Optimizing Audio Augmentations for Contrastive Learning of Health-Related Acoustic Signals

Louis Blankemeier BLANKEMEIER@GOOGLE.COM
Sebastien Baur SEBASTIENBAUR@GOOGLE.COM
Wei-Hung Weng CKBJIMMY@GOOGLE.COM
Jake Garrison JAKEGARRISON@GOOGLE.COM
Yossi Matias YOSSI@GOOGLE.COM
Shruthi Prabhakara SHRUTHIP@GOOGLE.COM
Diego Ardila ARDILA@GOOGLE.COM
Zaid Nabulsi ZNABULSI@GOOGLE.COM
Google Research, USA

Abstract

Health-related acoustic signals, such as cough and breathing sounds, are relevant for medical diagnosis and continuous health monitoring. Most existing machine learning approaches for health acoustics are trained and evaluated on specific tasks, limiting their generalizability across various healthcare applications. In this paper, we leverage a self-supervised learning framework, SimCLR with a Slowfast NFNet backbone, for contrastive learning of health acoustics. A crucial aspect of optimizing Slowfast NFNet for this application lies in identifying effective audio augmentations. We conduct an in-depth analysis of various audio augmentation strategies and demonstrate that an appropriate augmentation strategy enhances the performance of the Slowfast NFNet audio encoder across a diverse set of health acoustic tasks. Our findings reveal that when augmentations are combined, they can produce synergistic effects that exceed the benefits seen when each is applied individually.

Keywords: health acoustics, audio augmentation, contrastive learning

1. Introduction

Non-speech, non-semantic sounds, like coughing and breathing, can provide information for doctors to detect various respiratory diseases, cardiovascular diseases and neurological diseases (Boschi et al., 2017; Zimmer et al., 2022). Advances in deep learning-based machine learning (ML) allow us to develop medical assistants and continuous health monitoring applications by learning effective acoustic data representations (Alqudaihi et al., 2021).

Current approaches for learning health acoustic representations are mostly trained and evaluated on specific tasks. For example, Botha et al. (2018); Larson et al. (2012); Tracey et al. (2011); Pahar et al. (2021) trained models to detect tuberculosis using cough sounds via supervised learning. However, it can be challenging to adopt these models directly for other health acoustic tasks. Retraining task specific health acoustic models requires manual data collection and labeling by clinical experts, which can be time consuming and costly.

Researchers within the ML community have explored various self-supervised strategies to learn general purpose data representations that overcome the limitations of domain-specific representations (Balestriero et al., 2023). Among these approaches, contrastive learning has proven effective for generating robust representations across multiple data modalities, including images, videos, speech, audio, and periodic data (Chen et al., 2020a; Jiang et al., 2020; Qian et al., 2021; Oord et al., 2018; Yang et al., 2022). Selecting appropriate data augmentations is crucial for performant contrasting learning algorithms (Chen et al., 2020a) (see Related Works for details). Consequently, significant research has been conducted on the utility of various augmentations for images (Chen et al., 2020a), videos (Qian et al., 2021), and speech/audio (Al-Tahan and Mohsenzadeh, 2021; Jiang et al., 2020). However, the unique characteristics of health-related acoustic signals, such as coughs and breathing sounds, which differ in pitch

and tone from speech and music, raise questions about the applicability of existing contrastive learning and augmentation strategies in this specialized domain.

To address this research gap, our study systematically explores eight distinct audio augmentation techniques and their combinations in the context of health acoustic representation learning. We employ the self-supervised contrastive learning framework, SimCLR (Chen et al., 2020a), with a Slowfast NFNet backbone (Wang et al., 2022). After identifying the best combination of augmentations, we compare the performance of the resulting Slowfast NFNet against other state-of-the-art off-the-shelf audio encoders on 21 unique binary classification tasks across five datasets. This work offers two major contributions: (1) we identify augmentation parameters that work best when applied to health acoustics, and (2) we investigate the synergistic effects of combining audio augmentations for enhancing health acoustic representations using SimCLR.

