2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871897

Estimation of Respiratory Rate from Breathing Audio*

John Harvill[1], Yash Wani[2], Mustafa Alam[2], Narendra Ahuja[1], Mark Hasegawa-Johnson[1],
David Chestek[3], David G. Beiser[2]

Abstract— **The COVID-19 pandemic has fueled exponential growth in the adoption of remote delivery of primary, specialty, and urgent health care services. One major challenge is the lack of access to physical exam including accurate and inexpensive measurement of remote vital signs. Here we present a novel method for machine learning-based estimation of patient respiratory rate from audio. There exist non-learning methods but their accuracy is limited and work using machine learning known to us is either not directly useful or uses non-public datasets. We are aware of only one publicly available dataset which is small and which we use to evaluate our algorithm. However, to avoid the overfitting problem, we expand its effective size by proposing a new data augmentation method. Our algorithm uses the spectrogram representation and requires labels for breathing cycles, which are used to train a recurrent neural network for recognizing the cycles. Our augmentation method exploits the independence property of the most periodic frequency components of the spectrogram and permutes their order to create multiple signal representations. Our experiments show that our method almost halves the errors obtained by the existing (non-learning) methods.**

Clinical relevance—**We achieve a Mean Absolute Error (MAE) of 1.0 for the respiratory rate while relying only on an audio signal of a patient breathing. This signal can be collected from a smartphone such that physicians can automatically and reliably determine respiratory rate in a remote setting.**

I. Introduction

The COVID-19 pandemic has established the use of telemedicine as a critical health care delivery channel that is likely to expand in the future. A significant challenge faced in telemedicine care delivery is the accurate measurement of vital signs such as respiratory rate. Respiratory rate, defined as the number of breaths a person takes per minute, is one of four clinical vital signs. As such, it plays a central role in the physical examination and accurate diagnosis of patients. Changes in respiratory rate have been shown to be an important early indicator of clinical deterioration and increased mortality in a variety of disease states [1], [2]. Thus an accurate measurement of respiratory rate is critical for assessing patient stability. The gold standard for measuring respiratory rate is to count a patient's breaths over a 60 second interval. In the busy clinic or triage setting, this approach is inefficient and often abbreviated by observing breaths over shorter time intervals (e.g., 10 seconds) which can lead to inaccurate estimates [3], [4], [5]. Additionally, awareness of one measuring their own respiratory rate has been shown to change a patient's respiratory rate [6].

A variety of practical challenges for respiratory rate estimation are raised in the setting of a telemedicine visit due to poor lighting, low video quality, and camera angle which may hinder a practitioner's ability to manually assess a patient's respiratory rate and increase the potential for human error. Accordingly, there is an urgent need for a robust, low-cost method for estimating respiratory rate for the delivery of health care in busy and remote telemedicine settings.

One well-established automated approach for measuring respiratory rate in the monitored clinical setting is known as impedance pneumonography which measures changes in transthoracic impedance during the respiratory cycle via skin electrodes [7]. Yet this method requires expensive equipment that is only typically available in a monitored clinical setting such as an emergency department, intensive care unit and some general medical wards and thus does not lend itself to remote or resource-limited settings.

By contrast, an estimation system relying on audio signals alone provides an automatic, remote, and virtually free alternative for respiratory rate estimation that is amenable to hospital and telehealth settings. Existing work in this area involves both learning-based and purely signal processing-based techniques that have been used to estimate respiratory rate directly or related features like inspiration and expiration cycle boundaries which can be used for computation of respiratory rate [8], [9], [10], [11].

In this paper we make the following contributions: (1) Propose a novel breath cycle supervision technique that assigns a binary label to each acoustic frame. This allows learning-based methods to be used for respiratory rate estimation when labeled data is scarce. (2) Propose a novel input transformation that further mitigates overfitting in a low-data scenario by focusing on periodicity in the signal. (3) Create a partition of the only publicly-available labeled audio respiration dataset (ICBHI) that makes it useful for evaluation of respiratory rate estimation methods instead of its original purpose which is lung sounds classification. (4) Compare our proposed learning-based approach to existing signal processing-based techniques on our partition.

II. Related Work

Signal Processing-Based Methods: Multiple signal processing methods have been used to estimate respiratory

*This work was supported by C3.ai Digital Transformation Institute.

