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

Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19

Rubén Solera-Ureña[1], *Catarina Botelho*[1,2], *Francisco Teixeira*[1,2], *Thomas Rolland*[1,2], *Alberto Abad*[1,2], *Isabel Trancoso*[1,2]

[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, University of Lisbon (IST-UL), Portugal

rsolera@hlt.inesc-id.pt, {alberto.abad,isabel.trancoso}@inesc-id.pt

Abstract

In the last months, there has been an increasing interest in developing reliable, cost-effective, immediate and easy to use machine learning based tools that can help health care operators, institutions, companies, etc. to optimize their screening campaigns. In this line, several initiatives emerged aimed at the automatic detection of COVID-19 from speech, breathing and coughs, with inconclusive preliminary results. The ComParE 2021 COVID-19 Cough Sub-challenge provides researchers from all over the world a suitable test-bed for the evaluation and comparison of their work. In this paper, we present the INESC-ID contribution to the ComParE 2021 COVID-19 Cough Sub-challenge. We leverage transfer learning to develop a set of three expert classifiers based on deep cough representation extractors. A calibrated decision-level fusion system provides the final classification of coughs recordings as either COVID-19 positive or negative. Results show unweighted average recalls of 72.3% and 69.3% in the development and test sets, respectively. Overall, the experimental assessment shows the potential of this approach although much more research on extended respiratory sounds datasets is needed.

Index Terms: COVID-19 detection, Transfer Learning, Cough Representations, X-vector embeddings, CNN embeddings, PASE+ features

1. Introduction

The COVID-19 respiratory disease was declared a pandemic by the WHO in 11 March 2020 and has had dramatic personal, societal and economical consequences that extend until today. Until more efficient treatments are developed and a high percentage of population has access to vaccines, hygienic respiratory practices, social distancing and massive screening campaigns for early diagnosis are the most effective ways to contain it. Clinical diagnosis of COVID-19 relies on RT-PCR and antigen tests. However, these procedures present several disadvantages: significant monetary cost, intrusive and in-person collections of samples performed by professionals that strain public health systems, diagnosis delays due to saturation of laboratories, etc. For all these reasons, there is an increasing interest in developing reliable, cost-effective, immediate and easy to use tools that can help health care operators, institutions, companies, etc. to optimize their screening campaigns.

The automatic detection of diseases that affect or are reflected in speech or respiratory sounds is an emerging research field with potentially significant impact in medicine. In this line, the automatic detection of COVID-19 from speech, breathing or coughing can assist in mass-testing of populations by providing a preliminary screening tool that does not require the intervention of experts and may be remotely available in a worldwide scale. Several investigations based on different datasets have appeared during the last months, with varied and inconclusive preliminary results. Initiatives such as the ComParE 2021 COVID-19 Cough Sub-challenge (CCS) [1] provide researchers from all over the world a valuable space to share ideas and findings and to compare their results in a common test-bed.

In this work, we present the INESC-ID contribution to the ComParE 2021 COVID-19 Cough Sub-challenge, which is based on two fundamental pillars. We leverage **transfer learning** to develop a set of COVID-19 classification subsystems based on **deep cough representation extractors** (TDNN-F and CNN embeddings, and PASE+ features). Individual decisions of the three experts are fed to a calibrated decision-level fusion system. This ensemble of expert subsystems based on cough representations is expected to produce well-calibrated log-likelihood scores over a wide range of operating points. The output can be more easily interpreted by a human expert and incorporated into the decision making process.

The rest of this work is structured as follows: Section 2 presents a review of previous work on automatic detection of COVID-19 from speech and respiratory sounds. In Sections 3 and 4, we briefly describe the datasets used in this study and present our proposal for COVID-19 detection from coughs. Experimental results are presented in Section 5, and the paper ends with conclusions in Section 6.

2. Related work

Current research on automatic detection of COVID-19 from speech or respiratory sounds is based on previous work that show how they are differently affected by distinct respiratory diseases and can thus be used to detect pertussis, asthma, pneumonia, tuberculosis, among others [2]. Although there are no conclusive evidences for the moment, preliminary work points out to specific signatures of COVID-19 in coughs and speech that could be used to detect it even in apparently asymptomatic individuals and to distinguish it from other common respiratory illnesses. To address the scarcity of COVID-19-labeled data, many of these approaches rely on transfer learning, data augmentation, and class balancing techniques.