2. Related Works

In ML, data augmentation serves as a regularization technique to mitigate the risk of model overfitting (Zhang et al., 2021). Within the framework of contrastive learning, the objective is to learn data representations that minimize the distance between representations of semantically similar inputs and maximize the distance between representations of semantically dissimilar inputs. Data augmentations are critical for contrastive learning-based self-supervised learning (SSL), and eliminates the need for labeled data for representation learning. By applying a variety of augmentations to a single input, semantically consistent but distinct variations, commonly referred to as views, are generated (Von Kügelgen et al., 2021). The task then becomes pulling these related views closer together in the representational space, while concurrently pushing views derived from different, unrelated inputs farther apart, via a contrastive loss, such as InfoNCE in SimCLR (Chen et al., 2020a). This approach establishes a form of invariance in the model, rendering it robust to the augmentations applied during the training process. Augmentations have been widely explored as part of contrastive learning-based SSL methods such as SimCLR, BYOL (Grill et al., 2020), MoCo (Chen et al., 2020b), and SwAV (Caron et al., 2020). Data augmentations also enhance the performance of SSL methods broadly across different data modalities, including images (Chen et al., 2020a), videos (Qian et al., 2021), audio (Al-Tahan and Mohsenzadeh, 2021; Niizumi et al., 2021), speech (Jiang et al., 2020), and 1-dimensional signals (e.g., human physiological signals) (Yang et al., 2022). In this study, we turn our attention toward a relatively underexplored domain: the application of data augmentations strategies for contrastive learning of health acoustic signals.

The most closely related area of research to our focus on health acoustics is the research investigating augmentation strategies for speech and audio data. Early research by Ko et al. (2015) explored creating two augmented speech signals with speeds relative to the original of 0.9 and 1.1. This yielded performance improvements across four speech recognition tasks. Jansen et al. (2018) expanded upon this by introducing a triplet loss for audio representation learning, incorporating random noise, time/frequency translation, example mixing, and temporal proximity augmentations. Jiang et al. (2020) employed an adaptation of SimCLR for speech data, termed Speech SimCLR, where they applied a diverse set of augmentations: random pitch shift, speed perturbation, room reverberation and additive noise to the original waveform, as well as time and frequency masking to the spectrogram. Niizumi et al. (2021) developed a comprehensive audio augmentation module including pre-normalization, foreground acoustic event mixup, random resize cropping and post-normalization. Fonseca et al. (2021c) investigated a multi-modal approach by adopting augmentations from both vision and audio domains, including random resized cropping, random time/frequency shifts, compression, SpecAugment (Park et al., 2019), Gaussian noise addition, and Gaussian blurring. They also used sound separation techniques for sound event detection to enable targeted data augmentations (Fonseca et al., 2021b). Shi et al. (2022) explored the impact of noise injection as an augmentation strategy to bolster the robustness of speech models. CLAR identified six augmentation operations: pitch shift, noise injection in frequency domain, and fade in/out, time masking, time shift, time stretching in the temporal domain, and explored their utility for audio contrastive learning (Al-Tahan and Mohsenzadeh, 2021). In this study, we build upon these ideas to systematically investigate the optimal combination and sequence of augmentation strategies, with a specific focus on developing robust representations for health acoustics.

3. Methods

The study is structured into three phases. The first phase consists of finding the best parameters for each augmentation that we consider for use with SimCLR. In the second phase, we investigate various combinations of augmentations, where we apply one or two successive augmentations to create each view of the input. Here, we use the augmentation parameters that we select in the first phase. In the third phase, we compare the results of our best performing model to other state-of-the-art audio encoder models on the validation set used for comparing augmentations. We choose to hold out the test sets due to ongoing model development and these results may thus be optimistic. This evaluation involves 21 unique downstream tasks across five datasets and we investigate the quality of embeddings generated from each audio encoder using linear probing (Köhn, 2015). Our study employs SimCLR with a 63 million parameter SlowFast NFNet-F0 as the neural network backbone (Chen et al., 2020a; Wang et al., 2022).

