*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# COVID-19 Detection from Spectral features on the DiCOVA Dataset

*Kotra Venkata Sai Ritwik, Shareef Babu Kalluri, Deepu Vijayasenan*

Department of Electronics and Communication Engineering,
National Institute of Technology Karnataka Surathkal, India

{sairitwik270600,shareefbabu1,deepu.senan}@gmail.com

## Abstract

In this paper we investigate the cues of COVID-19 on sustained phonation of Vowel-/i/, deep breathing and number counting data of the DiCOVA dataset. We use an ensemble of classifiers trained on different features, namely, super-vectors, formants, harmonics and MFCC features. We fit a two-class Weighted SVM classifier to separate the COVID-19 audio from Non-COVID-19 audio. Weighted penalties help mitigate the challenge of class imbalance in the dataset. The results are reported on the stationary (breathing, Vowel-/i/) and non-stationary(counting data) data using individual and combination of features on each type of utterance. We find that the Formant information plays a crucial role in classification. The proposed system resulted in an AUC score of 0.734 for cross validation, and 0.717 for evaluation dataset.

**Index Terms**: COVID-19, acoustics, machine learning, respiratory diagnosis, healthcare

## 1. Introduction

The research on novel coronavirus (COVID-19) has become the majorly focused research in the pandemic situation spread around across 200 countries the world since last year 2020. Extensive research involved in identification of infected people with corona virus is challenging for health organizations as well as individuals. The speared of the virus is majorly with contacts and aerosol molecules released from sneezing, cough, and cold/flu[1].

The rise in body temperature, difficulty in breathing, cough and cold are the majorly observed symptoms of COVID-19; however, totally asymptomatic cases are also possible[1]. The clinical protocols in identifying whether the individual is infected with coronavirus include swab test [2], CT scans [3, 4], chest X-Ray Images [5, 6] etc. Along with clinical methods of identifying COVID-19, researchers have also explored machine learning (ML) and deep learning (DL) methods [7], in leveraging the bio-markers like speech and audio signals in screening the COVID-19 to help in the assessment of the viral infection.

Different studies and pathological investigations have proved that COVID-19 infected individuals exhibit difficulty while breathing and speaking. Furthermore, the changes in speech might not be identifiable by human perception, although the person is infected, but the Computer auditions (CA) [8, 9, 10, 11] can. In this work we tried to investigate detection of the COVID-19 from speech sounds counting numbers, breathing and speech sounds i.e, Vowel-/i/ from DiCOVA dataset [12].

There have been a few attempts in the literature to use cough patterns to classify pathological conditions such as bronchitis and pertussis[13, 11]. Many efforts have been attempted in collecting the data for detecting COVID-19 using web based applications as well as mobile apps [14, 15, 16] etc,. The research efforts made on the collected speech/cough/breathing data is mainly focused on the machine learning approaches.

According to research on COVID-19 patients, infected people's respiration activity is higher than that of other flu and common cold patients [17]. From a survey different types of features extracted from speech/breath/cough data; Mel frequency cepstral coefficients (MFCC), spectral features, temporal features and spectrograms, etc, are used for COVID-19 detection [18]. Other research has found that respiratory parameters have an effect on a person's stress levels [19], mood/emotion[20], and physiological state of individuals [21]. Various research attempts have been made to estimate the behavioural state of humans using speech data in situations such as cold, cough, discomfort, pain, sleepiness, and infant cries as part of the Interspeech Computational Para-linguistics Challenge (ComParE). Schuller et al. proposed a set of computer audition tasks for COVID-19 risk assessment using machine learning techniques, including speech analysis and sound analysis.[9]. Few other studies on patients with health problems like Obstructive Sleep Apnea Detection (OSA) used machine learning approaches using conventional features like formants, pitch, MFCCs, and Linear prediction cepstral coefficients etc,[22]. In this work, we focus on the DiCOVA Track-2 dataset, and explored different spectral features for detection of COVID-19 from the breathing patterns, numbers counting from one to twenty and Vowel-/i/.

The organization of the paper is as follows, Section 2 describes about the dataset used i.e, DiCOVA track 2 dataset. Section 3 details the feature extraction and statistical representation of each phoneme as well as sentence. Details of the classification method and the experiments and results are detailed in Section 4. Finally, the conclusions of the reported work and future directions of the proposed approach are presented in Section 5.