Most of previous work is based on convolutional neural networks (CNN) in different settings. In [3], a pre-trained VGGish model [4] is used as a generic audio feature extractor. In [5] and [6], CNNs are firstly trained for cough detection and then fine-tuned for COVID-19 detection. An ensemble with two DNN-based experts and a CNN that is directly trained from scratch to detect COVID-19 has been proposed in [7]. Other types of models are also used for COVID-19 detection. Pinkas

http://dx.doi.org/10.21437/Interspeech.2021-1702

et al. [8] present a three-stage system consisting of a self-supervised attention-based transformer that generates embeddings from the inputs, a set of recurrent neural network (RNN) classifiers specialized in one or multiple types of input, and a final support vector machine (SVM) meta-model. Some works also leverage the information about self-reported symptoms. In [9], they are encoded as one-hot vectors and combined either at the feature- or decision-levels with traditional speech features. Pal et al. [10] describe a system consisting of a DNN that generates cough embeddings from traditional handcrafted features and a transformer-based self-attention network that generates symptoms- and demographic-data embeddings, which are concatenated and fed to a fully-connected layer to produce the final decision.

3. Corpora description

The imperative need for COVID-19 datasets of speech and respiratory sounds has been partially addressed by several initiatives, such as those conducted by the University of Cambridge [11, 3, 9], the École Polytechnique Fédérale de Lausanne [12, 13], the Indian Institute of Science, Bangalore [14, 15], the Carnegie Mellon University [16], the MIT AudioID Laboratory [17], and the Virufy initiative [18], besides our own on-going efforts at INESC-ID/IST-Univ. Lisbon [19]. This work uses the first two datasets. The COVID-19 COUGH (C19C) corpus [3, 9] is provided in the ComParE 2021 COVID-19 Cough Sub-Challenge as a test-bed for the evaluation of the systems presented by the participants. The COUGHVID corpus [13] is used to train and/or fine-tune the transfer learning-based cough representation extractors described in Section 4.1. For both datasets, silence segments are removed using a modified version of a cough segmenter developed by the COUGHVID team[1].

3.1. COVID-19 COUGH (C19C) corpus

The COVID-19 COUGH (C19C) corpus is a curated subset of the Cambridge COVID-19 Sound database [3, 9], a crowd-sourced corpus with examples of breathing, coughs and speech recorded ”in-the-wild”. The C19C corpus contains 725 cough recordings from 397 participants and self-reported COVID-19 status labels (positive/negative), distributed in three speaker-independent, gender-balanced subsets: train (71 positives/215 negatives), development (48 positives/183 negatives) and a blind test set (208 samples).

In our preliminary analysis of this corpus, we noticed that some files present a reduced bandwidth of 4 kHz, hypothetically corresponding to audio samples originally recorded at a sampling rate of 8 kHz. Namely, 13, 8 and 8 narrow-band files were detected in the train, development and test subsets, respectively. This condition certainly reflects the reality of many real-world applications. However, we noticed that all the narrow-band recordings in the train and development subsets correspond to the COVID-19 positive class. From our analysis of the baselines and our own systems, we consider that this might be affecting their performance by making the training process pay attention to this spurious condition. For this reason, we decided to create a second version of the dataset by removing all the narrow-band recordings in the original train and development subsets, even at the cost of reducing the number of COVID-19 positive examples. The resulting dataset, denoted as ”C19C$_{\text{fullband}}$”, contains 273 samples in the train subset (58 positives/215 negatives) and 223 in the development subset (40 positives/183 negatives). The test subset is kept untouched so as to stick to the original definition and evaluation conditions of the challenge.

3.2. COUGHVID corpus

The COUGHVID corpus [13] is a non-curated, publicly open dataset. Recordings were performed using a lossy codification and present a variety of conditions (sampling rate, bandwidth, number of channels, and quality). Volunteers recorded their coughs and reported their COVID-19 status (positive/symptomatic/healthy), age, gender, and medical condition. A small fraction of the dataset was annotated by expert pulmonologists with information regarding type of cough (wet/dry/inconclusive), presence of audible dyspnea, wheezing, stridor, choking, and nasal congestion, diagnosis (upper/lower respiratory tract infection/obstructive lung disease/COVID-19/healthy cough), and severity (healthy cough/mild/severe). It comprises 27550 recordings, 15125 of which are classified as coughs by an automatic cough detector developed by the COUGHVID team (probability of cough $\geq$0.8). 10763 of them have self-provided gender and COVID-19 status annotations, distributed as following: 680 COVID-19 positives (395 male/285 female), 8270 healthy (5632 male/2638 female), and 1813 symptomatic (1114 male/699 female).

