Deep Learning against COVID-19: Respiratory Insufficiency Detection in Brazilian Portuguese Speech

Edresson Casanova[1], Lucas Gris[2], Augusto Camargo[3], Daniel da Silva[2], Murilo Gazzola[1], Ester Sabino[4], Anna S. Levin[4], Arnaldo Candido Jr[2], Sandra Aluisio[1], Marcelo Finger[3†]

[1]Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos/SP, Brazil
[2]Federal University of Technology – Paraná, Medianeira/PR, Brazil
[3]Instituto de Matemática e Estatítica, University of São Paulo, São Paulo/SP, Brazil
[4]Faculdade de Medicina, University of São Paulo, São Paulo/SP, Brazil

Abstract

Respiratory insufficiency is a symptom that requires hospitalization. This work investigates whether it is possible to detect this condition by analyzing patient's speech samples; the analysis was performed on data collected during the first wave of the COVID-19 pandemic in 2020, and thus limited to respiratory insufficiency in COVID-19 patients. For that, a dataset was created consisting of speech emissions of both COVID-19 patients affected by respiratory insufficiency and a control group. This dataset was used to build a Convolution Neural Network to detect respiratory insufficiency using speech emission MFCC representations. Methodologically, dealing with background noise was a challenge, so we also collected background noise from COVID-19 wards where patients were located. Due to the difficulty in filtering noise without eliminating crucial information, noise samples were injected in the control group data to prevent bias. Moreover, we investigated (i) two approaches to address the duration variance of audios, and (ii) the ideal number of noise samples to inject in both patients and the control group to prevent bias and overfitting. The techniques developed reached 91.66% accuracy. Thus we validated the project's Leading Hypothesis, namely that it is possible to detect respiratory insufficiency in speech utterances, under real-life environmental conditions; we believe our results justify further enquiries into the use of automated speech analysis to support health professionals in triage procedures.

1 Introduction

This work started as part of the academic initiative to help in the effort to deal with the COVID-19 pandemic in a severely affected region in Brazil. COVID-19 is an infectious disease caused by the virus SARS-CoV-2. This illness is mainly associated to severe acute respiratory syndrome, although it is harmful to other organs, like heart, kidney and brain. About 82% of cases are mild or moderate, while the rest are severe or grave, demanding hospitalization or intensive care. The most vulnerable groups are people over the 60's, and people with specific medical conditions such as diabetes, obesity, hypertension and heart disease. According to WHO[1], in August 3 2020, more than 19.2 million people in the world had contracted COVID-19, with a Case Fatality Ratio of CFR=2.8%. Respiratory Insufficiency (RI) is a symptom that requires hospitalization, which is aggravated due to a frequent COVID-19 condition called *silent hypoxia*, low blood oxygen concentration without breath shortness (Tobin et al., 2020).

This work ***leading hypothesis*** states that *it is possible to detect respiratory insufficiency by analyzing spoken utterances in real-life conditions*, typically a moderately large sentence, thus subscribing to the view of *speech as a biomarker*. This work aims at validating this leading hypothesis using deep learning techniques.

If the hypothesis holds, it will motivate further enquiries on the use of automated speech analysis to support health professionals; with infectious diseases such as COVID-19, a serious concern involves deciding whether an RI suspect should stay in social isolation or be directed to a medical facility. Project SPIRA[2][3] was initiated to investigate the feasibility of supporting medical triage of patients with COVID-19 symptoms by remotely detecting respiratory insufficiency through automated speech utterance analysis, where no other resources are

†Corresponding author:mfinger@ime.usp.br .

[1]https://covid19.who.int, visited May 24 2021.
[2]https://spira.ime.usp.br/
[3]In Portuguese, ***S**istema de detecção **P**recoce de **I**nsuficiência **R**espiratória por análise de **A**udio* – system for early detection of respiratory insufficiency via audio analysis.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 625–633
August 1–6, 2021.

available other than a phone line or a cellphone app. A positive result may motivate further research into speech-based remote detection of respiratory problems originating from other causes, such as heart condition, airway obstruction, severe asthma, H1N1, etc.