Audio Augmentations We investigate eight augmentations (Figure 1). These include the following time-domain augmentations: crop and pad, noising, Brownian tape speed (Weng et al., 2023), scaling, pitch shift, time stretch, and circular time shift. Additionally, we experiment with SpecAugment which is applied after the transformation of audio inputs into spectrograms (Park et al., 2019). A description of each augmentation strategy is provided in Appendix Table A1.

Each of the augmentations offers a tunable parameter space to allow for varying degrees of transformational intensity. To identify the optimal hyperparameters for each specific augmentation, we first conduct an exhaustive grid search. After we determine the best augmentation parameters, we explore the potential synergistic effects from the sequential application of either one or two successive augmentations. Since we include 8 augmentations, experimenting with every permutation of one or two augmentations would result in 64 experiments. However, in this work, SpecAugment was only applied after the time domain augmentations which reduced the number of 2-step augmentations to 57.

Datasets For this study, we curate a training dataset, YT-NS (YouTube Non-Semantic), consisting of two-second long audio clips extracted from one billion non-copyrighted YouTube videos, totalling about

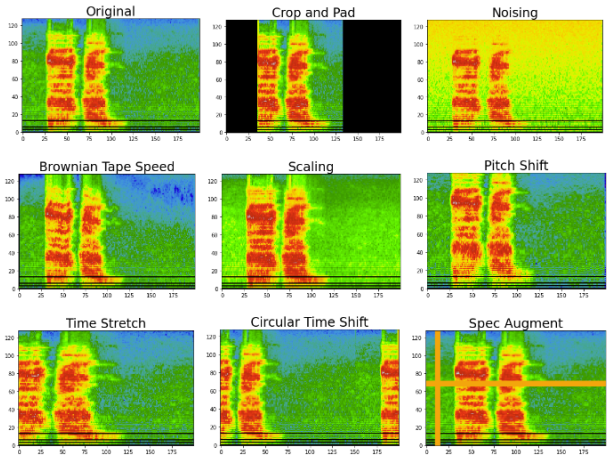

Figure 1: Mel spectrograms generated from various augmentations applied to the same health acoustic sample. One two-second example from the CoughVID dataset (Orlandic et al., 2021) is acquired and modified by each augmentation method.

255 million 2s clips or 142k hours. We apply a convolutional neural network-based health acoustic detector model, trained on two public health acoustic AudioSet derivatives, FSD50K and Flusense, as well as another health acoustic dataset. We use this model to filter two-second audio clips from these one billion videos for the following health acoustic signals: coughs, speech, laughing, throat-clearing, baby coughs, and breathing. Estimated numbers of each of these clips is provided in Appendix Table A2. The Slowfast NFNet encoder is trained solely using this dataset.

For evaluation, we use five publicly available datasets, FSD50K (Fonseca et al., 2021a), Flusense (Al Hossain et al., 2020), PSG (Korompili et al., 2021), CoughVID (Orlandic et al., 2021), and Coswara (Bhattacharya et al., 2023). We describe evaluation datasets in Appendix Table A3.

Evaluation 21 unique downstream binary classification tasks across five datasets are leveraged to evaluate the quality of health acoustic representations generated from the learned audio encoders, including 13 human acoustic event classifications, five sleep apnea-specific tasks, and three cough relevant tasks. The cough tasks include COVID detection, sex classification, and smoking status classification.

For phases 1 and 2 of our study where we identify the best parameters for each augmentation, as well as the best combination of augmentations, we develop a composite score that aggregates performance across the various downstream tasks. The PSG, CoughVid, and Coswara datasets are segmented into two-second clips. For Flusense, we preprocess the data by segmenting variable length clips using the labeled timestamps. For FSD50K and Flusense, we adopt a lightweight evaluation strategy where we randomly sample a single two second long clip from each longer clip. We take the average area under the receiver operating characteristic curve (AUROC) across these tasks and use this composite measure to rank augmentation strategies.

