TCS R&I-SNLP DiCOVA 2021 Challenge System Report

Swapnil Bhosale[§], Upasana Tiwari[§], Rupayan Chakraborty, Sunil Kumar Kopparapu

TCS Research, Tata Consultancy Services Limited, India

email:{bhosale.swapnil2, tiwari.upasana1, rupayan.chakraborty, sunilkumar.kopparapu}@tcs.com

Abstract

Cough sounds as a descriptor have been used for detecting multiple respiratory tract infections based on its intensity, duration of intermediate phase between two cough sounds, repetitions, dry vs wet etc. However, COVID-19 diagnosis using only cough sounds is challenging because of cough being a common symptom among many non COVID-19 health diseases and lack of availability of enough samples belonging to COVID-19 positive subjects. Towards this direction we propose two different approaches. In our first approach, we explore the robustness of multi-domain representation by performing the early fusion over a wide set of temporal, spectral and tempo-spectral handcrafted features, followed by training a Support Vector Machine (SVM) based classifier. In our second approach, we employ a contrastive loss function to learn a latent space from low-level Mel Filter Cepstral Coefficients (MFCCs) where representations belonging to samples having similar cough characteristics are closer. This helps learn representations for the highly varied COVID-negative class (healthy and diseases other than COVID-19), by not limiting its representations fall into a single cluster, rather multiple smaller clusters. Using only the data provided as a part of the challenge, we compare our performance with the provided baseline, our first approach achieves an absolute improvement of 0.74% and 1.07%, whereas our second approach shows an improvement of 2.09% and 3.98% ,over the blind test and validation set, respectively.

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

1. System Description

Unlike detection of other respiratory tract infections, while detecting COVID-19 infections, it is necessary to be able to identify markers that discriminate beyond forced cough (simulated) and ailment cough. The large variations within the COVID-negative (healthy and diseases other than COVID-19) samples, make it difficult to learn representations into two non-overlapping clusters. Moreover, most COVID-19 datasets include high imbalance among the COVID-positive and healthy subjects. Fig. 1 represents the framework of our proposed approach.

1.1. Methodology Overview

In order to address the above mentioned challenges, we propose a few-shot based approach such that the classification task is now converted to a 2-way N-shot learning problem where the ways represent 2 classes, i.e. COVID-positive and COVID-negative, and N shots represents, N audio samples each from both the classes, provided as reference samples. Each query sample is then inferred depending on the distance between the learned representations of query and the provided reference samples. Interestingly, these representations are learned using a episodic triplet loss, which enables the model to learn the intra-clusters among the highly varied COVID-negative class. Moreover, during training the loss in each episode is computed w.r.t to a fixed number of shots for each class, thereby maintaining a balance between the number of samples for both the classes [1] [2]. It is important to note that, in a supervised approach each training sample is equivalent to a single training configuration, i.e. the same sample shown multiple times to a model (during training) will make the model overfit (memorize it), however, in this case a single sample may yield different loss values (and hence different possibilities of attaining convergence), depending on the reference samples (also called as support set) provided with it in each training iteration (also called as episode).

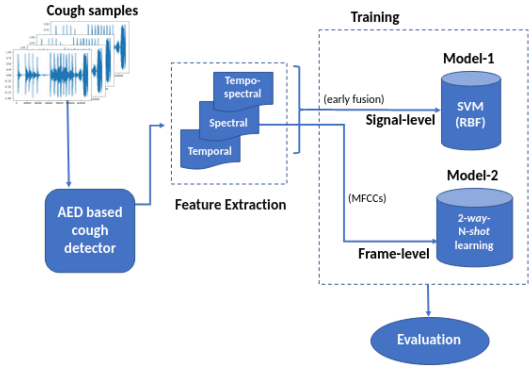


Figure 1: *Framework*

In addition to the above approach, we also followed a feature-based machine learning technique where we explored the robustness of multi-domain cough biomarkers with a simple Support Vector Machine (SVM) classifier. Here, we experimented with a wide set of handcrafted features capturing the temporal, spectral and tempo-spectral descriptors (discussed in detail in Section 1.3) to learn the cough patterns that can discriminate between COVID-positive and COVID-negative samples.

