the precision of the vocal sound classes. We run a McNemar's test and confirm the improvement is statistically significant ($p < 0.05$). All these demonstrate that the proposed Vocal-Sound dataset, while consisting of non-spontaneous sounds, can be used as training material to effectively improve the vocal sound classification performance in realistic use cases.

6. CONCLUSIONS

In this paper, we introduce VocalSound, a new dataset consisting of over 21,000 audio recordings of laughter, sighs, coughs, throat clearing, sneezes, and sniffs. Compared with existing generic audio event datasets, the proposed dataset has more vocal sound samples and richer speaker information. Our experiments show that the VocalSound dataset can noticeably improve vocal sound recognition performance. We hope the new dataset can contribute to future research on building accurate and robust vocal sound recognizers.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] L. Kvapilova, V. Boza, P. Dubec, M. Majernik, J. Bogar, J. Jamison, J. C. Goldsack, D. J. Kimmel, and D. R. Karlin, "Continuous sound collection using smartphones and machine learning to measure cough," *Digital Biomarkers*, 2019.

[2] N. Simou, N. Stefanakis, and P. Zervas, "A universal system for cough detection in domestic acoustic environments," in *European Signal Processing Conference*, 2021.

[3] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *International Conference on Ubiquitous Computing*, 2011.

[4] 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 *ICASSP*, 2017.

[5] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[6] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[7] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM Multimedia*, 2015.

[8] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *arXiv preprint arXiv:2010.00475*, 2020.

[9] E. Fonseca, S. Hershey, M. Plakal, D. P. Ellis, A. Jansen, and R. C. Moore, "Addressing missing labels in large-scale sound event recognition using a teacher-student framework with loss masking," *IEEE Signal Processing Letters*, 2020.

[10] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[11] 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 *DCASE*, 2017.

[12] L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *arXiv preprint arXiv:2009.11644*, 2020.

[13] J. Laguarta, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, 2020.

[14] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in *ACM KDD*, 2020.

[15] M. Cohen-McFarlane, R. Goubran, and F. Knoefel, "Novel coronavirus cough database: Nococoda," *IEEE Access*, 2020.

[16] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, 2020.

[17] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen *et al.*, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," in *Interspeech*, 2021.

[18] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough against covid: Evidence of covid-19 signature in cough sounds," *arXiv preprint arXiv:2009.08790*, 2020.

[19] B. Kim, M. Ghei, B. Pardo, and Z. Duan, "Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology," in *DCASE*, 2018.

[20] M. Cartwright and B. Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[21] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.