## 2. Dataset

The dataset is taken from the DiCOVA challenge Track-2 dataset as described in [12]. There are 3 different modes of audio recordings in the data, namely Breathing-Deep, counting of digits from one to twenty at a normal pace and sustained utterance of the vowel /i/ from 1199 subjects. The training split consists of 990 speakers with 930 speakers in the COVID-19 negative class and 60 speakers in COVID-19 positive class. The evaluation data consists of 188 speakers in the COVID-19 negative class and 21 speakers in COVID-19 positive class. All the audio files are of 44.1 kHz, mono channel type in .flac format. We convert the .flac files to .wav and down sample them to 16kHz and 8kHz.

## 3. Feature Description

In this work, we explored features from the speech at different time resolutions. We experimented with both short and long temporal windows to extract different information pertaining to



http://dx.doi.org/10.21437/Interspeech.2021-1031

MFCCs, formant locations, pitch and harmonic frequency locations from the input signal.

In case of the non-stationary input signal (counting digits data), we use a super-vector like feature representation from the short term spectrum. In case of the stationary input signals (Breathing, vowel -/i/ ) we use the average MFCC across the file as feature representation. In combination with these, statistics of frame-wise formant and harmonic locations are also used as features. We describe each individual feature below

### 3.1. Super vector features

Super-vectors[23] represent mean of the short term features for different sound classes. These were originally used in speaker recognition. We use these super vectors to extract utterance level features from the non-stationary input signal – i.e., the counting-digits data. Here each user counts from one to twenty and is considered one single utterance.

We use the ASpIRE chain model to compute the frame level posterior probabilities of each utterance in the Counting-Normal dataset. The ASpIRE model is a Time-Delayed Neural Network (TDNN) trained on the Fisher English dataset [24]. Data augmentation by means of different room impulse responses and noises are incorporated in the model to increase its robustness [25]. The ASpIRE model takes short term mel spectrum as the input and computes the frame-level posteriors of different context-dependent phonemes. We windowed the audio files into 25ms long, and a shift of 10ms and extracted the 40-dimensional MFCC coefficients. In addition; we computed the frame level phoneme posteriors from the ASpIRE chain model [26]. Although ASpIRE model outputs context dependent phoneme posteriors, we sum over the contexts to obtain context independent phoneme posteriors. We thus obtain a 39 dimensional posterior vector corresponding to the 39 TIMIT phonemes. Silence phoneme is discarded. We then compute the normalized first order statistics of each phoneme in the following manner.

Consider an input sequence of short term mel-spectral features from a speech utterance $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_T}\}$. The frame level posteriors output by the model are $\{\mathbf{p_1}, \ldots, \mathbf{p_T}\}$ where:

$$\mathbf{p_j} = \begin{bmatrix} \mathrm{p}_j^1 \\ \mathrm{p}_j^2 \\ \vdots \\ \mathrm{p}_j^M \end{bmatrix} \tag{1}$$

with $p_j^i$ denoting the posterior probability of phoneme $i$ at the frame index $j$. Once the posterior probabilities are calculated, we compute the normalized first order statistics of each phoneme as :

$$\mathbf{f^i} = \frac{1}{\sum_j p_j^i} \sum_j p_j^i \mathbf{x_j} \tag{2}$$

We concatenate all frames of $\mathbf{f^i}$ of each phoneme to obtain a super vector $\mathbf{F} = [\mathbf{f^1}, \mathbf{f^2}, \ldots, \mathbf{f^M}]$ which represents the utterance level feature.

Each phoneme's first order statistics is a 40 dimensional vector,they are concatenated to form a $40 \times 39 = 1560$ dimensional super-vector for each utterance. These 1560 dimensional super-vectors are used as features to fit a support vector machine (SVM) model.

### 3.2. Average MFCC features

In the case of Vowel-/i/ and breathing data, the vocal tract is stationary. Therefore is not very meaningful to compute phoneme posteriors. In this the average of the short term MFCC features across all frames is used as input features. We keep all the coefficients without cepstral truncation. Since we compute 40 dimensional MFCC features, the averaged feature also will be a 40 dimensional feature vector. We call these the average MFCC features.

### 3.3. Fundamental Frequency and Formants

This set of features is based on a longer temporal context. We use the pitch and formant information as input features. We use the PEFAC algorithm [27] which uses both noise rejection and normalization while ensuring temporal continuity in the pitch estimates. Formant frequencies are computed by capturing the peaks of an $18^{th}$ order auto regressive (AR) model. This results in nine peak locations. We use the first four peak locations to capture the formant frequencies denoted by F1, F2, F3, F4. The percentile values (5,25, 50, 75, 95) of formants across the entire utterance are computed as features and are used to train a SVM model for classification [28]. We call these as the formants features.

