The DiCOVA 2021 Challenge: an encoder-decoder approach for COVID-19 recognition from coughing audio

Gauri Deshpande, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Deshpande, Gauri, and Björn W. Schuller. 2021. "The DiCOVA 2021 Challenge: an encoder-decoder approach for COVID-19 recognition from coughing audio." In *Interspeech 2021, Brno, Czechia, 30 August - 3 September 2021*, edited by Hynek Heřmanský, Honza Černocký, Lukáš Burget, Lori Lamel, Odette Scharenborg, and Petr Motlicek, 931–35. Baixas: International Speech Communication Association (ISCA). https://doi.org/10.21437/interspeech.2021-811.

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

The DiCOVA 2021 Challenge – An Encoder-Decoder Approach for COVID-19 Recognition from Coughing Audio

Gauri Deshpande[1,2], *Björn W. Schuller*[2,3]

[1]TCS Research Pune, India
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
gauri1.d@tcs.com, schuller@ieee.org

Abstract

This paper presents the automatic recognition of COVID-19 from coughing. In particular, it describes our contribution to the DiCOVA challenge – Track 1, which addresses such cough sound analysis for COVID-19 detection. Pathologically, the effects of a COVID-19 infection on the respiratory system and on breathing patterns are known. We demonstrate the use of breathing patterns of the cough audio signal in identifying the COVID-19 status. Breathing patterns of the cough audio signal are derived using a model trained with the subset of the UCL Speech Breath Monitoring (UCL-SBM) database. This database provides speech recordings of the participants while their breathing values are captured by a respiratory belt. We use an encoder-decoder architecture. The encoder encodes the audio signal into breathing patterns and the decoder decodes the COVID-19 status for the corresponding breathing patterns using an attention mechanism. The encoder uses a pre-trained model which predicts breathing patterns from the speech signal, and transfers the learned patterns to cough audio signals.

With this architecture, we achieve an AUC of 64.42 % on the evaluation set of Track 1.

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

1. Introduction

COVID-19 is an infectious disease caused by a newly discovered spread via the Coronavirus SARS-CoV-2 having adverse effects on the functioning of human respiratory system [1]. Individuals infected with COVID-19 experience mild to moderate respiratory illness. To a large extent, identifying the presence of COVID-19 has been attempted from cough audio signals [1, 2, 3, 4, 5, 6]. The authors of [1, 7, 8, 9] have analysed breathing audio signals along with coughing of COVID-19 subjects. The speech signal is analysed in [10, 11, 12] with a focus on vowels, alphabets, & counting from 1 to 10.

The DiCOVA 2021 challenge – Track 1 provides cough audio signals & Track 2 provides breathing, sustained phonation, and speech signals for the detection of a subject's COVID-19 status [13]. As reported in [1] and [7], breathing patterns are found more effective in identifying the bio-markers of COVID-19 in human produced audio. Also, [8, 9] have demonstrated the combined analysis of coughing and breathing giving better performance. However, capturing low amplitude breathing signals in noisy environments is challenging. To overcome the problem of data collection in the right context, we propose to use the human-audio signals and extract their breathing patterns. The main contributions of this paper can be summarised as: First, we demonstrate the technique to extract the breathing pattern of a cough audio signal. Second, we introduce the use of such breathing patterns derived from audio for detecting COVID-19 bio-markers. To the best of our knowledge, it is the first work on using audio-based breathing parameters in the context of COVID-19.

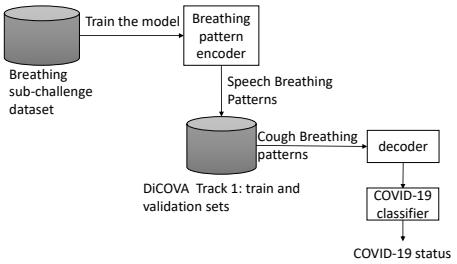

Figure 1: *Conceptual model of the proposed COVID-19 recogniser informed by a data-trained breathing predictor from audio.*