4. Methodology

4.1. Transfer learning-based representations

Data augmentation and class balancing with synthetic data are common techniques in many machine learning areas. However, they must be applied very carefully in the COVID-19 detection task, since it is still not clear how this disease affects speech and respiratory sounds, and whether the use of these techniques could alter specific biomarkers of COVID-19. Thus, in this work, we decided to just leverage transfer learning to develop a set of deep cough feature extractors that generate TDNN-F embeddings, CNN embeddings and PASE+ features.

4.1.1. TDNN-F embeddings

X-vector embeddings are current state-of-the-art speaker representations [20]. Not only do they outperform previously proposed representations (e.g., *d-vectors* [21], *i-vectors* [22]), but they have also been shown to contain enough information to be successfully applied to other paralinguistic and extra-linguistic tasks. In particular, several authors have applied *x-vectors* to the automatic detection of obstructive sleep apnea [23], Parkinson’s disease [24, 25], Alzheimer’s disease [26], and depression [24]. The success of *x-vector* representations motivated us to explore their applicability to coughs, hypothesizing that *x-vector*-like embeddings trained using coughs, instead of speech, are capable to encode relevant information about the cough signal and to transfer medically meaningful information.

Based on this hypothesis, we implemented a reduced version of the Factorized TDNN (TDNN-F)-based network [27], proposed for speaker recognition by Villalba et al. [28]. The network architecture is summarized in Table 1. Each block, with the exception of the statistics pooling layer, corresponds to a TDNN, TDNN-F or dense layer, followed by a Leaky-ReLU activation, a batch normalization layer and a dropout layer. Cough embeddings are 128-dimensional vectors obtained at the output of the final dense layer (layer block 7).

We considered that available COVID-19-labeled data was still scarce for training an architecture like this one from scratch. For this reason, training of the network was performed in two

[1] https://c4science.ch/diffusion/10770

Table 1: *TDNN-F embedding network architecture*

Layer	Layer type	Ctx. 1	Ctx. 2	Size	Inner size
1	TDNN	t-2:t+2	-	512	-
2	TDNN-F	t-2,t	t,t+2	1024	256
3	TDNN-F	t	t	1024	256
4	TDNN-F	t-2,t	t,t+2	1024	256
5	Dense	t	-	2048	-
6	Stats. Pool.	full seq.	-	2×2048	-
7	Dense (embedding)	-	-	128	-

stages. In a first step, the network was trained for age estimation (regression) and gender classification (binary classification). To this end, we used the subset of the COUGHVID dataset for which information about the age and gender of the speakers is available, split into train (7145 recordings), development (1531), and test (1531) subsets. We hypothesized that, similarly to speaker identity, a representation that allows us to estimate age and gender will also carry health-related information. In a second step, we leveraged the annotations by expert pulmonologists to fine-tune the previously trained network in a multi-task classification setting (five tasks): *cough type* (dry/wet); presence of *dyspnea* (yes/no); presence of *wheezing* (yes/no); *diagnosis* (healthy cough/lower tract infection/upper tract infection/obstructive disease/COVID-19); and *severity* (healthy cough/mild/severe). The reason behind this fine-tuning step is the fact that these tasks are much closer to COVID-19 classification than age and gender. Just a fraction of the dataset was annotated by one to four experts. Thus, experts' annotations were aggregated into a single label by majority voting, discarding those recordings where a tie occurred. This resulted in 1285 recordings for train, 265 for development, and 256 for test.

The TDNN-F network was implemented in Pytorch. Training and fine-tuning used the Adam optimizer with a weighted sum of the loss functions of the involved tasks. Classification tasks used binary cross-entropy loss, while the regression task used the mean squared error loss. In the second stage, each class was weighted with the inverse of its frequency in the training subset to address the unbalanced nature of this dataset. Input features consisted in 30 MFCCs computed every 10 ms from 25 ms-length frames. Cepstral mean and variance normalization was applied using a 3 s-length sliding window. These steps were performed following the *egs/voxceleb/v2* Kaldi recipe [29].