1.2. Pre-processing

Analysing the spectrograms of audio waveforms (see Fig. 2), we observe that majority of the samples show high spectral spread under 8 kHz. Thus, we resampled each audio at 16 kHz using librosa [3] as our audio processing library.

1.2.1. Localizing cough-relevant regions using a pre-trained Audio Event Detection (AED) model

Owing to its crowdsourced nature, the provided data consists of outliers in terms of both, the duration of the samples (see Fig.3) and the amount of undesirable audio content other than cough-related sounds (i.e. background noise, human speech etc) present. Towards this direction, we utilized a pre-trained

[§]Both the authors have an equal contribution

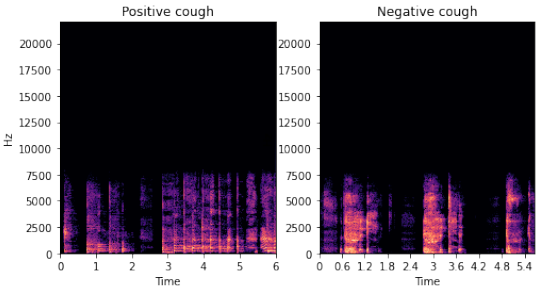
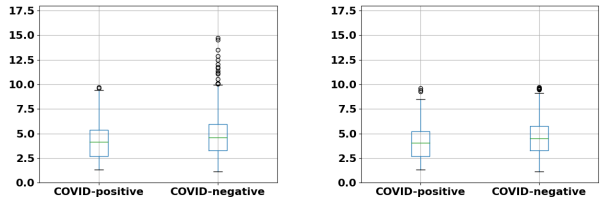


Figure 2: *Spectrogram of a positive and a negative cough signal sampled at 44.1 kHz*



(a) *Original (1040 samples)* (b) *Pre-processed (1073 samples)*

Figure 3: *Variation in duration of samples. The points marked using circles depict the outliers (on the basis of the duration).*

AED model, YAMNet [4], trained on generic audio events (including cough sounds), to localize two types of regions, namely,

- COUGH SOUNDS events in the same subtree of actual 'cough' event from knowledge graph [1], i.e. semantically similar, ex: throat clearing + other events (not in the subtree) but acoustically similar, ex: plop.
- OTHER all audio events other than those mentioned above.

Once individual regions are obtained, a smoothing operation is applied to remove discontinuous regions (regions lesser than 500 ms). We specify two criteria based on the number of detected regions and their corresponding time durations, (a) no COUGH SOUNDS region detected throughout the duration of the audio sample, and (b) detected OTHER region is more than 4 secs. Audio files satisfying these criteria, are further split into smaller audio samples (to be used as individual training samples) based on detected regions. It is important to note, that this even filters smaller duration audio files with undesirable audio content. This gives a a total of 1073 samples, fairly consistent across their duration (see Fig. 3b) as well as inclusion of audio content relevant to cough sounds.

1.3. Feature Description

1.3.1. Discrete Wavelet Transformation (DWT) Based Tempo-spectral Descriptors

DWT is a signal decomposition technique that represents the temporal changes in spectral dynamics of a signal. Cough sounds, being non-stationary in nature, can be better represented by tempo-spectral structure that captures the dynamic changes in signal. Here, we use *db3* as a mother wavelet due to their efficient time frequency localization properties in cough sound analysis [6, 7]. The choice for level of decomposition is based on prominent frequency components of the signal.

[1] please refer to Audioset ontology at `https://research.google.com/audioset/ontology/index.html`.