This research started as a response to the peak of the first COVID-19 wave in 2020, when health infrastructure was overloaded, so no doctors nor nurses were available for data collection, and there was no triage point available for research. Thus, COVID-19 patient utterances were collected mostly by medical students at COVID-19 wards from patients with blood oxygenation below 92%, as an indication of respiratory insufficiency, and control data was collected by voice donations over the internet, assumed healthy, with no access to blood oxygenation. Recordings were made in out-of-studio conditions, using portable recording equipment employed in noisy wards. On the other hand, conditions for healthy voice donations over the Internet and using diverse sound equipment had a large variation. This *audio data in-the-wild* approach was assumed from the start as part of the challenge of validating the leading hypothesis. Part of the methodological novelty of this work lies on how to deal with these conditions. This task required a multidisciplinary group involving medical doctors, linguists, speech therapists and computer scientists, all of which were aware of those conditions and challenges facing us.

This work proposes a machine learning method to detect respiratory insufficiency by analyzing voice audio recordings of sentences long enough to feature respiratory pauses in speech. The test is very cheap, requiring only a voice sample from each patient and maybe employed where no other medical equipment is available.In order to tackle the audio analysis, we propose the use of deep artificial neural networks over Mel Frequency Cepstral Coefficients (MFCCs) (Logan et al., 2000) extracted from patient's audios.

The code and datasets are publicly available at https://github.com/SPIRA-COVID19/SPIRA-ACL2021, under a CC BY-SA 4.0 license.

This paper is organized as follows: Section 2 discusses related work. In Section 3, the dataset acquired, the preprocessing steps, the noise insertion procedure, the proposed model and experiments are described, respectively. Afterwards, the models obtained are evaluated and discussed in Section 4. Finally, Section 5 presents the conclusions and final thoughts.

2 Related work

COVID-19 is a recent disease. However, even before the eruption of the pandemic, we could already find in the literature a few explorations of speech as a biomarker (Botelho et al., 2019; Trancoso et al., 2019; Nevler et al., 2019), with some recent recommendations (Robin et al., 2020).

Several initiatives can be found on the Web that record human voice in order to assess the presence and the gravity of COVID-19, e.g. the COVID-19 Sounds data collection initiative (Brown et al., 2020) and startup initiative aiming to develop a pre-diagnostic tool[4]. Those works aim to diagnose COVID-19 from voice or breathing or coughing sounds, and there are some initial positive results on COVID-19 detection in asymptomatic individuals (Laguarta et al., 2020). Unlike our approach, no work aimed specifically at respiratory insufficiency or at patient triage, but they propose to employ some form of artificial intelligence processing.

In similarity to our goals, there have been recent proposals of applications for the triage of patients using natural language processing of texts extracted from radiology reports (Hassanpour et al., 2017) and patient questionnaires (Spasić et al., 2019). So language, both as text and now as speech, is being used for patient screening.

Moreover, Neural Networks and Convolutional Neural Networks (CNNs) have been used in noisy environments mostly, but not exclusively, for fault diagnosis (Zhang et al., 2018; Munir et al., 2019), noise reduction in voice processing (Maas et al., 2012) and medical ECG diagnosis (Acharya et al., 2017). On the other hand, *noise injection* was a technique used in the past to avoid overfitting in training Neural Networks (Matsuoka, 1992; Grandvalet et al., 1997; Zur et al., 2009), as opposed to avoiding classification biases, as in our approach.

3 Methodology

In order to build a neural network model for the proposed task, it is necessary to gather a dataset containing voices of healthy individuals and COVID-19 patients (Section 3.1). The resulting dataset required several preprocessing treatments and noise treatment, as discussed in Sections 3.2 and 3.3. The next step was to propose several neural models to

[4]https://www.voicemed.io/

investigate the best one for the task (Section 3.4) evaluated according to experiments carried over the dataset (Section 3.5).

3.1 Dataset

The dataset creation was composed of two parts and an "appendix". The first part consisted of audios gathered via Web by a system specifically designed for this task[5], from May to July of 2020. Healthy volunteers were asked to donate audio samples via a web interface. This allowed us to build our control group. In order to do that, the system URL was disclosed through local news and social networking. The resulting dataset part is composed, after elimination of blank samples, of more than 6 thousands voice donors. No blood oxygen saturation information was available for the control group.