4.1.2. CNN embeddings

Reported CNN-based approaches for COVID-19 detection suffer from some limitations such as the use of CNNs as generic audio feature extractors not tuned to this task, or the use of relatively small datasets to train/fine-tune those networks. Here, we leverage transfer knowledge from a model (VGGish) trained on a vast corpus for audio classification, and then fine-tune it for COVID-19 detection using the COUGHVID dataset.

The VGGish model [4] is an adaptation for audio classification of the VGG network [30]. It comprises four blocks, each one with one or two convolutional layers followed by a pooling layer. The output of the last pooling layer is flattened and followed by two fully-connected (FC) layers and an output layer. This model was originally trained with 5.4M hours of YouTube data. In this work, we used a simplified version as shown in Table 2. Layers 1 to 7 correspond to the original architecture, with pre-trained weights from the original model. The top-level FC layers in the original model were here substituted by lower-dimensionality layer 9 to facilitate fine-tuning with limited data. Layer 8 flattens and reduces dimensionality.

Table 2: *Architecture of the simplified VGGish*

Layer	Layer type	Output shape
1	Conv2D	(96,64,64)
2	MaxPooling2D	(48,32,64)
3	Conv2D	(48,32,128)
4	MaxPooling2D	(24,16,128)
5-6	Conv2D (x2)	(24,16,256)
7	MaxPooling2D	(12,8,256)
8	GlobalAvg.Pooling2D	(256)
9	FullyConnected	(64)
10	FullyConnected	(1)

This CNN architecture is used in this work in two different settings. In both cases, the generated embeddings are 256-dimensional vectors (output of layer 8). When used as a pre-trained generic feature extractor, weights are directly loaded from the original model. Layers 9 and 10 are included on top to allow for fine-tuning for COVID-19 detection using a balanced subset of the COUGHVID dataset (680 positive and 680 negative cough recordings; 80% of this subset is used for training and 20% for development). In this case, the weights of these two layers are initialized randomly and the whole CNN is fine-tuned for 150 epochs using cross-entropy loss, the Adam optimizer with a learning rate of 10^{-5} and a batch size of 64. Inputs to the VGGish network consisted in log Mel-spectrogram features computed every 0.24 s from 0.96 s-length segments.

4.1.3. PASE+ features

Recent representation learning approaches have focused on self-supervised learning where targets are learned directly from the signal. The problem-agnostic speech encoder (PASE+) model [31, 32] encodes raw speech and passes it to multiple regressors and discriminators (workers). Each worker provides additional prior information to the encoder by giving a different view of the raw input signal. PASE+ features are extracted at the end of the encoder after joint training of the encoder and the workers.

The PASE+ architecture begins with a shared encoder consisting of a SincNet-based layer, followed by seven blocks with a one-dimensional convolutional layer, batch normalization and PreLU activation. It follows a Quasi-RNN layer that learns long-term dependencies. The final encoder representation is the sum of the linearly projected intermediate features computed by each of the seven convolutional blocks, using skipped connections, and the Quasi-RNN output. This allows the transfer of information from different levels of abstraction as well as the improvement of gradient flows. The output of the encoder is fed to twelve workers. Each worker is a small feed-forward neural network that solves a single task. Some tasks consist in reconstruction by minimizing the mean square error of the frame and its corresponding six neighbouring context frames. The reconstructed targets are waveform, log power spectrum (LPC), MFCCs, prosody, filter banks (fbanks) and gammatone. PASE+ has also workers where the targets are estimated over longer windows (200 ms instead of 25 ms) for LPS, MFCC, fbanks and gamatone. Finally, two workers are binary discrimination tasks: Info max local and Info max global.

In our experiments, we used two different PASE+ feature extractors. The first PASE+ extractor was pre-trained for 150 epochs on Librispeech [33]. The second one was trained from scratch on COUGHVID data for 150 epochs. We used a batch size of 64 with a learning rate of 0.0005 and 0.001 for the workers and the encoders, respectively. 256-dimensional feature vectors are extracted for each frame every 10 ms.