[1] John Harvill (student), Narendra Ahuja (Faculty), and Mark Hasegawa-Johnson (Faculty) belong to the Department of Electrical and Computer Engineering at University of Illinois Urbana-Champaign {harvill2, n-ahuja, jhasegaw}@illinois.edu.

[2] Yash Wani (student) Mustafa Alam (postgrad) and David Beiser (Faculty) belong to the Pritzker School of Medicine at the University of Chicago yashwani@uchicago.edu, {malam, dbeiser}@medicine.bsd.uchicago.edu.

[3] David Chestek (Faculty) belongs to the University of Illinois Chicago School of Medicine dchest2@uic.edu.

rate from audio alone. Dafna et al. implemented an autocorrelation-based approach which maximizes correlation between the original audio signal and the same signal delayed in time [8]. The autocorrelation is multiplied by a breathing interval probability function (BIPF) which serves to emphasize more physiologically-likely breathing intervals based on prior knowledge. Ren et al. extract the envelope of the audio signal by computing the maximum absolute value of samples within an audio frame [9]. The breathing rate is then estimated by examining the time at which a similarity function between the original audio and time-delayed audio is maximized.

Learning-Based Methods: A number of deep learning-based algorithms have also been created in order to estimate respiration rate. Nallanthighal et al. [10] collected audio data from subjects performing a number of tasks including spontaneous speech, reading a script and general speech. Then spectral features like spectrograms and log Mel spectrograms calculated from the audio signals were fed into CNN and LSTM models, using respiratory belt sensor values as ground truths. Jácome et al. [11] used the Faster R-CNN architecture to classify regions in input spectrograms into three regions: background, inspiration, and expiration. Unlike Nallanthighal et al. who used audio gathered from an external microphone, lung sounds gathered from a microphone-enabled stethoscope were used as input audio sources.

III. Data

Respiratory Sound Database ICBHI[1] 2017: The Respiratory Sound Database from ICBHI 2017 [12] was collected for classification of various breathing disorders. The dataset[2] consists of breathing samples with full inspiration/expiration intervals labeled. To the best of our knowledge, this is the only publicly-available dataset of breathing sounds. In the original dataset, there are many irregular breathing samples that have properties inconsistent with normal breathing data that the algorithms presented in this paper would expect as input, i.e. aperiodic. To have an accurate representation of performance on regular breathing data, which is what would be expected in most settings, we want to test our system on relatively periodic and clean[3] breathing samples only. We manually remove samples from the dataset for our purposes according to four general criteria: (1) samples with noticeable heart beats (2) samples with too many random background sounds (3) too much background noise or hum present in the audio sample (4) irregular or non-periodic breathing pattern. Of the original 914 audio samples, only 419 remained after manual removal according to the above criteria. We also trim audio segments from the second to the second-to-last interval label, because we find inconsistent labels at the beginning and end of many samples that do not correspond to full breathing cycles.

Preprocessing and Feature Extraction: We use the log Mel spectrogram as our feature representation for audio in all approaches except the envelope baseline (discussed in the following section). We downsample audio recordings to 16kHz and use a window of length 1024, a hop size of 10ms, and 80 Mel frequency bins. We clip all signal components below -120dB and normalize to the range 0 to 1 where 0 corresponds to -120dB and 1 corresponds to 0dB.

IV. Baselines

Autocorrelation approach: We adapted the autocorrelation-based approach from [8] to serve as a baseline. We compute the log Mel spectrogram as discussed previously, and then periodicities are calculated for a given frequency component i using the following equation:

$$P_{i,l} = \frac{1}{T \times FR - l} \sum_{n=0}^{T \times FR - l} (X_{i,n} - \bar{X}_i) \times (X_{i,n+l} - \bar{X}_i) \tag{1}$$

where T is the window's length in seconds, FR is the frame rate in Hertz, $\bar{X}_i$ is the mean of component i, and l is the time-lag in seconds. The periodicities are then smoothed with a low-pass filter. We choose the top 20 most periodic frequency components per breathing sample[4] and take the mean of these 20 components per frame to get the overall periodicity signal. We multiply the periodicity signal by the BIPF [8] and then calculate the breathing rate estimate using the first local maximum in the resulting signal.