2. Related Work

Several studies in the past have worked with cough audio signals for the detection of respiratory diseases. A dataset of cough audio from 20 patients having pertussis, 11 having croup, and 7 containing wheezing sounds corresponding to other diseases such as bronchiolitis and asthma is used in [16] and [17]. In these two studies, the authors have diagnosed pertussis cough and classified cough from other non-cough sounds. The authors report a sensitivity of 90.31 %, a specificity of 98.14 %, and an F1-score of 88.70 %, using three spectral domain features and logistic regression as classifier in classifying cough from non-cough sounds. With a feature set of Mel-Frequency Cepstral Coefficient (MFCC), Zero Crossing Rate (ZCR), crest factor, energy, dominant frequency, and other spectral parameters such as roll-off, skewness, kurtosis, centroid, spread, decresase, flatness, slope, standard deviation and band power, the pertussis cough is identified with 100 % accuracy on a subset of data comprising of 10 pertussis and 11 non-pertussis cough audio

[1]https://www.who.int/health-topics/coronavirus

[2]https://coughvid.epfl.ch
[3]opensigma.mit.edu
[4]https://www.coughagainstcovid.org
[5]https://voicemed-791a3.firebaseapp.com
[6]https://voca.ai/corona-virus

http://dx.doi.org/10.21437/Interspeech.2021-811

Table 1: *COVID-19 detection from speech, breathing and coughing audio data. The modes of audio used by the groups :- S: Speech; B: Breath; C: Cough. The features used by the groups :- MFCC; VFO: Vocal Fold Oscillation. The ML techniques used are:- SVM: Support Vector Machine; LR: Logistic Regression; AUC: Area Under the Curve.*

Group Name	Mode	COVID-19 subjects' count	Features & ML Techniques	Performance
Cambridge [1]	S, B, C	235	MFCC with SVM, LR	0.8 AUC
Coswara [14]	S, B, C	104		
Coughvid[2] [15]	C	632		
AI4COVID [2]	C	70	MFCC with CNN, SVM	Sensitivity 0.77
MIT[3] [3]	C	2660	MFCC with CNN	98 % Accuracy
CoughAgainstCovid[4] [4]	C	2001	MFCC with CNN	0.31 Specificity at 0.9 Sensitivity
VoiceMed[5]	C	165	Spectrogram with CNN	88 % Accuracy
Voca[6] [10]	S	30	MFCC with CNN	70 % Accuracy
CMU [11]	S	299	VFO with CNN	0.9 AUC

recordings only. In-clinic and outside clinic research studies are conducted in [18] with speech from 70 and 131 participants respectively. The authors report a classification accuracy of 75 % with a RandomForest classifier for the prediction of pulmonary disorders and a mean absolute error of 9.8 % for the the ratio of a person's vital capacity to expire in the first second of forced expiration to the full forced vital capacity (FEV1/FVC) prediction task using an eight dense layered neural network. The seven most relevant features identified by the authors are frequency of pause while speaking, shimmer, absolute jitter, relative jitter, maximum of Fast-Fourier Transform (FFT) of inspiratory sound in frequencies from 7.8 kHz to 8.5 kHz, mean of phonation period to inspiratory period ratio, and average phonation time. Yadav et al. [19] used the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) baseline acoustic features [20] for the classification of 47 asthmatic and 48 healthy individuals with a classification accuracy of 75.4 % using voiced speech sounds. The authors compared the performance exhibited by these features with that of only MFCCs, and report an absolute improvement of 18.28 % over the accuracy given by only MFCCs. Recently, several efforts for the identification of COVID-19 cough from non-COVID-19 cough are seen [21, 22, 23]. A summary is presented in Table 1.

Breathing signal analysis has already found its significance in detecting COVID-19 bio-markers. The COVID-19 detection (from non-COVID-19 asthmatic cough) performance reported by Cambridge University[7] in [1] states an Area Under the Curve (AUC) of 0.8. Similarly, the authors of [7] also found breathing signals performing better with an absolute improvement of 2.4 % Unweighted Average Recall (UAR) than coughs in classifying COVID-19 subjects vs healthy subjects. They reported a UAR of 76.1 % using breathing sound and 73.7 % for coughing sound from the data set collected by the Cambridge University [1]. Another study using a subset of data collected by Cambridge University [1] is presented in [8]. The data comprises of coughing and breathing audio recordings from 62 COVID-19 positive and 293 healthy participants. The authors applied an end-to-end deep network on the joint representation of coughing and breathing audio signals and reported an AUC of 0.846. Recently, there have been multiple efforts in identifying breathing patterns of speech signals using regression techniques. In the Breathing Sub-challenge of the Interspeech 2020 Computational Paralinguistics (ComParE) [24], a Pearson correlation of r = 0.507 on the development, and r = 0.731 on the test data set is presented. The winners of this challenge [25], reported r = 0.763 between the speech signal and corresponding breathing values of the test set. Further, inhalation events are detected from the breathing pattern identified from speech signals in [26]. Motivated by the outstanding performance of state-of-the art approaches, we explored the use of these breathing patterns for detection of COVID-19 cough.