For phase 3, we segment the PSG data into 10 second clips, and for FSD50K and Flusense, we crop or zero pad each clip to 10 seconds. We adopt a sliding window approach for FSD50K, Flusense, and PSG, where embeddings are generated for two-second windows with a step size of one second. We apply mean pooling to the resulting embeddings to generate our final output embedding.

For all phases, we use linear probing to evaluate the quality of the generated representations. We use logistic regression with cross-validated ridge penalty, which is trained to predict binary labels from the frozen precomputed embeddings (Köhn, 2015). We report AUROC for all tasks and use the DeLong method to compute the 95% confidence intervals (CIs) (DeLong et al., 1988).

Baseline Models For comparative evaluation, we consider several off-the-shelf audio encoders, each trained on semantic or non-semantic speech data. Specifically, our baseline models include TRILL (Shor et al., 2020), which is a publicly available ResNet50 architecture trained on an AudioSet subset that is enriched with speech labels. FRILL (Peplinski et al., 2020) is a light-weight MobileNet-based encoder distilled from TRILL. BigSSL-CAP12 (Shor et al., 2022) leverages a Conformer-based architecture, trained on YouTube and LibriLight.

4. Results

Optimal augmentation parameters In Appendix Table A1, we display the optimal parameters for each augmentation derived from the associated grid searches. We find that up to a certain threshold, generally more intense augmentation parameters yield better performance.

Comparing augmentations Comparing the left and right panels of Figure 2 shows that many augmentations perform better in combination than individually. Our analysis indicates that the most effective single augmentation strategy is SpecAugment (left panel in Figure 2). The most effective 2-step augmentation strategy involves applying circular time shift , followed by time stretch, as depicted in Figure 2. Interestingly, circular time shift does not perform well on its own and each of these augmentations individually underperform SpecAugment. However, circular time shift and time stretch are synergistic when applied together. The right panel of Figure 2 shows that on average, time stretch is the most useful first augmentation, excluding SpecAugment which is always applied second or alone. SpecAugment is the most useful second augmentation on average.

Comparing to baselines Appendix Tables 4, 5 demonstrate performance of the best SimCLR model versus the baseline models on the validation set used for the comparison of augmentations. Overall, the performance of the SimCLR model is similar to BigSSL-CAP12, despite training on about 10x less hours of data and using a model that is nearly 10x smaller, and outperforms off-the-shelf audio encoders.

Figure 2: Evaluation performance for comparing augmentation combinations. (Left) from single augmentations. (Right) two augmentations applied where rows represent the first augmentation and columns represent the second augmentation.

5. Discussion and Conclusion

We investigated a comprehensive list of augmentations for use in the health acoustic domain. We demonstrated the synergistic benefit of the circular time shift and time stretch augmentations. Circular time shift and time-stretching may synergistically improve model generalizability by introducing a diverse range of temporal patterns for the same sound.

There are few limitations worth noting. We decided to keep our test sets held out for ongoing model development, thus our comparisons to baselines may be optimistic. We also confined our analysis to a single Slowfast NFNet architecture. This leaves open the possibility that different architectures could yield varying results. Future research may focus on other augmentations, including frequency domain augmentations, as well as augmentations that better leverage health acoustic inductive biases. Additionally, incorporating labels during training (Khosla et al., 2020), such as health signal type, may further improve the learned representations.

Acknowledgments

We thank Yun Liu from Google Research for his critical feedback, Shao-Po Ma for his preliminary work on the PSG dataset, CoughVID and Project Coswara teams for making the datasets publicly available, and the Google Research team for software and hardware infrastructure support. PSG, CoughVID and Coswara are licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License and follow the Disclaimer of Warranties and Limitation of Liability in the license.

References

Forsad Al Hossain, Andrew A Lover, George A Corey, Nicholas G Reich, and Tauhidur Rahman. Flusense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–28, 2020.

Haider Al-Tahan and Yalda Mohsenzadeh. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2530–2538. PMLR, 2021.