### 3.4. Harmonics

This is also a set of features based on a longer context. Basically we are motivated to use these features since, they could capture jitter, shimmer etc. Thirty harmonics are extracted from an 80 order AR model computed over a time window of length 60ms and shift 10ms. Similar to formants we compute the $5^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ and $95^{th}$ percentile values across the entire utterance along with the standard deviation of the percentiles to be used as features. We call these the harmonic features.

## 4. Experiments and Results

We performed all our experiments on the DiCOVA track 2 dataset. The dataset details are mentioned in Section 2.

We extract the short term MFCCs from the input speech files. These are used to compute super-vector features in case of counting data. The super-vector computation uses the phoneme posteriors predicted by the ASPIRE model. The statistics are aggregated across all frames to compute the utterance level features (Eqn 2.) In case of the other two inputs (i.e., breathing and vowel /i/) the signal is stationary and therefore the MFCC is simply averaged to obtain the utterance level feature.

The pitch, formant frequencies, and harmonic frequencies are computed at the frame level using a long temporal context (80ms ). Different percentiles (5, 25, 50, 75, 95) of each of the features (pitch, formant frequencies, harmonic frequencies etc.,) are used as the utterance level features. We exclude the pitch information in the case of breathing data.

These features are used to learn individual SVM models. In order to assess the individual feature performances, and tune the hyper-parameters of the system (SVM penalty terms, feature combination scheme), we perform a 5-fold cross validation using the folds given in the DiCOVA Track-2 dataset mentioned (see Section 2 ) for each feature separately.

We used a class-weighted SVM model with a Radial Basis Function (RBF) as kernel to detect COVID-19 speakers. The penalty parameter $C$ is weighted differently for positive and negative class. The weights are inversely proportional to the

number of samples in the class. This allows the margin to be softer on the COVID-19 positive samples but forces the margin to be harder on the COVID-19 negative samples and minimize false negatives.

Table 1: *Sensitivity, Specificity and AUC metrics on Counting-Normal data using different features on cross validation data (CV) and evaluation data (Eval) of DiCOVA dataset.*

| Counting-Normal | | | |
|---|---|---|---|
| Features | Sensitivity | Specificity | AUC |
| Super-vectors (CV) | 0.610 | 0.614 | 0.642 |
| Formants (CV) | **0.629** | **0.626** | 0.666 |
| Harmonics (CV) | 0.610 | 0.609 | 0.641 |
| All Features(CV) | 0.619 | 0.617 | **0.675** |
| Super-vectors (Eval) | 0.619 | 0.570 | 0.638 |
| Formants (Eval) | **0.714** | 0.617 | **0.697** |
| Harmonics (Eval) | 0.524 | **0.654** | 0.652 |
| All Features (Eval) | 0.638 | 0.638 | 0.683 |

The model is evaluated using Sensitivity, Specificity and Area Under the Receiver Operating Characteristic curve (AUC). Sensitivity and Specificity are given by:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

Where TP denotes number of true positives, FP denotes the false positives, FN denotes the false negatives, TN denotes the true negatives.

We perform hyper-parameter optimization for the SVM penalty parameter across the 5-folds to maximize the AUC value. We thus aim for maximization of the AUC as the metric to assess the classifier's performance. We compute the equal error rate point in the ROC that is the place where the false positive rates and false negative rate are equal, where:,

$$\text{False Positive Rate } = \frac{FP}{(FP + TN)}$$

$$\text{False Negative Rate } = \frac{FN}{(FN + TP)}$$

Sensitivity, specificity and AUC values for counting-digits,Vowel-/i/ and Breathing-Deep dataset are shown in Table 1, Table 2 and Table 3 respectively. The ROC-curves for counting-digits, Vowel-/i/ and Breathing-Deep are shown in Figure 1, Figure 2 and Figure 3. respectively.

Table 1 shows the result of the counting speech data for different features. We note that formant features perform the best on the counting-digits data followed by harmonics and super-vectors. Katrin *et.al* has reported that the combination of multiple features can characterise COVID-19 in significant way when compared with individual feature set [29]. We also perform a similar feature combination on all three computed feature vectors (super-vectors, formants and harmonics) by averaging the scaled probabilities from each of the individual SVMs. The feature combination results in an average AUC of 0.675 on the validation and AUC of 0.683 on evaluation data. Even though the combination is better than the super-vectors and harmonics, the formant features alone yield the best performance on the evaluation data.