Envelope approach: The second baseline is an envelope-based approach adapted from [9]. Due to the sensitivity of this approach to amplitude variations, the raw audio signal is first denoised using a python library called `noisereduce` [13], [14]. The denoising approach relies on example acoustic frames containing noise only, which we select by choosing acoustic frames whose features have a normalized variance above a pre-decided threshold. The denoised signal is then bandpass filtered to exclude low and high frequency cutoffs unrelated to breath sounds. The audio signal is then divided into frames, and the maximum absolute value of audio samples in a given frame is calculated. Cubic spline interpolation is used to ensure that the envelope and the original audio signal have the same length. A similarity function $f(t)$ is then computed between the original envelope $e(l)$ and a time-lagged envelope $e(l+t)$ using the following equation:

$$f(t) = \frac{\sum_{l=0}^{N-T_{max}-1} |e(l) - e(l+t)|}{N - T_{max}} \tag{2}$$

for $T_{min} \leq t \leq T_{max}$ where N is the total length of the envelope and T_{min} and T_{max} represent the lower and upper

[1] International Conference on Biomedical and Health Informatics.

[2] All human subjects data (breathing audio) used in this study was already publicly available online and was not collected by any of the authors of this work. Therefore there was no need to consult our Institutional Review Board for approval.

[3] Audio samples with little background noise, i.e. high Signal-to-Noise Ratio (SNR).

[4] For each frequency component, we compute the autocorrelation at various shifts. Then we compute the variance of those autocorrelation calculations. Frequency components with the highest variance in autocorrelation are the most periodic.

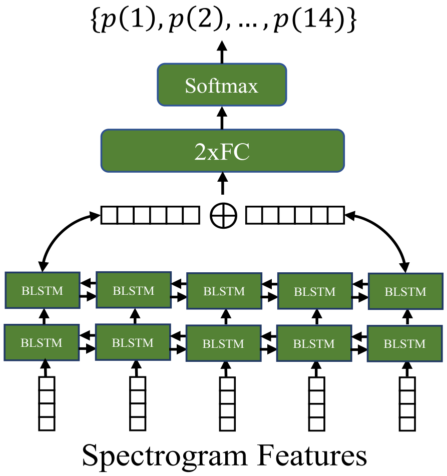

Fig. 1. Neural network architecture for interval count estimator.

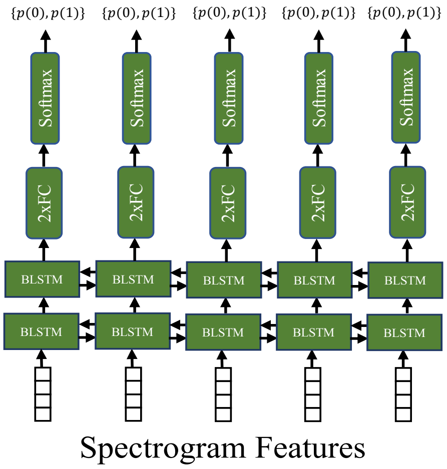

Fig. 2. Neural network architecture of proposed approach.

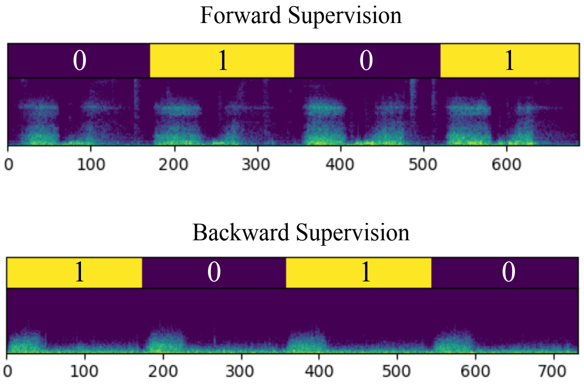

Fig. 3. Provided are an example of forward supervision and backward supervision. Note that for forward supervision, the label of the first inspiration/expiration interval (from left to right) is zero. For backward supervision we have the opposite case, where the last interval (again left to right) is labeled zero.

bounds of the time lag. The breathing rate is ultimately estimated from the first local minimum of the similarity function.