[7] https://www.covid-19-sounds.org/en

3. System Description

As reported in [1] and [7], the breathing patterns are found to be more effective in identifying the bio-markers of COVID-19 in human produced audio. With this motivation, we propose to represent the cough audio signals as breathing patterns and use them for detection of COVID-19 status.

3.1. Methodology Overview

As seen in the Figure 1, we use the UCL-SBM dataset provided in the breathing sub-challenge of the Interspeech 2020 Computational Paralinguistics ChallengE (ComParE) [24], to train an encoder model which predicts breathing patterns of an incoming audio signal. This pre-trained encoder is further used to predict the breathing patterns of the cough audio signals shared in the Track 1 of the DiCOVA challenge [13] . These cough-breathing patterns are then used as feature vectors to train a decoder model. The decoder decodes the COVID-19 status from cough-breathing patterns.

3.2. Pre-processing

The breathing sub-challenge dataset comprises of spontaneous speech recordings of 49 English speakers, with 16 kHz sampling rate. The speakers had a respiratory belt attached to capture their linear voltage readings corresponding to the changes in thoracic circumference associated with the respiration. The breathing voltage readings are 6 000 normalised values for a duration of 4 minutes per speaker, which corresponds to 25 samples per second or a reading for every 40 msec duration. To establish a correlation with the given breathing patterns, a time domain feature vector of 27 features is extracted for every 40 msec frame of the speech signal. The speech signals are down-sampled to 8 kHz sampling rate, with single channel, 16-bit sample size to extract these features. Figure 3 and Section 3.3 explains the procedure to extract features.

DiCOVA challenge Track 1 and 2 provides imbalanced datasets in five folds. Each fold in Track 1 has 772 COVID-19 negative and 50 COVID-19 positive cough audio samples. In Track 2, each fold consists of counting normal, breathing, and