In the second part, we collected audios from patients infected by SARS-CoV-2 from June to July of 2020. This collection was performed in COVID-19 wards in two university hospitals, in São Paulo city, Brazil, restricted to patients with blood oxygenation level (SpO2) inferior to 92%, as an indication of respiratory insufficiency. This allowed us to collect 536 samples from patients in different age groups. Several problems led to discarding patient voice samples, chiefly among which were collectors whispering during collection; a large set of collection instructions was assembled during the period in which voice collection took place. It is important to note that São Paulo is a local and international hub, with a large migrant and immigrant population. Hospitals received COVID-19 patients from the city as well as from adjoining regions. Collection was absolutely anonymous, so no one knows who were the patients and controls, and no ethnographic information is available. On the other hand, this allowed us to release the data.

As a COVID-19 ward is a noisy environment, an "appendix" was built for this dataset, consisting of samples of pure background noise at the ward (no voice), typically collected at the start of a collection session. This is an important piece of information, as the ward noise is very different from the background noise found in the control group, and consists of a *data bias* that has to be controlled during experiments.

The gathered audios contain three utterances:

- Utterance 1, a moderately long sentence containing 31 syllables, designed by linguists to allow for spontaneous breathing breaks, while being relatively simple to be spoken, even by low literacy voice donors: *"O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa."* ("Love of neighbor helps in strengthening the fight against Coronavirus.");

- Utterance 2, a well known nursery rhyme for donors having reading difficulties, due to lack or reading glasses in hospital, or other types of reading impediments: *"Batatinha quando nasce, espalha a rama pelo chão, nenezinho quando dorme põe a mão ao coração"* ("When small potatoes germinate, branches sprout on the ground; when baby sleeps, hands rest over the heart");

- Utterance 3, a widely known song, on the lines of "Happy birthday to you": *"Parabéns a você, nesta data querida, muitas felicidades, muitos anos de vida"* ("Happy birthday to you, on this dear date, lots of happiness, many years of life").

Collecting longer utterances was totally impractical in a COVID-19 ward. The collection had to be adapted to what was possible in that context.

We identified several issues with the original dataset that need to be addressed. First, there is class imbalance, as we have fewer positive instances (COVID-19 patients) than negative ones (healthy individuals from the control group). Second, it is sex imbalanced, as a greater number of healthy women participated in the process than healthy men. Additionally, there are more men in COVID-19 wards than women. Third, there is an age imbalance, as there are more elderly in hospital care than young people in our observations. Fourth, we also detected utterance imbalance, as utterance 1 was more common among patients; healthy people typically recorded all proposed utterances. Fifth, the control group presented popping and crackling noise, possible due to the characteristics from the recording devices. Furthermore, as mentioned above, wards tend to be noisy environments.

We addressed most of the dataset issues by sample balancing, taking advantage of the greater number of control group samples. Only audios from utterance 1 were selected and the number of samples used in experiments was balanced by class and sex, but not by age, to avoid drastically reduc-

[5] https://spira.ime.usp.br/coleta/

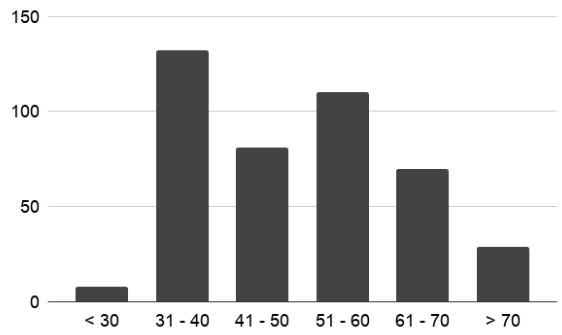

Figure 1: Distribution of Ages in the Dataset

ing the available data. Overall age distribution is presented in Figure 1.

We also had to discard audio containing the collector's (whispering) voice. The most serious issue for bias removal, though, is the presence of ward background noise in patient audios; we observed that it is easier to insert ward noise in the control group than to remove it from the patients' signal. This process will be addressed in Section 3.2.

The dataset was divided in training, validation and test, as is usual in statistical learning. We selected audios with the best signal-noise ratio to use in the test set, and the second best audios were used for validation. The aim of this partitioning is to detect training overfitting.

Information of the resulting filtered dataset is presented in Table 1.