Table 1: *Performance result of our approach on validation and test set*

Approach	Validation			Test		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Baseline	68.89	81.6	43.42	69.85	80.49	53.65
1	69.96	86.4	39.38	70.59	80.49	45.83
2	72.87	92.0	39.79	71.94	80.49	47.40

Since cough sounds are known to have more energy in lower frequencies (varying from 20 Hz - 50 Hz in different studies) [8, 9, 10] , we decomposed the signal till 10th level using PyWavelets tool [11]. Thereafter, we extract the set of six features for each frequency band, namely, Energy, Entropy, Root Mean Square (RMS), Recoursing Energy Efficiency (REE), Logarithmic REE (LREE), Absolute Logarithmic REE (ALREE) [12]. Thus, we get a 60-dimensional feature vector.

1.3.2. Spectral descriptors

In order to capture the spectral properties in a cough sound we first compute 64 low level descriptors (LLDs) with 20 ms of window length and 10 ms of overlap, followed by computing the delta and delta-delta coefficients in order to extract the temporal mutual information among the adjacent frames. A list consists of the LLDs: 13 Mel-Frequency Cepstral Coefficients (MFCC) components, Zero-crossing, Spectral Centroid, Roll-off Frequency, Spectral Flux, Spectral entropy, Spectral Spread, 12 Chroma components, Jitter (ratio, percentage, factor), 26 Mel Spectrograms, Fundamental Frequency (F_0), Log energy, Entropy. The usage of these features are well studied in previous works on cough sound detection [13, 14, 15, 16, 17]. All LLDs are smoothed over time with a symmetric moving average filter with length of 3 frames. To capture the distributions beyond the mean, we further computed high-level descriptors (HLDs), statistical features over LLDs. A complete list of 25 HLDs is: mean, median, range, maximum, minimum, position of minimum and maximum, percentile, 1st, 2nd and 3rd quartile, interquartile range, standard deviation, variance, skewness, kurtosis, linear and quadratic regression coefficients. In total, we get a 4800-dimensional feature vector. We use pyAudioAnalysis library to implement this feature set [18]

1.3.3. Temporal descriptors

To capture the temporal characteristics of a cough signal, we extract the hjorth parameters (mobility and complexity) which have been proved to be a powerful biomarker in respiratory sounds [19]. Motivated by the wide usage of fractal dimension in lung sound analysis [20, 9], we compute the Petrosian Fractal Dimension (PFD) and Higuchi Fractal Dimension (HFD) of cough signals. Furthermore, we use Detrended Fluctuation Analysis (DFA) that captures the self-similarity within a time-series over a longer period, whose application has been explored in respiratory sounds [21]. We use an open source python module, PyEEG [22], to compute this 5-dimensional feature vector.

1.4. Classifier Description

- Approach-1 : In this approach, we perform an early fusion over a wide set of handcrafted features (as discussed in Section 1.3) to get a multi-domain representation of a cough signal. This results into 4865-dimensional (60+4800+5) feature vector. Further, we train a SVM with RBF kernel on the resultant feature vector. To prevent biased learning as a result of high imbalance in the data (COVID-positive=50, COVID-negative=772), we assign a higher weight to the minority class and lower to the majority class within the cost function. For this,

we use an utility provided by python's $sklearn$ which automatically adjusts weights inversely proportional to class frequencies in the data.

- Approach-2 : In this approach we use features similar to the baseline approach [23], i.e. 39 dimensional MFCCs combined with the delta and delta-delta features. Additionaly, for an audio sample x with frame-level feature vectors, $x_i \in \mathcal{R}^{117}$, features from w succeeding and w preceeding frames are concatenated. In our experiments, we use w=3 and thus, our input to the model is, $\bar{x_i} \in \mathcal{R}^{(7*117)}$ (7:{3+1+3}) vector. The embedding block (referred as $f_\phi(.)$ in [2]) consists of a stack of 2 dense layers, with 800, 512 hidden units each with ReLU activation. During training, in each episode a triplet loss is computed using the output from the last dense layer as the latent representation. Post-training, the output of last dense layer is used as embeddings to train a logistic regression classifier. Similar to the baseline approach, our model provides the frame-level scores which are then averaged to obtain a prediction for the entire audio sample. We use number of shot as 5 (i.e. each training episode consists of 5 reference samples from each class) and margin = 0.5, where margin defines the maximum difference between the euclidean distances between anchor and positive sample, and between anchor and negative sample (discussed in [2]).