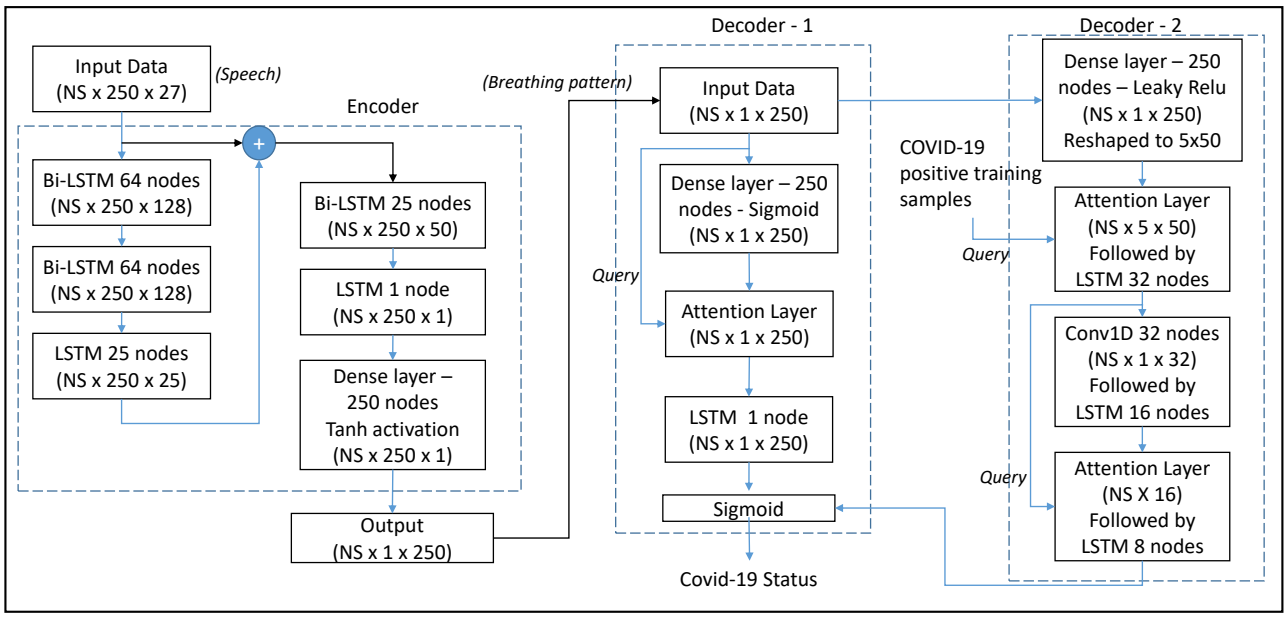

Figure 2: *Encoder-Decoder Architecture. Two different versions of a decoder network are presented. Decoder* 1 *is a simpler network, with which the results on the evaluation set of Track 1 are submitted. Decoder* 2 *is an improved version giving 10 % higher AUC on the Track 1 validation set. NS: Number of steps.*

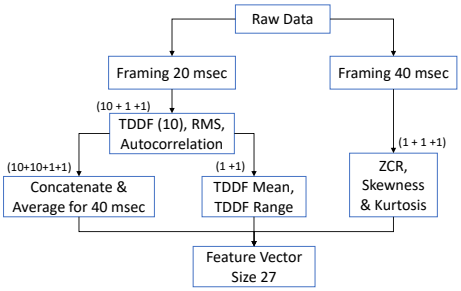

Figure 3: *Feature representation for the training encoder.*

vowel-e pronunciation data. Techniques such as time-stretch, and pitch-change for augmenting audio data might lead to the loss of COVID-19 bio-markers, as they change the audio signal properties. To balance the two classes, the samples from the minority class are augmented by repetition such that we obtain equally numbered samples in both of the classes. To detect the breathing patterns of cough-audio signals, again 27 features are extracted. As the encoder model described in Section 3.4 needs a batch size of 250, the cough features are augmented by repetition till they form an array of size, 250 x 27. The breathing patterns of cough-audio signals, as obtained at the output of encoder, are represented with a series of 250 values.

3.3. Feature Description

Figure 3 explains the steps for extracting 27 time-domain features from the speech signal to train the encoder. As seen in the Figure, Zero Crossing Rate (ZCR), skewness, and kurtosis are extracted from 40 msec frames, and Time Domain Difference Features (TDDF) [27], Root Mean Square (RMS), and frame autocorrelation are extracted for every 20 msec frame of the speech. These TDDFs from 20 msec frame are concatenated and an average is calculated for RMS and autocorrelation.

3.4. Classifier Description

The encoder uses a stacked Bi-directional Long Short Time Memory (Bi-LSTM) architecture to encode the speech signals into breathing patterns. The encoder architecture is as shown in Figure 2. The deep network uses a batch size of 250, and has a skip connection after three layers. The 'tanh' activation at the output layer gives breathing values in the range of -1 to 1.

We explored and present here the details of the two decoder architectures as shown in Figure 2. Decoder 1 uses a dense layer with 'sigmoid' activation and converts the range for breathing values into 0 to 1. An attention layer identifies the significant breathing values using 'tanh' and 'sigmoid' layers outcome. The last layer is again a sigmoid activation which acts as a classifier to detect the (binary) COVID-19 status. Decoder 2 has a 'leaky-ReLU' (Rectified Linear Unit) activation at the input layer and uses a 1-dimensional convolution layer (Conv1D) along with stacked attention and LSTM layers. In this Decoder 2 network, we pass the training samples with COVID-19 positive status as query to the first attention layer. Also, dropout factor of 0.4 is used with the attention and Conv1D layers to avoid overfitting.

4. Results and Conclusion

At the output of the encoder, the Pearson correlation of true values with the predicted values is obtained, where we receive an r value of 0.47 on the devel set. With further observation, it is found that 4 out of 16 devel set files are having a correlation below 0.3, while another 12 had an r value above 0.5, giving an average of 0.57 for the r value while calculating for every file (or speaker), thus showing a drop of 0.1 for the entire data-set.