Interval count estimator: The interval count estimator baseline is learning-based, and thus requires a supervision signal. As a simple approach, we choose to model respiratory rate estimation as a classification problem over the number of inspiration/expiration regions present in the audio. For this dataset, the maximum number of regions is 14, so we predict over 14 possible classes where each class index indicates the number of regions in an audio sample[5]. We use a neural network composed of two Bi-directional Long Short-Term Memory (BLSTM) layers and two fully-connected (FC) layers (see Figure 1). The hidden states in the BLSTM and FC layers both have dimension 100. The final state of the forward direction and the first state of the backward direction from the BLSTM are concatenated to create a fixed-length representation to summarize the audio. This representation is then passed through the FC layers where a softmax is computed over the possible classes. We optimize the network using the crossentropy loss, train with the Adam optimizer, and use a learning rate of 0.0001.

V. Methods

Binary Framewise Supervision: Since we have little training data, we want to keep the supervision signal as simple as possible. Due to the periodic nature of respiration, we hypothesize that representing transitions between different regions at the frame level, rather than predicting the total number of inspiration/expiration regions at the sample level, may provide better supervision for this task. We choose to supervise a neural network with a binary framewise signal, where transitions are indicated by switching the class label. We always label each frame corresponding to the first inspiration/expiration region with class 0. We also train a separate system where we reverse the spectrogram along the time axis and label it in the same fashion. We refer to these labeling techniques as forward and backward supervision, respectively. For a visual depiction of the supervision signals, refer to Figure 3.

We use a neural network similar in size and structure to the Interval Count Estimator baseline for our proposed approach (see Figure 2). We use two BLSTM layers and two FC layers both with hidden dimension 100. Unlike the Interval Count Estimator, the FC layers are applied to each hidden state output by the BLSTM, because classification is performed at the frame level. We then optimize the network with crossentropy at each frame, using the Adam optimizer and a learning rate of 0.0001.

Frequency Permutation: Due to the underlying periodicity of the breathing signal, many frequency components should be relatively periodic individually. Note that this is a necessary assumption for the autocorrelation baseline [8]. Thus, we hypothesize that the framewise relationships between different frequency components are not important. For tasks such as automatic speech recognition (ASR), this is not the case, but for this task we can capitalize on

[5] The ground truth count is obtained by counting the number of inspiration/expiration regions in the annotation for an audio sample.

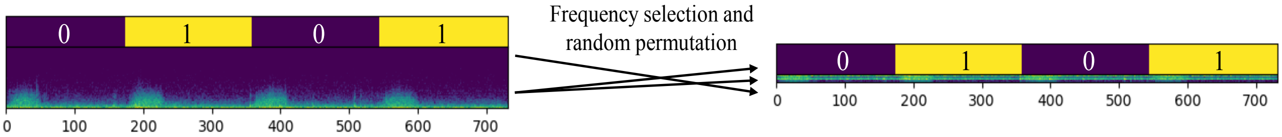

Fig. 4. Given is a visual depiction of the process of frequency permutation of the input spectrogram (80 frequency bins on left). The 20 most periodic frequency components of the sample are selected. 10 components are then randomly selected and shuffled before passing to the model for prediction (right).

frequency independence to propose another form of data augmentation that helps us avoid overfitting. Instead of providing the spectrogram as-is, we randomly permute the frequency components each time we pass a training example to our model. Periodicity information is still present in the input after permutation, but the exact form of the training example will be different every time. We hypothesize that the model learns to look for periodicity in every channel, without being able to overfit to any frequency-specific pattern that may be found in the data.

Since we want our model to focus on periodicity information, we want to avoid distraction that would be caused by providing relatively non-periodic components. To do this, we follow the same approach proposed in the autocorrelation baseline, where we first select the top 20 most periodic components in the audio sample. We further augment the data by randomly choosing a subset of 10 of these frequency components, and randomly permute this subset before passing it as input to our model (see Figure 4).

Respiratory Rate Estimation from Learned Signal: Once training is complete, we need to compute the respiration rate from the framewise class probabilities output by the model. This sequence should alternate between outputting classes 0 and 1 with relatively high probability. We empirically observe that predictions are confident and accurate for the earlier inspiration/expiration regions, and become less confident over time. Thus, we choose to make our estimate of the underlying period of the breath by the length of the first region. The length of this region is determined as the first frame at which the probability of class 0 drops below 0.5. To enhance our prediction, we also train a separate model where we flip the audio and make the same predictions. Thus, we rely on the predicted length of the first and last inspiration/expiration regions, called breathing intervals, to make the final prediction, which is the mean of both individual model predictions (see Figure 5).