Table 3: *Performance results (unweighted average recall-UAR) on the COVID-19 COUGH (C19C) corpus*

System	dev	$dev_{fullband}$	$test$
ComParE 2021 CCS Sub-challenge Baseline			
openSMILE	61.4	53.0	65.5
openXBOW$_{2000}$	64.7	56.5	72.9
DeepSpectrum+SVM	63.3	57.3	64.1
auDEEP$_{\text{-60 dB}}$	67.6	57.3	67.6
End2You	61.8	-	64.7
Fusion of Best	-	-	73.9
TDNN-F Embeddings			
Trained COUGHVID$_{\text{Step1}}$	68.8	63.6	-
Fine-tuned COUGHVID$_{\text{Step2}}$	68.1	62.3	-
CNN Embeddings			
Pre-trained YouTube	66.9	62.4	-
Fine-tuned COUGHVID	71.2^{+}	65.6	62.3^{+}
PASE+ Features			
Trained Librispeech	63.1	61.7	-
Trained COUGHVID	67.4	$\mathbf{66.8}^{+}$	64.1^{+}
Calibrated Fusion			
Fusion of experts	$\mathbf{72.3}^{+}$	66.1	69.3^{+}

4.2. COVID-19 condition classification

Given the limited amount of COVID-19 data available for training our system, our approach leverages transfer learning to obtain rich representations of cough, as described before. Thus, complexity of the system is transferred from the back-end (powerful complex model) to the front-end (rich representation extractors). In this work, three SVMs are used on top of the TDNN-F embeddings, CNN embeddings and PASE+ features, respectively, to produce expert decisions. Given the different nature of these representations, their respective pipelines are slightly different. File-wise TDNN-F embeddings are directly fed to the SVM. The CNN-based extractor generates a sequence of embeddings for each recording, computed from cough segments of 0.96 s with a shift of 0.24 s. Here, the sequence of embeddings is fed to the SVM and a final decision is taken by majority voting. PASE+ features are generated every 10 ms, with a receptive field of about 150 ms. In this case, an average feature vector computed across the whole sequence of features is fed to the SVM classifier. These three SVMs were trained on both the C19C and C19C$_{\text{fullband}}$ datasets. Different kernels (linear/RBF), data normalizations (zero mean and unit variance/[0,1] range) and class balancing methods (none/downsampling of the majority class/class weighting) were explored. The optimal configuration and hyperparameters were determined based on development results using a grid-search. Finally, system dependent scaling factors (and an offset) are estimated through linear logistic regression on the development subsets to combine the soft decisions of each individual system. The regression approximates log-likelihood ratios, thus, a theoretically determined decision threshold can be used for making hard decisions.

5. Experimental results and analysis

Table 3 shows performances of the ComParE 2021 CCS baselines and our own system. All systems where trained separately on both the C19C and C19C$_{\text{fullband}}$ train subsets and evaluated on the corresponding *dev* and *dev$_{fullband}$* subsets. Reported test results correspond to our best individual systems trained on the C19C and C19C$_{\text{fullband}}$ datasets (marked with $^{+}$). All results are expressed in terms of the unweighted average recall (UAR).

From the results, we can see that our proposal achieves competitive performance compared to the baseline systems. In the case of the TDNN-F x-vector embeddings-based expert, training of the network on a gender classification and age regression task (step 1) with COUGHVID data achieves a UAR of 63.6% on *dev$_{fullband}$*. Fine-tuning performed in step 2 using a multi-task setting does not seem to enrich the cough representations. We hypothesize that the reason is some degree of overfitting for some of the subtasks for which not enough data is available. CNN embeddings pre-trained on YouTube videos show a reasonable performance, which improves to 65.6% after fine-tuning using COVID-19-specific data. The best performance among the three expert systems is achieved by that based on PASE+ features trained with COUGHVID data, with development and test UAR of 66.8% and 64.1%, respectively. This is the feature extraction method that benefits most from the use of COVID-19-specific data, with an absolute improvement of the UAR of 5.1% with respect to the PASE+ extractor trained on the larger Librispeech dataset. A comparison of the results obtained by the PASE+-based system on the *dev* and *dev$_{fullband}$* subsets point out to a higher robustness of these features.

The last row shows the results obtained from the fusion of the best x-vector (Trained COUGHVID$_{\text{Step1}}$), CNN (Fine-tuned COUGHVID), and PASE+ (Trained COUGHVID) experts. The system trained on the C19C datasets achieves 72.3% UAR on development (1.1% absolute improvement from the best expert) and 69.3% on test. Reported underperformance on the C19C$_{\text{fullband}}$ deserves more analysis, but might be due to the low number of COVID-19 positive examples.