Kawther S Alqudaihi, Nida Aslam, Irfan Ullah Khan, Abdullah M Almuhaideb, Shikah J Alsunaidi, Nehad M Abdel Rahman Ibrahim, Fahd A Alhaidari, Fatema S Shaikh, Yasmine M Alsenbel, Dima M Alalharith, et al. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. *Ieee Access*, 9:102327–102344, 2021.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C Chandrakiran, Sahiti Nori, KK Suhail, Sadhana Gonuguntla, et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific Data*, 10(1):397, 2023.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269, 2017.

GHR Botha, Grant Theron, RM Warren, Marisa Klopper, Keertan Dheda, PD Van Helden, and TR Niesler. Detection of tuberculosis by automatic cough sound analysis. *Physiological measurement*, 39(4):045005, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021a.

Eduardo Fonseca, Aren Jansen, Daniel PW Ellis, Scott Wisdom, Marco Tagliasacchi, John R Hershey, Manoj Plakal, Shawn Hershey, R Channing Moore, and Xavier Serra. Self-supervised learning from automatically separated sound scenes. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 251–255. IEEE, 2021b.

Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E O'Connor, and Xavier Serra. Unsupervised contrastive learning of sound event representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 371–375. IEEE, 2021c.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous. Unsupervised learning of semantic audio representations. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 126–130. IEEE, 2018.

Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*, 2015.

Arne Köhn. What's in an embedding? analyzing word embeddings through multilingual evaluation. 2015.

Georgia Korompili, Anastasia Amfilochiou, Lampros Kokkalas, Stelios A Mitilineos, Nicolas-Alexander Tatlas, Marios Kouvaras, Emmanouil Kastanakis, Chrysoula Maniou, and Stelios M Potirakis. Psg-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Scientific data*, 8(1):197, 2021.

Sandra Larson, Germán Comina, Robert H Gilman, Brian H Tracey, Marjory Bravard, and José W López. Validation of an automated cough detection algorithm for tracking recovery of pulmonary tuberculosis patients. 2012.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, 2021.

Madhurananda Pahar, Marisa Klopper, Byron Reeve, Rob Warren, Grant Theron, and Thomas Niesler. Automatic cough classification for tuberculosis screening in a real-world environment. *Physiological Measurement*, 42(10):105014, 2021.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel. Frill: A non-semantic speech embedding for mobile devices. *arXiv preprint arXiv:2011.04609*, 2020.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*, 2020.

Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3169–3173. IEEE, 2022.

Brian H Tracey, Germán Comina, Sandra Larson, Marjory Bravard, José W López, and Robert H Gilman. Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis. In *2011 Annual international conference of the IEEE engineering in medicine and biology society*, pages 6017–6020. IEEE, 2011.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. Towards learning universal audio representations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4593–4597. IEEE, 2022.

Wei-Hung Weng, Sebastien Baur, Mayank Daswani, Christina Chen, Lauren Harrell, Sujay Kakarmath, Mariam Jabara, Babak Behsaz, Cory Y McLean, Yossi Matias, et al. Predicting cardiovascular disease risk using photoplethysmography and deep learning. *arXiv preprint arXiv:2305.05648*, 2023.

Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. *arXiv preprint arXiv:2210.03115*, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Alexandra J Zimmer, César Ugarte-Gil, Rahul Pathri, Puneet Dewan, Devan Jaganath, Adithya Cattamanchi, Madhukar Pai, and Simon Grandjean Lapierre. Making cough count in tuberculosis care. *Communications medicine*, 2(1):83, 2022.

Appendix A. SimCLR hyperparameters

For training, we use 32 TPU-v3 cores with a batchsize of 4096. We use an AdamW optimizer with default parameters and a learning rate of 1.6e-3. We train all models for at least 300k steps, saving checkpoints every 5k steps. We select checkpoints that exhibit the best performance on the validation data after applying an exponential moving average, with a bias correction and a weight of 0.5, to the validation curves.