1.5. Results

We compare our approaches with the provided baseline (replicated using the provided scripts). As can be seen from table 1, our both approaches surpass the baseline when evaluated over the validation samples as well as the blind test samples. We hypothesize, this is primarily because of two reasons, firstly, because of the robustness of handcrafted multi-domain representation in our first approach and the balanced episodes generated as a part of the few-shot pre-training accompanied by the triplet loss function which enables the intra class learning within the highly varied COVID-19 negative class in our second approach. Secondly, we use a balanced subset of the provided validation folds (since the provided validation folds are skewed), to decide on an early stopping criteria that guides the model convergence. It is important to note that in both our approaches we use the samples provided as a part of the DiCOVA challenge, and do not use any external data. The performance of our approaches could be further improved with access to larger databases with a fairly equal class distribution.

2. References

[1] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, 2019.

[2] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Semi supervised learning for few-shot audio classification by episodic triplet mining," *arXiv preprint arXiv:2102.08074*, 2021.

[3] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8. Citeseer, 2015.

[4] Tensorflow, "Yamnet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, 2020.

[5] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP 2017*.

[6] K. Kosasih, U. R. Abeyratne, V. Swarnkar, and R. Triasih, "Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 62, 2014.

[7] T. H. Pingale and H. Patil, "Analysis of cough sound for pneumonia detection using wavelet transform and statistical parameters," in *IEEE International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017.

[8] K. K. Lee, S. Matos, K. Ward, G. F. Rafferty, J. Moxham, D. H. Evans, and S. S. Birring, "Sound: a non-invasive measure of cough intensity," *BMJ open respiratory research*, vol. 4, 2017.

[9] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, vol. 2, 2008.

[10] 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, 2020.

[11] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "Pywavelets: A python package for wavelet analysis," *Journal of Open Source Software*, vol. 4, 2019.

[12] M. Murugappan, N. Ramachandran, Y. Sazali *et al.*, "Classification of human emotion from eeg using discrete wavelet transform," *Journal of Biomedical Science and Engineering*, vol. 3, no. 04, 2010.

[13] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one*, vol. 11, 2016.

[14] 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 *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[15] S. Z. H. Naqvi and M. A. Choudhry, "An automated system for classification of chronic obstructive pulmonary disease and pneumonia patients using lung sound analysis," *Sensors*, vol. 20, 2020.

[16] T. Drugman, J. Urbain, N. Bauwens, R. Chessini, C. Valderrama, P. Lebecque, and T. Dutoit, "Objective study of sensor relevance for automatic cough detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, 2013.

[17] M. J. Rahman, E. Nemati, M. Rahman, K. Vatanparvar, V. Nathan, and J. Kuang, "Efficient online cough detection with a minimal feature set using smartphones for automated assessment of pulmonary patients," in *International Conference on Ambient Computing, Applications, Services and Technologies*, 2011.

[18] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, 2015.

[19] A. Rizal, R. Hidayat, and H. A. Nugroho, "Lung sound classification using empirical mode decomposition and the hjorth descriptor," *American Journal of Applied Sciences*, vol. 14, 2017.

[20] A. Rizal, H. A. Nugroho, and R. Hidayat, "Fractal dimension for lung sound classification in multiscale scheme." *Journal of Computer Science*, vol. 14, 2018.

[21] G. C. Donaldson, T. A. Seemungal, J. R. Hurst, and J. A. Wedzicha, "Detrended fluctuation analysis of peak expiratory flow and exacerbation frequency in COPD," *European Respiratory Journal*, vol. 40, 2012.

[22] F. S. Bao, X. Liu, and C. Zhang, "Pyeeg: an open source python module for eeg/meg feature extraction," *Computational Intelligence and Neuroscience*, vol. 2011, 2011.

[23] 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.