6. Conclusions and future work

In this paper, we present the INESC-ID contribution to the ComParE 2021 COVID-19 Cough Sub-challenge. We leverage transfer learning from related tasks and datasets to develop a set of three expert classifiers based on deep cough representation extractors: TDNN-F embeddings, CNN embeddings, and PASE+ features. Our results show competitive performance compared to the baseline systems, although they are still far from those required to become a reliable tool to assist COVID-19 screening. The reduced amount of data and possible spurious effects, like the bandwidth one reported, suggest a cautious interpretation of the results.

Larger datasets will allow us to learn better cough representations, as well as to complement them with better-suited back-end classifiers. Recurrent neural networks combined with attention mechanisms appear as an appropriate way to capture temporal dynamics of cough recordings and focus attention on the most relevant segments. It would also be interesting to explore how the multi-task TDNN-F network-based cough embeddings compare to embeddings extracted with the same network architecture trained for cough-based speaker identification. Finally, the performance exhibited by the PASE+ features nominate them as a good alternative input representation to MFCCs for the TDNN-F x-vector embeddings extractor.

7. Acknowledgements

This work has been supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), under projects UIDB/50021/2020 and CMU/TIC/0069/2019 and grant SFRH/BD/149126/2019, and by EU funds through Horizon 2020 R&I programme, under the Marie Sklodowska-Curie Grant Agreement No. 766287.

8. References

[1] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. M. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *Proceedings of INTERSPEECH 2021*, Brno, Czechia, Sept. 2021, to appear.

[2] 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, pp. 1–20, Sept. 2016.

[3] 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*, Virtual Event, CA, USA, 2020, p. 3474–3484.

[4] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[5] 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," *preprint arXiv:2009.08790*, 2020.

[6] 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.

[7] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, and A. Khanzada, "Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough," *preprint arXiv:2011.13320*, 2021.

[8] 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.

[9] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8328–8332.

[10] A. Pal and M. Sankarasubbu, "Pay Attention to the Cough: Early Diagnosis of COVID-19 Using Interpretable Symptoms Embeddings with Cough Sound Signal Processing," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, ser. SAC '21, 2021, p. 620–628.

[11] University of Cambridge, UK. COVID-19 Sounds App. [Online]. Available: https://covid-19-sounds.org

[12] École polytechnique fédérale de Lausanne, Switzerland. COUGHVID. [Online]. Available: https://coughvid.epfl.ch

[13] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *preprint arXiv:2009.11644*, 2020.

[14] Indian Institute of Science, Bangalore, India. Project Coswara. [Online]. Available: https://coswara.iisc.ac.in

[15] 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 Interspeech 2020*, Shanghai, China, 2020, pp. 4811–4815.

[16] Carnegie Mellon University, PA, USA. COVID Voice Detector. [Online]. Available: https://cvd.lti.cmu.edu

[17] MIT AudioID Laboratory, Cambridge, MA, USA. MIT Covid-19 Initiative. [Online]. Available: https://opensigma.mit.edu

[18] Virufy initiative. [Online]. Available: https://virufy.org

[19] Instituto Superior Técnico-University of Lisbon and INESC-ID Lisbon, Lisbon, Portugal. Detecção da COVID-19 a partir de tosse e fala. [Online]. Available: https://www.inesc-id.pt/covid19

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[21] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in dnn-based speech synthesis using d-vectors." in *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 3404–3408.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[23] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Álvarez, and L. A. Hernández-Gómez, "Modeling Obstructive Sleep Apnea Voices Using Deep Neural Network Embeddings and Domain-Adversarial Training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 240–250, 2019.

[24] J. Correia, F. Teixeira, C. Botelho, I. Trancoso, and B. Raj, "The In-the-Wild Speech Medical Corpus," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021, pp. 6973–6977.

[25] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using X-Vectors to Automatically Detect Parkinson's Disease from Speech," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1155–1159.

[26] S. Zargarbashi and B. Babaali, "A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language," *preprint arXiv:1910.00330*, 2019.

[27] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Proceedings of Interspeech 2018*, Hyderabad, India, 2018, pp. 3743–3747.

[28] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hawaii, US, Dec. 2011.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, 2015.

[31] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in *Proceedings of Interspeech 2019*, Graz, Austria, 2019, pp. 161–165.

[32] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-Task Self-Supervised Learning for Robust Speech Recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6989–6993.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.