3.2 Pre-processing

In general, the majority of the audios in the dataset was sampled at 48kHz. We pre-processed these files using Torch Audio 0.5.0 in the following way. First, for dimensionality reduction reasons, we resampled these audios at 16kHz. Second, we extracted the MFCCs using a 400ms window employing Fast Fourier Transform (FFT) (Brigham and Morrow, 1967), with hop length 160 and 1,200 FFT components, of which we retained only 40 coefficients. Before the MFCC feature extraction process though, we need to address the difference of duration present in our data.

The duration of our samples in the dataset varies, in which audios from the positive class are slightly longer than audios from the negative class, as presented in Table 1. We have developed two approaches to deal with this phenomenon. First, we applied padding in the instances during training. This is equivalent to complete the audios with silence so that all audios have the same duration. Second, we have extracted fixed length fragments from the audios. This approach aims to prevent the model from performing the classification giving too much importance to the audio length. In order to augment the training data, windowing with 1 second steps was applied to extract audio fragments.

3.3 Noise Insertion During Training

Ward noise is a serious bias source, as confirmed by our preliminary experiments (Section 4). In this scenario, a neural network can be biased during training by focusing only on background noise. One possible alternative would be noise filtering, but besides the possibility of inserting extra biases due to differential noise suppression in patient and control audio samples, there is also the possibility of suppressing important low-energy information that allows for the distinction between healthy and respiratory affected speech samples.

To address this issue, we decided to record pure background noise samples from COVID-19 wards and to inject into patients and control group audios. In total, 16 samples with approximately 1 minute each were recorded.

The inserted noise can also be a cause of bias, as the model can extract specific features from the noise recording. To avoid this kind of bias, we decided to inject noise in all samples. We had the option of inserting in training, validation and testing samples, which will be described in Section 4. We can also control the amount of noise samples inserted in each audio.

In our experiments, we investigate the ideal number of noise samples to inject in both patients and the control group. This had a big impact in overfitting prevention, as a form of unbiased learning, as described in Section 4. During training, at each epoch, audio samples can be injected with one or more distinct noise samples. Each time a given audio is used for training, noise samples are drawn from the noise base. Besides that, the start point of each noise sample is also randomized. Finally, we also draw a factor to change the intensity of the sample. This factor is constrained by a maximum amplitude value, which was determined from the analysis of patient audio noises. The aim is to insert noises as similar as possible to already pre-existing noise. We also executed the same experiment three times with different random seeds to obtain better measures of the noise insertion impact.

Table 1: Filtered dataset information

Sets	Control			Patients			Total Audios	Total Duration (s)
	Male	Female	Mean Duration (s)	Male	Female	Mean Duration (s)		
Training	59	84	8.15	83	66	13.18	292	3110
Validation	8	8	7.75	8	8	10.78	32	296
Test	22	26	7.77	28	32	9.43	108	983

The test and validation sets were created in such a way to allow overfitting detection as they are composed mostly of audios with very limited amount of noise. As a result, we cannot apply k-fold Cross Validation and similar methods. We compensate this by running the same experiment three times with different random seeds. This fact, together with the dynamic noise insertion during training, allows us to obtain averaged accuracy for each experiment.

3.4 Proposed Model

Several models were tested in preliminary experiments and we describe the one that led to the best results.

This process involved three main aspects: (a) the topology and model parameters; (b) the main hyper-parameters; (c) regularization. The last is especially important, since our dataset contains several issues that can lead to overfitting.

Regarding topology and model parameters, preliminary experimental results showed that CNNs applied to MFCCs are useful to analyze this kind of problem. Other preliminary experiments investigated spectrograms and topologies like fully-connected and recurrent networks, which showed lower performance than the chosen topology. Figure 2 presents the chosen model's main features including layers, filters, kernels, number of neurons and activation functions. The following conventions are adopted in the figure: kernel size is represented by K; convolutional dilation size (Yu and Koltun, 2015) is represented by D; and fully connected layers are represented by FC. The input size is omitted because these parameters changed according to the experiment and will be detailed in Section 3.5. We investigated the use of Mish activation function (Misra, 2019), due to its regularization effects during training, which helps prevent overfitting.