Table 2: *Sensitivity, Specificity and AUC metrics on Vowel-/i/ data using different features on cross validation data (CV) and evaluation data (Eval) of DiCOVA dataset.*

| Vowel-/i/ | | | |
|---|---|---|---|
| Features | Sensitivity | Specificity | AUC |
| MFCC (CV) | **0.648** | **0.649** | **0.687** |
| Formants (CV) | 0.600 | 0.597 | 0.619 |
| Harmonics (CV) | 0.600 | 0.624 | 0.632 |
| All Features (CV) | 0.629 | 0.628 | 0.677 |
| MFCC (Eval) | 0.476 | 0.622 | 0.603 |
| Formants (Eval) | 0.573 | **0.633** | 0.583 |
| Harmonics (Eval) | **0.619** | 0.559 | 0.586 |
| All Features (Eval) | 0.571 | 0.574 | **0.611** |

Table 3: *Sensitivity, Specificity and AUC metrics on Breathing-Deep data using different features on cross validation data (CV) and evaluation data (Eval) of DiCOVA dataset*

| Breathing Deep | | | |
|---|---|---|---|
| Features | Sensitivity | Specificity | AUC |
| MFCC (CV) | **0.619** | 0.597 | 0.614 |
| Formants (CV) | 0.610 | **0.613** | 0.651 |
| All Features (CV) | 0.600 | 0.608 | **0.654** |
| MFCC (Eval) | 0.524 | 0.554 | 0.592 |
| Formants (Eval) | **0.667** | 0.590 | **0.717** |
| All Features(Eval) | 0.619 | **0.676** | 0.691 |

From Table 2 we note that MFCC features perform the best on the Vowel-/i/ data followed by harmonics and formants. Similarly, we performed feature combination on MFCC, formants and harmonics as described above. This results in an AUC of 0.677 on validation and an AUC of 0.611 on evaluation data. In this case, the feature combination is marginally better than the best performing feature on evaluation data.

From Table 3 we note that formants perform the best on the Breathing-Deep data followed by MFCC features. The combination of MFCC and formant features on the Breathing-Deep data has obtain an average AUC of 0.654 on validation and an AUC of 0.691 on evaluation data. Here the feature combination outperforms the MFCC feature, although it is marginally worse than formants.

Table 4: *AUC of each feature combined across tasks on validation data (CV) and evaluation data (Eval) of DiCOVA dataset*

| Features | AUC |
|---|---|
| Super-vectors+MFCC (CV) | 0.709 |
| Formants (CV) | **0.713** |
| Harmonics (CV) | 0.664 |
| Super-vectors+MFCC (Eval) | 0.630 |
| Formants (Eval) | **0.752** |
| Harmonics (Eval) | 0.626 |