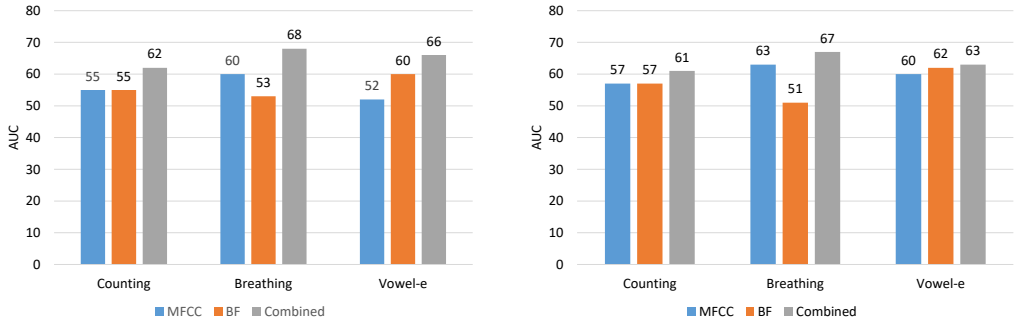

Figure 4: *Track-2 validation (left) & evaluation (right) set average performance for breathing, counting, and vowel-e using MFCCs, BF: breathing features and Comb: feature set combining MFCCs & BF.*

4.1. Track 1 Results

The breathing parameters of the cough audio signals are passed as an input feature vector of length 250 to the decoder. As shown in Figure 2, we explored two different decoder architectures. The result obtained with the Decoder 1 network is submitted at the DiCOVA challenge, in which we achieved an AUC of 64.4 % on the evaluation set and an average AUC of 47.2 % on the validation set of the Track 1 data. With Decoder 2, a further complex attention network, we obtain an absolute improvement of 10 % on the validation set from 47.2 % to 57.4 % AUC[8]. For the Track 1 evaluation set, using Decoder 1, the model gives an average specificity of 40.1 at the sensitivity of 80.4. We made two more submissions to the DiCOVA challenge – Track 1 evaluation set. In the first submission, we trained a RandomForest classifier using the breathing patterns extracted from speech signals. It gave an average AUC of 69.11 % on the validation set, however, a lesser AUC of 60.66 % on the evaluation set.

In the second submission, MFCCs gave an average AUC of 53.84 % on the validation set and an AUC of 55.12 % on the evaluation set of Track 1 using Decoder 1 network. With Decoder 2 network, MFCCs give an average AUC of 51.4 % on validatio set Track 1. As seen in Table 2, breathing features give an absolute improvement of 6 % over MFCCs using Decoder 2. Combining the two feature sets further improve the result to 57.2 % and 61.1 % using Decoder 1 and Decoder 2 network respectively.

4.2. Track 2 Results

We have also evaluated this system's performance on the Track 2 dataset. With the same encoder-decoder (Decoder 1) architecture, average AUC on the five folds of Track 2 validation and evaluation sets are as shown in Figure 4. As seen for both validation and evaluation sets, breathing features extracted from counting and vowel-e audio signals are performing better than that from the breathing audio signals. This seems to be corollary of using breathing features extracted from speech signals for training the decoder. With the complex attention based decoder network (Decoder 2) mentioned in Section 4.1, we could not find major improvement in the Track 2 results. On comparing the performance exhibited by MFCCs on this dataset, using Decoder 1 network, it is seen that again breathing features perform better with an absolute improvement of 2 % for vowel-e data. In case of counting-normal data, both MFCCs and breathing features, have similar performance. MFCCs are found to perform better than breathing features for breathing audio data with an absolute improvement of 12 % on the evaluation set. The feature set combining MFCCs and breathing features, improves the performance across all the modalities.

[8]Note that this result was obtained after the challenge's closure of deadline.