Regarding the main hyper-parameters, we have used the Binary Cross-Entropy as loss, and Adam optimizer (Kingma and Ba, 2014). The initial learning rate was set to 10^{-3}, and the Noam's decay scheme (Vaswani et al., 2017) was applied on each 1,000 steps. For each experiment presented in Section 3.5, we trained the model for 1,000 epochs using a batch size of 30.

Regarding regularization, overfitting mitigation is a major concern given our dataset noise characteristics. Therefore, several approaches for regularization were applied. Besides Mish as an activation function, we used three other strategies. First, a global weight decay of 0.01 was applied. Second, a dropout of 0.70 was used in all layers, except in the output layer. Last, we applied group normalization (Wu and He, 2018) after each convolutional layer. The group normalization was applied on pairs of convolution filters. Therefore, the number of groups is half the number of filters.

3.5 Experiments

For the experiments we explored three main aspects with respect to noise insertion and duration variance: (a) overfitting impact; (b) padding vs windowing approach (using four second windows or adding padding); and (c) the ideal number of noise samples. Table 2 presents the proposed experiments and their results.

First we investigated if the model can overfit when trained over original audios (experiments 1.x). In this series of experiments, we trained the model using both approaches of duration variance.

Second we analyzed two approaches to address the duration variance: audio padded to the maximum length of the dataset; windowing using the approach described in Section 3.2 (experiments 2.x). Specifically, we presented padding application only in experiments 1.1 and 2.1 because experiments showed that the windowed approach led to more robust results. When padding is used, the accuracy is calculated as usual. However, in windowed experiments, several audio fragments are extracted and their predictions averaged for the classification decision. Regarding window size, we have chosen four seconds, considering our smallest audio in the

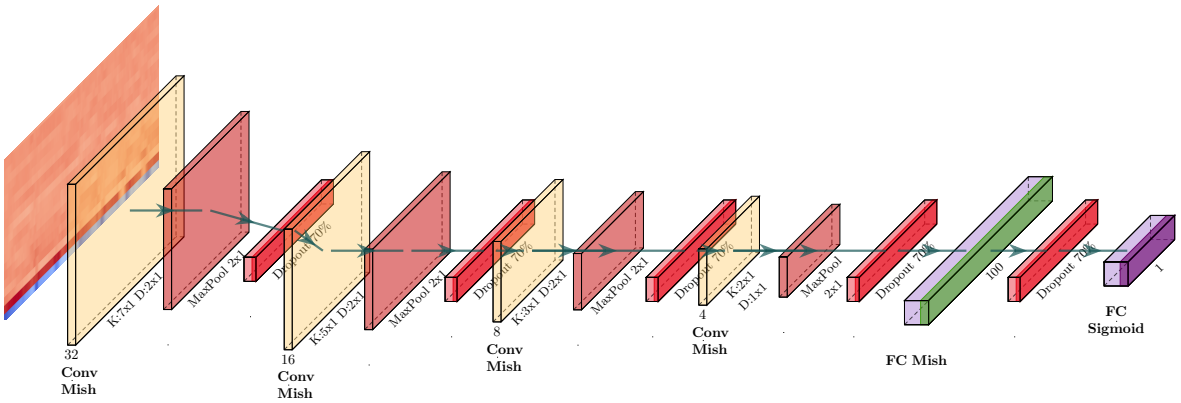

Figure 2: CNN topology proposed with four convolutional layers and two fully connected layers