Appendix B. Appendix Tables

Augmentation	Apply On	Description	Best Parameters	Grid Search (Cartesian product of the lists)
Crop and pad	Temporal	Crops the audio signal and then zero-pads to the input length.	Probability = 1.0 Min fraction = 0.1 Max fraction = 0.5	Probability = [0.8, 1.0] Min fraction = [0.1, 0.3, 0.5] Max fraction = [0.3, 0.5, 0.7] Only when max fraction >min fraction
Noising	Temporal	Adds gaussian noise to the audio signal.	Probability = 1.0 Mean = 0.2 Stddev = 0.2	Probability = [0.8, 1.0] Mean = [-0.2, 0.0, 0.2] Stddev = [0.2, 0.4, 0.6]
Brownian tape speed	Temporal	Simulates playing back the signal on a tape while the playback speed at each time step is drawn from a normal distribution.	Probability = 0.8 Magnitude = 20	Probability = [0.8, 1.0] Magnitude = [2, 10, 20]
Scaling	Temporal	Modifies the audio gain.	Probability = 0.8 Min factor = 0.25 Max factor = 1.75	Probability = [0.8, 1.0] Min factor = [0.25, 0.75, 1.25] Max factor = [0.75, 1.25, 1.75] Only when max factor >min factor
Pitch shift	Temporal	Moves the pitch of the audio up or down without changing its speed.	Probability = 0.8 Min factor = 1.25 Max factor = 1.75	Probability = [0.8, 1.0] Min factor = [0.25, 0.75, 1.25] Max factor = [0.75, 1.25, 1.75] Only when max factor >min factor
Time stretch	Temporal	Slows and speeds up the audio signal without changing its pitch.	Probability = 0.8 Min time stretch = 0.75 Max time stretch = 1.75	Probability = [0.8, 1.0] Min time = [0.25, 0.75, 1.25] Max stretch = [0.75, 1.25, 1.75] Only when max factor >min factor
Circular time shift	Temporal	Translates the audio signal temporally without truncating the signal, while wrapping along the time axis.	Probability = 1.0	Probability = [0.8, 1.0]
SpecAugment	Spectrogram	Applies masking to the temporal and frequency axes.	Probability = 1.0 Time mask max frames = 24 Time mask count = 20 Frequency mask max bins = 20 Frequency mask count = 5	Probability = [0.8, 1.0] Time mask max frames = [24, 36] Time mask count = [10, 20] Frequency mask max bins = [10, 20] Frequency mask count = [3, 5]

Table A1: Description of the augmentation strategies used in the study.

Class	Estimated # Audio Clips
Cough	77,000,000
Speech	65,100,000
Laughing	77,240,000
Throat clearing	3,300,000
Baby cough	800,000
Breathing	31,500,000

Table A2: Number of YouTube audio clips used for training.

Dataset	Tasks	Number of examples for training linear probes	Number of examples for evaluation	Reference
FSD50K	Health acoustic event (6 tasks)	32,652	8,313	Fonseca et al. (2021a)
Flusense	Health acoustic event (7 tasks)	7,537	1,779	Al Hossain et al. (2020)
PSG	Apnea, arousal events (5 tasks)	7,320	3,625	Korompili et al. (2021)
CoughVID	COVID, sex (2 tasks)	44,249	15,083	Orlandic et al. (2021)
Coswara	COVID, sex, smoking status (3 tasks)	10,230	4,285	Bhattacharya et al. (2023)

Table A3: Evaluation datasets statistics.