Table 2: *Track-1 performance reported in average AUC. D1: Decoder 1, D2: Decoder 2, BF: Breathing Features, Comb: Combined set of MFCC & BF.*

Set	D1			D2		
%	MFCC	BF	Comb.	MFCC	BF	Comb
Val	53.8	47.2	57.2	51.4	57.4	61.1
Test	55.1	64.4	−−	−−	−−	−−

4.3. Conclusion and Future Work

The work presented in this paper introduces the concept of encoding speech audio signals into breathing patterns. Further, these breathing patterns are used for identification of COVID-19 bio-markers. This is a preliminary attempt to examine the significance of breathing-pattern representation of an audio signal for one of the many possible applications. It is seen that the breathing features outperform MFCCs for cough and vowel-e audio data. In case of counting, both have similar results. In case of breathing audio data, MFCCs are found to perform better. However, the feature set combining both the features throughout performs better than the individual feature set. We encourage researchers to augment this concept with recent deep learning techniques to accomplish better results for speech analysis based applications including detection of COVID-19.

With the availability of more COVID-19 positive data, we would like to augment the dataset and observe its performance. Also, a better performing encoder as reported in [25] can be used to analyse the impact of better correlated breathing patterns on the detection accuracy of COVID-19 positive cough, speech, and breath signals. The early and late fusion techniques can also be tried to uplift the performance. Overall, we find some value in using a pre-trained breathing model, yet, future work will need to combine it elegantly with other modelling.

5. Acknowledgement

The authors gratefully acknowledge support from the German Research Foundation (DFG) under the Reinhart Koselleck-Project AUDI0NOMOUS (grant No. 442218748).

6. References

[1] 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2020, p. 3474–3484. [Online]. Available: https://doi.org/10.1145/3394486.3412865

[2] 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*, vol. 20, p. 100378, 2020.

[3] 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*, vol. 1, pp. 275–281, 2020.

[4] 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.

[5] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "Covid-19 cough classification using machine learning and global smartphone recordings," *arXiv preprint arXiv:2012.01926*, 2020.

[6] R. Dunne, T. Morris, and S. Harper, "High accuracy classification of covid-19 coughs using mel-frequency cepstral coefficients and a convolutional neural network with a use case for smart home devices," *ResearchSquare Preprint*, 2020.

[7] B. W. Schuller, H. Coppock, and A. Gaskell, "Detecting covid-19 from breathing and coughing sounds using deep neural networks," *arXiv preprint arXiv:2012.14553*, 2020.

[8] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. W. Schuller, "End-2-End COVID-19 Detection from Breath & Cough Audio," *BMJ Innovations*, vol. 7, 2021, 8 pages, to appear.

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

[10] T. Dubnov, "Signal analysis and classification of audio samples from individuals diagnosed with covid-19," Ph.D. dissertation, UC San Diego, 2020.

[11] S. Deshmukh, M. A. Ismail, and R. Singh, "Interpreting glottal flow dynamics for detecting covid-19 from voice," *arXiv preprint arXiv:2010.16318*, 2020.

[12] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, and V. Aharonson, "Sars-cov-2 detection from voice," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 268–274, 2020.

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

[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 *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 4811–4815.

[15] 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.

[16] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "Automatic cough detection in acoustic signal using spectral features," in *Proceedings of the 41st Annual International Conference of the Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, 2019, pp. 7153–7156.

[17] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one*, vol. 11, no. 9, 2016.

[18] K. San Chun, V. Nathan, K. Vatanparvar, E. Nemati, M. M. Rahman, E. Blackstock, and J. Kuang, "Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Austin, TX, USA: IEEE, 2020, pp. 1–10.

[19] S. Yadav, M. Keerthana, D. Gope, U. K. Maheswari, and P. K. Ghosh, "Analysis of acoustic features for speech sound based classification of asthmatic and healthy subjects," in *Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 6789–6793.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[21] 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.

[22] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for covid-19," *arXiv preprint arXiv:2005.08579*, 2020.

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

[24] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2042–2046.

[25] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, and A. Karpov, "Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2072–2076.

[26] A. D. MacIntyre, G. Rizos, A. Batliner, A. Baird, S. Amiriparian, A. Hamilton, and B. W. Schuller, "Deep attentive end-to-end continuous breath sensing from speech," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2082–2086.

[27] G. Deshpande, V. S. Viraraghavan, and R. Gavas, "A successive difference feature for detecting emotional valence from speech," in *Proceedings of Speech, Music and Mind 2019, SMM19, Satellite Workshop of Interspeech*, Vienna, Austria, 2019, pp. 36–40.