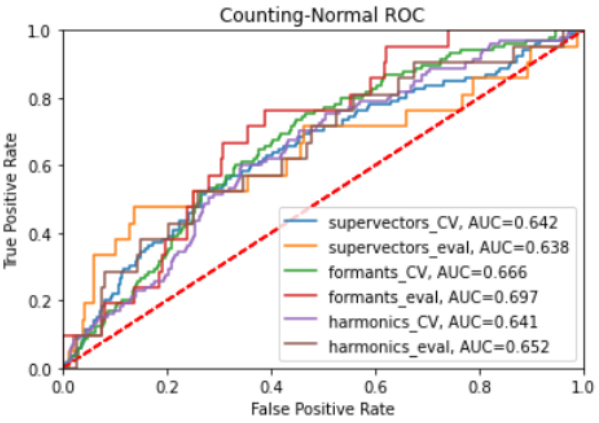


Figure 1: *ROC of counting data of DiCOVA dataset*

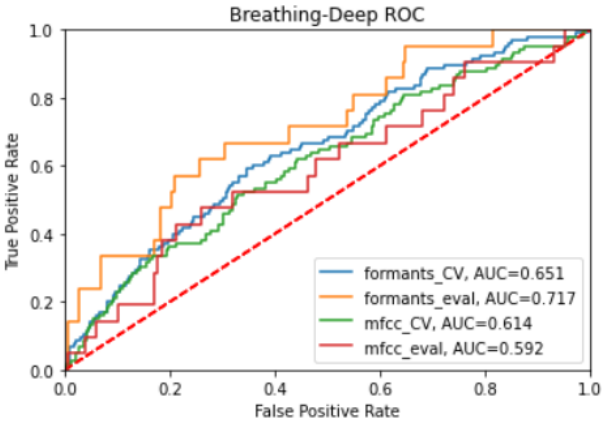


Figure 3: *ROC of breathing data of DiCOVA dataset*

We also try to study the performance of each feature across the different tasks. We perform different combinations for all the tasks. First, we consider formant features across the three different tasks. Next, we combined super-vectors and MFCC (SV+MFCC) features across all the three different tasks. Later, we also consider harmonic features across the counting and vowel tasks.

We observe from Table 4 that formant information plays an important role in the classification. The ROC curves are shown in the Figure 5. We find that the harmonic features performance is inferior to the other features, and therefore dropped this feature from further evaluation.

We now to try to combine the results across three different input signals ( Counting, breathing and vowel - /i/). Again we simply average the individual probabilities of each feature from different input signals. From each of the three signals we have a SVM probability for short term spectral based features and Formant features. We average these six probability values to get the final combination. This results in an average AUC of 0.734 on validation and an AUC of 0.717 on the evaluation data. The ROC curve is shown in Figure 4.

This feature combination has improved the performance compared to any single feature best result from any of the input signals. Thus there is some complementary information between the different signals and features.

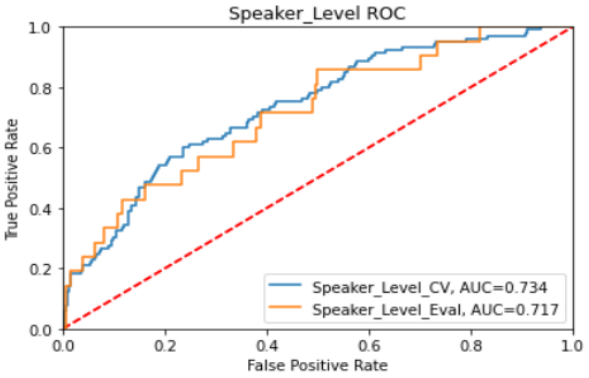


Figure 4: *ROC of feature combination of DiCOVA dataset*

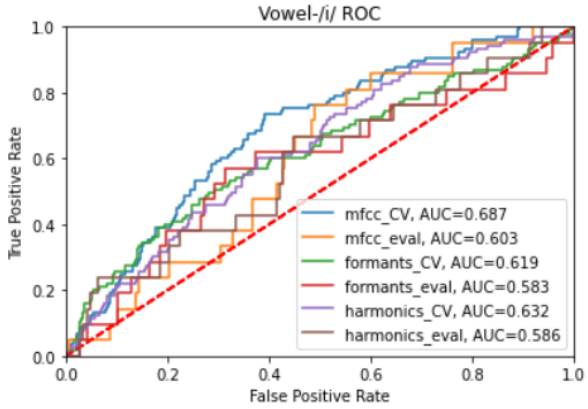


Figure 2: *ROC of Vowel -/i/ data of DiCOVA dataset*

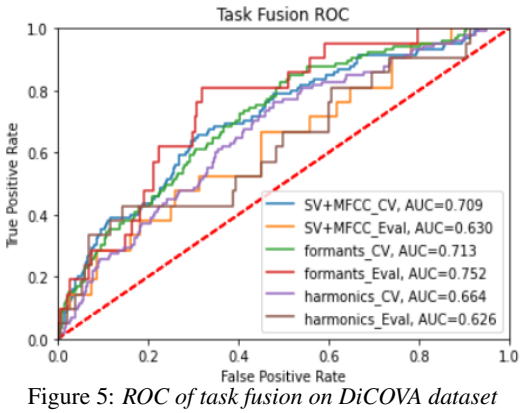


Figure 5: *ROC of task fusion on DiCOVA dataset*

## 5. Conclusion

We have designed an SVM system with features based on formants, harmonic frequencies and short term MFCC for COVID-19 detection in the DiCOVA Challenge Track-2 Dataset. We used super-vector based features for non-stationary input (counting) and average MFCC features for stationary input (breathing, Vowel-/i/) as short term features. We then combined the SVM models of formants and short term MFCC features across all the signals, by averaging the individual SVM probabilities. The proposed system resulted in an AUC score of 0.734 for cross validation, and 0.717 for evaluation dataset that is better than any of the individual features. In the future, we would like to study additional features and assess the performance of our system across different datasets.