VI. Experiments

We run experiments with all baselines and our proposed model under three different noise conditions. For the first condition, no noise is added to the signal such that the signal-to-noise (SNR) ratio is infinity. For the other two conditions, we add white noise such that the SNR is -10dB or -20dB to simulate performance in adverse recording conditions. We split the data into a training and test split, where the training set is used to calculate statistics for the autocorrelation baseline BIPF and train neural networks for the Interval Count Estimator baseline and our proposed approach. We

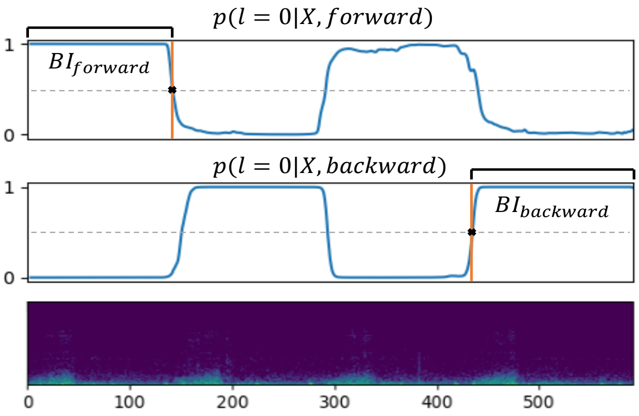

Fig. 5. Visual depiction of estimation of repiratory rate from predicted framewise probabilities using forward and backward supervised models. $BI_{forward}$ refers to the forward breathing interval, $BI_{backward}$ refers to the backward breathing interval, and X refers to the input provided to the neural network (spectrogram is shown in the figure for intuition).

then estimate the respiratory rate for all methods using the test set. Our training set consists of 335 samples (40 of which are held out for validation), and the test set consists of 84 samples.

VII. Results

Mean Absolute Error (MAE) and Standard Deviation (STD) for each method under each noise condition are reported in Table I. We compute statistical significance using a dependent t-test for paired samples.

While statistical significance varies, our method is best for all noise conditions both for MAE and STD, implying more accurate and stable predictions compared to the other techniques. The autocorrelation approach is the most competitive baseline compared to our proposed approach, but its performance suffers in worse noise conditions. We hypothesize that the Interval Count Estimator does a poor job most likely due to weak sample-level supervision and the relatively small amount of data. Since class predictions are integer counts, errors of even one class cause large deviations from the correct respiration rate for samples with few inspiration/expiration regions.

VIII. Ablation

We want to compare the results of our proposed system to a version where the frequency channels are not permuted to determine the importance of permutation for avoiding overfitting. We run an additional experiment where we provide the entire input spectrogram to the model instead

MAE			
SNR	∞	-10dB	-20dB
Autocorrelation	1.9*	4.3	6.4**
Envelope	4.6**	8.1**	7.4**
Interval Count	7.7**	8.7**	8.7**
Proposed	**1.0**	**4.1**	**4.0**
STD			
SNR	∞	-10dB	-20dB
Autocorrelation	4.6	5.3	5.6
Envelope	6.4	5.7	5.6
Interval Count	6.7	7.2	7.2
Proposed	**1.7**	**4.1**	**3.8**

TABLE I

COMPARISON OF METHODS. THE BEST VALUE AT EACH SNR LEVEL IS BOLDED. FOR MAE, STATISTICAL SIGNIFICANCE BETWEEN THE PROPOSED APPROACH AND ALL OTHER METHODS IS DENOTED BY * FOR $p < 0.1$ OR ** FOR $p < 0.05$ ON EACH OF THE OTHER METHODS, RESPECTIVELY.

of the random subset of 10 frequency channels from the 20 most periodic channels. Results for this experiment are given in Table II. We find that the system without frequency permutation performs worse than the system with frequency permutation at a statistically-significant level. This provides strong evidence that providing the entire spectrogram as input to the model with so few training examples allows the model to overfit and leads to worse generalization than the model with permuted frequency channels.