Table 2: Proposed experiments and results

Description	Exp.	Duration Approach	Noise Samples		Accuracy (without noise in test samples)	Accuracy (with noise in test samples)	Training time (h)
			Patient	Control			
Overfitting Analysis	1.1	Padding	0	0	98.15 ± 0.93	50.93 ± 0.53	4.25
	1.2	Windowing	0	0	98.15 ± 0.53	50.93 ± 0.93	9.07
Duration Variance	2.1	Padding	0	1	61.11 ± 8.40	74.07 ± 1.93	4.77
	2.2	Windowing	0	1	66.67 ± 3.74	86.11 ± 2.98	9.58
Noise Insertion Analysis	3.1	Windowing	1	1	80.56 ± 2.45	68.52 ± 1.41	6.57
	3.2	Windowing	1	2	84.26 ± 6.17	83.33 ± 3.34	12.27
	3.3	Windowing	2	2	88.89 ± 0.53	85.19 ± 0.93	13.00
	3.4	Windowing	2	3	74.07 ± 5.10	85.19 ± 1.85	13.67
	3.5	Windowing	3	3	91.67 ± 2.98	87.04 ± 0.93	14.70
	3.6	Windowing	3	4	62.96 ± 8.35	74.07 ± 2.45	11.85
	3.7	Windowing	4	4	88.89 ± 1.41	83.33 ± 1.07	10.40
	3.8	Windowing	4	5	56.48 ± 5.10	72.22 ± 9.99	9.83
	3.9	Windowing	5	5	70.37 ± 15.8	69.44 ± 9.27	10.55
	3.10	Windowing	5	6	51.85 ± 3.51	61.11 ± 2.98	11.18
	3.11	Windowing	6	6	74.07 ± 10.7	74.07 ± 8.83	11.98
	3.12	Windowing	6	7	50.00 ± 0.53	54.63 ± 3.51	12.63

dataset contains 4.6 seconds and new data samples can be even smaller.

Third we examine the ideal number of noise samples to be inserted to prevent overfitting (experiments 3.x), using the best duration approach according to experiments 2.x. For each experiment, we tested the model using both noise insertion and no noise insertion to analyze performance.

Our model was implemented using Pytorch 1.5.1. We ran the experiments on a NVIDIA Titan V GPU with 12GB RAM in a server with Intel(R) Core(TM) i7-8700 CPU and 16GB of RAM.

4 Results and Discussion

To better understand bias and overfitting we used a test set containing only audios with a minimal amount of noise. The accuracy of each experiment is presented in Table 2, both with and without artificial insertion of ward noise in test samples.

Experiments 1.x showed the model is biased without noise insertion in the training set. We note a high accuracy in experiments 1.1 and 1.2 without noise in training and testing; in contrast, when noise is inserted in all test samples, it classifies all samples as coming from patients. We interpret this as a strong indication that the model is biased by the presence of noise in the patient samples.

Experiments 2.x showed that windowing (2.2) is preferable over padding (2.1), as described in Section 3; the model performs better when the windowed approach is used, that is, 66% using windowing against 61% using padding. We consider this as evidence of susceptibility to bias by padding. In fact, padding inserts a considerable amount of silence, specially in patient samples, and the windowed approach works as a data augmentation technique, as more instances are generated in this process.

Experiment 3.x were used to determine the optimal amount of noise insertion. Note that sometimes better results were obtained without noise in test samples and sometimes the other way around. In general, the bias is greatly reduced by inserting at least one noise sample on the negative instances. As expected, the insertion of too much noise decreases the model performance. The best overall accuracy was obtained in experiment 3.5, which reached 91% accuracy in the task. For experiment 3.5, we obtained F1 = 0.90, without noise insertion; with noise insertion, F1 = 0.87.

Figure 3 presents the loss variation of the best model (experiment 3.5) during training. Early stopping is used to get the best iterations after approximately 20k steps.

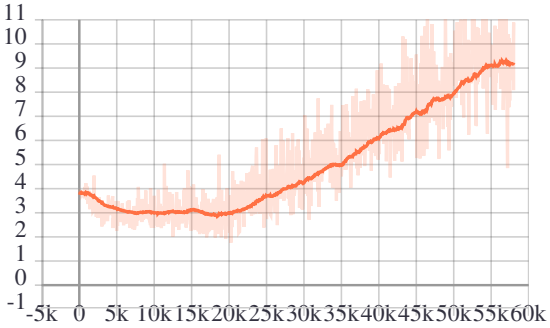

Figure 3: Validation Loss for Experiment 3.5 during Training

Figure 4 shows the model performance over the number of noise samples inserted. With respect to the number of noise samples, our experiments suggest that a number of 2 to 4 noise insertions in each audio provides best accuracy. In each case, two possibilities have been tested, namely the insertion of an equal number of noise samples in each training audio, and the insertion of one extra noise sample to control audio, assuming that patient audios already have the original ward noise. It was initially expected that the insertion of an extra noise sample in control audios would produce better results; surprisingly, the opposite effect was observed. The possible explanation for this observation is that there are times when wards are calmer and silent and the insertion of noise in control audios leads to bias. This is especially true for testing samples, due to the criteria used to build the testing set.

