COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening

*first authors, †alphabetical order. Contact: {tx229@, covid-19-sounds@cl.cam.ac.uk

Highlights
Audio signal based COVID-19 screening is non-invasive, affordable, and promising.

To facilitate the advancement and openness of audio-based machine learning for respiratory health, we release a dataset consisting of 53,449 audio samples crowd-sourced from 36,116 participants through our COVID-19 Sounds app.

Crowd-sourced Data Collection
Early in the pandemic, we released apps for data collection, asking participants to report their COVID-19 test results along with other meta-data as well as record their breathing, cough, and voice.

After a year of data collection, we release a comprehensive dataset in terms of global reach, number of languages, broad demographic range, span of health conditions, and number of audio modalities.

Tasks and Results

- **Task 1 - Respiratory symptom prediction**: distinguish between healthy controls and generic respiratory abnormalities.
- **Task 2 - COVID-19 prediction**: distinguish between tested-negative controls and tested-positive patients.

Our results across both tasks show that using all modalities achieves the best performance through transfer learning, followed by the cough modality. These findings showcase the promise of audio-based models for respiratory prediction and the potential of detecting COVID-19 biomarkers (AUC>0.7).

Discussion
Based on the collected data, we have investigated a series of crucial research questions. Specifically, we have explored:
- The power of acoustic features from respiratory sounds as well as symptoms to detect COVID-19 [1,2].
- Transfer learning based end-to-end deep neural network to detect COVID-19 from limited labelled sounds [3];
- Leveraging ensemble learning to improve predictive confidence toward more robust COVID-19 detection [4];
- The impact of confounding factors on experimental results when utilizing sounds for COVID-19 screening [5];
- The potential of audio signals to capture COVID-19 progression [6].

Our Publications

Data Sample Overview
Sample data from a participant (a male aged over 30 with smoking history) who recently tested COVID-19 and displayed symptoms including wet cough, headache, and sore throat.

(a) Breathing recording  (b) Cough recording  (c) Voice recording

---

https://www.covid-19-sounds.org/en/ We share the data for research purpose. Contact: covid-19-sounds@cl.cam.ac.uk