IX. FUTURE WORK

We explore a novel way to supervise a neural respiratory rate estimation system through binary framewise labels, but our approach to estimate the breathing rate from the predicted framewise probabilities of the trained model can be improved. We use only the first and last predicted breathing interval boundaries to estimate the breathing rate, which may be suboptimal to an approach that takes all predicted interval boundaries into account. An improved method for using the predicted framewise probabilities from our model to determine respiratory rate is left to future work.

	MAE	STD
No Freq Permutation	1.8**	3.0
Freq Permutation	**1.0**	**1.7**

TABLE II

ABLATION STUDY WITH SNR$= \infty$. STATISTICAL SIGNIFICANCE AT $p < 0.05$ IS DENOTED WITH ** FOR MAE BETWEEN "NO FREQ PERMUTATION" AND "FREQ PERMUTATION" OPTIONS.

X. CONCLUSIONS

In this paper, we propose two techniques that can be used to improve training of deep neural networks for respiratory rate estimation. Binary framewise labeling can be used to provide a stronger supervision signal to the neural network. Additionally, frequency permutation, which relies on the principle that frequency components should be periodic individually, can be used as a data augmentation method to avoid model overfitting for this task. We show that the proposed approach performs better than all baselines and that a system with frequency permutation outperforms a system without frequency permutation. Given that the ICBHI Respiratory Sounds Database is the only known public database of labeled respiratory sounds, and is relatively small, our proposed approach offers added benefit in that it performs extremely well with very few labeled training examples. The proposed approach demonstrates how a supervised respiratory rate estimation system can be built within a low-data setting.

REFERENCES

[1] John F Fieselmann, Michael S Hendryx, Charles M Helms, and Douglas S Wakefield, "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients," *Journal of general internal medicine*, vol. 8, no. 7, pp. 354–360, 1993.

[2] Thomas R Gravelyn and John G Weg, "Respiratory rate as an indicator of acute respiratory dysfunction," *Jama*, vol. 244, no. 10, pp. 1123–1125, 1980.

[3] Helen Ansell, Alannah Meyer, and Shona Thompson, "Why don't nurses consistently take patient respiratory rates?," *British Journal of Nursing*, vol. 23, no. 8, pp. 414–418, 2014.

[4] Tracy Flenady, Trudy Dwyer, and Judith Applegarth, "Rationalising transgression: a grounded theory explaining how emergency department registered nurses rationalise erroneous behaviour.," *Grounded Theory Review*, vol. 15, no. 2, 2016.

[5] Tracy Flenady, Trudy Dwyer, and Judith Applegarth, "Accurate respiratory rates count: So should you!," *Australasian Emergency Nursing Journal*, vol. 20, no. 1, pp. 45–47, 2017.

[6] Andrew Hill, Eliza Kelly, Mark S Horswill, and Marcus O Watson, "The effects of awareness and count duration on adult respiratory rate measurements: an experimental study," *Journal of clinical nursing*, vol. 27, no. 3-4, pp. 546–554, 2018.

[7] Catherine Redmond, "Trans-thoracic impedance measurements in patient monitoring," *EDN Network*, 2013.

[8] Eliran Dafna, Tal Rosenwein, Ariel Tarasiuk, and Yaniv Zigel, "Breathing rate estimation during sleep using audio signal analysis," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 5981–5984.

[9] Yanzhi Ren, Chen Wang, Jie Yang, and Yingying Chen, "Fine-grained sleep monitoring: Hearing your breathing with smartphones," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1194–1202.

[10] Venkata Srikanth Nallanthighal and H Strik, "Deep sensing of breathing signal during conversational speech," 2019.

[11] Cristina Jácome, Johan Ravn, Einar Holsbø, Juan Carlos Aviles-Solis, Hasse Melbye, and Lars Ailo Bongo, "Convolutional neural network for breathing phase detection in lung sounds," *Sensors*, vol. 19, no. 8, pp. 1798, 2019.

[12] BM Rocha, Dimitris Filos, L Mendes, I Vogiatzis, E Perantoni, E Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al., "A respiratory sound database for the development of automated classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 33–37.

[13] Tim Sainburg, "timsainb/noisereduce: v1.0," June 2019.

[14] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, pp. e1008228, 2020.