Dataset	Task	TRILL	FRILL	BigSSL-CAP12	SimCLR (ours)
FSD50K	Breathing	0.973 (0.958, 0.988)	0.974 (0.961, 0.987)	**0.983 (0.973, 0.993)**	0.982 (0.969, 0.995)
	Cough	0.988 (0.982, 0.994)	0.986 (0.979, 0.993)	0.998 (0.996, 1)	**0.999 (0.998, 1)**
	Laughter	0.985 (0.978, 0.992)	0.984 (0.977, 0.992)	**0.994 (0.988, 0.999)**	0.991 (0.983, 1)
	Sneeze	0.913 (0.757, 1)	0.960 (0.896, 1)	**0.997 (0.995, 1)**	0.988 (0.969, 1)
	Speech	0.970 (0.958, 0.982)	0.972 (0.962, 0.983)	**0.982 (0.974, 0.991)**	0.978 (0.967, 0.988)
	All Respiratory sounds	0.978 (0.972, 0.985)	0.979 (0.973, 0.985)	0.985 (0.977, 0.994)	**0.990 (0.984, 0.995)**
Flusense	Breathe	0.732 (0.602, 0.861)	0.741 (0.614, 0.869)	0.769 (0.636, 0.902)	**0.816 (0.706, 0.925)**
	Cough	0.656 (0.614, 0.698)	0.658 (0.616, 0.700)	0.675 (0.635, 0.716)	**0.703 (0.662, 0.743)**
	Gasp	0.731 (0.624, 0.837)	0.721 (0.618, 0.824)	0.766 (0.676, 0.855)	**0.777 (0.675, 0.880)**
	Sneeze	0.719 (0.663, 0.776)	0.717 (0.66, 0.773)	0.780 (0.731, 0.829)	**0.789 (0.740, 0.838)**
	Sniffle	0.734 (0.671, 0.798)	0.727 (0.662, 0.791)	0.717 (0.648, 0.787)	**0.762 (0.697, 0.827)**
	Speech	0.711 (0.670, 0.751)	0.701 (0.659, 0.742)	**0.764 (0.724, 0.804)**	0.751 (0.715, 0.788)
	Throat clearing	0.811 (0.692, 0.931)	0.756 (0.620, 0.891)	**0.914 (0.863, 0.964)**	0.788 (0.671, 0.905)
PSG	OSA	0.681 (0.643, 0.720)	0.697 (0.658, 0.736)	**0.770 (0.735, 0.806)**	0.700 (0.663, 0.738)
	Central	0.640 (0.441, 0.838)	0.695 (0.537, 0.852)	**0.725 (0.553, 0.896)**	0.690 (0.510, 0.870)
	Mixed	0.728 (0.658, 0.797)	0.732 (0.663, 0.800)	**0.788 (0.726, 0.850)**	0.719 (0.654, 0.783)
	Hypopnea	0.497 (0.445, 0.549)	0.537 (0.485, 0.588)	**0.639 (0.590, 0.688)**	0.549 (0.500, 0.597)
	Arousal	0.716 (0.674, 0.759)	0.732 (0.691, 0.772)	0.728 (0.686, 0.770)	**0.784 (0.746, 0.822)**

Table A4: Performance comparison (AUROC with 95% confidence intervals) on downstream tasks in FSD50K, Flusense and PSG datasets. OSA: obstructive sleep apnea.

Task	Dataset	TRILL	FRILL	BigSSL-CAP12	SimCLR (ours)
COVID	CoughVID	0.613 (0.592, 0.634)	0.611 (0.59, 0.632)	0.621 (0.6, 0.642)	**0.622 (0.601, 0.643)**
	Coswara	0.573 (0.54, 0.607)	0.591 (0.557, 0.625)	0.597 (0.565, 0.628)	**0.769 (0.752, 0.785)**
Smoker	Coswara	0.62 (0.589, 0.651)	0.579 (0.548, 0.609)	**0.624 (0.594, 0.654)**	0.591 (0.560, 0.621)
Sex	CoughVID	0.839 (0.831, 0.847)	0.83 (0.822, 0.838)	0.847 (0.839, 0.855)	**0.862 (0.854, 0.869)**
	Coswara	0.872 (0.861, 0.883)	0.827 (0.814, 0.84)	0.900 (0.890, 0.910)	**0.903 (0.893, 0.913)**

Table A5: Performance comparison (AUROC with 95% confidence intervals) on downstream tasks in CoughVID and Coswara datasets.