# 6. References


[1] "WHO Coronavirus Disease (COVID-19) Dashboard," https://covid19.who.int/, 2020, [Online; accessed 10-Feb-2021].

[2] C. for Disease Control and Prevention, "Coronavirus disease (COVID-19) Test for Current Infection," https://www.cdc.gov/coronavirus/2019-ncov/testing/diagnostic-testing.html, 2020.

[3] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid AI development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *arXiv preprint arXiv:2003.05037*, 2020.

[4] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (covid-19) classification using ct images by machine learning methods," *arXiv preprint arXiv:2003.09424*, 2020.

[5] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *arXiv preprint arXiv:2003.09871*, 2020.

[6] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect covid-19 disease from x-ray images using capsule networks," *Chaos, Solitons & Fractals*, vol. 140, p. 110122, 2020.

[7] A. Hassan, I. Shahin, and M. B. Alsabek, "Covid-19 detection system using recurrent neural networks," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 2020, pp. 1–5.

[8] T. F. Quatieri, T. Talkar, and J. S. Palmer, "A framework for biomarkers of covid-19 based on coordination of speech-production subsystems," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 203–206, 2020.

[9] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "Covid-19 and computer audition: An overview on what speech & sound analysis could contribute in the sars-cov-2 corona crisis," *arXiv preprint arXiv:2003.11117*, 2020.

[10] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," *Proceedings INTERSPEECH. Shanghai, China: ISCA*, 2020.

[11] I. D. Miranda, A. H. Diacon, and T. R. Niesler, "A comparative study of features for acoustic cough detection using deep architectures," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2601–2605.

[12] A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy *et al.*, "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[13] A. Windmon, M. Minakshi, S. Chellappan, P. Athilingam, M. Johansson, and B. A. Jenkins, "On detecting chronic obstructive pulmonary disease (copd) cough using audio signals recorded from smart-phones." in *HEALTHINF*, 2018, pp. 329–338.

[14] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara – a database of breathing, cough, and voice sounds for COVID-19 diagnosis," in *Proc. INTERSPEECH, ISCA*, 2020.

[15] "COVID-19 Sounds App by University of Cambridge University ," https:https://www.covid-19-sounds.org/en, 2020.

[16] "EPFL Cough for COVID-19 Detection," https://coughvid.epfl.ch/, 2020, [Online; accessed 07-Aug-2020].

[17] Y. Wang, M. Hu, Q. Li, X.-P. Zhang, G. Zhai, and N. Yao, "Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with covid-19 in an accurate and unobtrusive manner," *arXiv preprint arXiv:2002.05534*, 2020.

[18] G. Deshpande and B. W. Schuller, "Audio, speech, language, & signal processing for covid-19: A comprehensive overview," *arXiv preprint arXiv:2011.14445*, 2020.

[19] V. Perciavalle, M. Blandini, P. Fecarotta, A. Buscemi, D. Di Corrado, L. Bertolo, F. Fichera, and M. Coco, "The role of deep breathing on stress," *Neurological Sciences*, vol. 38, no. 3, pp. 451–458, 2017.

[20] R. A. Hameed, M. K. Sabir, M. A. Fadhel, O. Al-Shamma, and L. Alzubaidi, "Human emotion classification based on respiration signal," in *Proceedings of the International Conference on Information and Communication Technology*, 2019, pp. 239–245.

[21] M. J. Griffiths, D. F. McAuley, G. D. Perkins, N. Barrett, B. Blackwood, A. Boyle, N. Chee, B. Connolly, P. Dark, S. Finney *et al.*, "Guidelines on the management of acute respiratory distress syndrome," *BMJ open respiratory research*, vol. 6, no. 1, p. e000420, 2019.

[22] M. C. Botelho, I. Trancoso, A. Abad, and T. Paiva, "Speech as a biomarker for obstructive sleep apnea detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5851–5855.

[23] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[24] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[25] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[26] K. V. S. Ritwik, S. B. Kalluri, and D. Vijayasenan, "Covid-19 patient detection from telephone quality speech data," 2020.

[27] S. Gonzalez and M. Brookes, "Pefac-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[28] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Communication*, vol. 121, pp. 16–28, 2020.

[29] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. W. Schuller, "The voice of covid-19: Acoustic correlates of infection," 2020.