5 Conclusions and future work

In the effort to tackle the COVID-19, we have developed a method to classify real-life speech audio

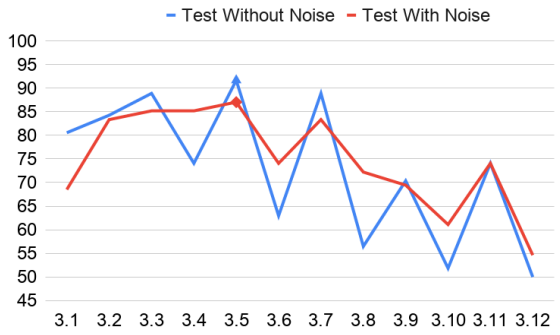

Figure 4: Sample Noise Analysis for the Best Experiments

signals on whether or not that signal originated from a person suffering from respiratory insufficiency. In this effort, we obtained 91.67% accuracy, thus validating the hypothesis that such a detection is feasible, and that human speech can be treated as a biomarker in this case.

One important consequence of this work was the construction of a dataset containing voice samples of COVID-19 patients with respiratory insufficiency and also a set of samples of environmental noise, which were central in treating real-life sound samples. Noise insertion was chosen as the more adequate option when contemplating the biases that would be incurred by filtering procedures. In particular, it made sense to add ward noise to the existing ward samples as a way to balance the biases that were incurred by the necessary addition of ward noise to control data. In this way, all data (patient and control) suffered from similar manipulation, avoiding editing bias, and experiments showed that such a procedure produced best results. This aimed at preventing the models from memorizing ward noise and editing distortion information instead of COVID-19 features.

There was a considerable difficulty to collect voice data from infected patients during the pandemic. The size of the patient dataset reflects the limitations on collections in COVID-19 wards. Moreover, the use of audio from different environments was absolutely unavoidable, as we only had access to patients in COVID-19 wards, where no control subjects were available. Therefore, control data had to be collected in a different environment. As a result, the amount of data was scarce, and data augmentation techniques were designed for such a setting; our results indicate that it was not an excessive amount of data augmentation, as con-

sistent results were obtained over a large variety of experiments. We hope that with the weakening of the emergency situation, it could become easier to collect data from patients with respiratory insufficiency.

Future work includes augmenting the dataset with audios collected at the triage point, whether in hospital admission rooms, or through a remote admission system. In this way, speech audio signals from both sufferers and non-sufferers of respiratory insufficiency would be obtained under similar conditions. This would allow us to extend this study to other respiratory illnesses besides COVID-19. Also, other neural architectures can be explored, as well as smarter feature engineering.

Acknowledgments

This work was supported by Fapesp project 2020/06443-5 (SPIRA). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Marcelo Finger was partly supported by Fapesp projects 2019/07665-4 and 2014/12236-1 and CNPq grant PQ 303609/2018-4.

We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used in part of the experiments presented in this research.

References

U. Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. 2017. Application of deep convolutional neural network for automated detection of myocardial infarction using ecg signals. *Information Sciences*, 415-416:190 – 198.

M Catarina Botelho, Isabel Trancoso, Alberto Abad, and Teresa Paiva. 2019. Speech as a biomarker for obstructive sleep apnea detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5851–5855. IEEE.

E Oran Brigham and RE Morrow. 1967. The fast fourier transform. *IEEE spectrum*, 4(12):63–70.

Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2020. Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. *arXiv preprint arXiv:2006.05919*.

Yves Grandvalet, Stéphane Canu, and Stéphane Boucheron. 1997. Noise injection: Theoretical prospects. *Neural Computation*, 9(5):1093–1108.

Saeed Hassanpour, Curtis Langlotz, Timothy Amrhein, Nicholas Befera, and Matthew Lungren. 2017. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: A tool to estimate diagnostic yield. *AJR. American Journal of Roentgenology*, 208:1–4.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

J. Laguarta, F. Hueto, and B. Subirana. 2020. Covid-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:275–281.

Beth Logan et al. 2000. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, pages 1–11.

Andrew Maas, Quoc V. Le, Tyler M. O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng. 2012. Recurrent neural networks for noise reduction in robust asr. In *INTERSPEECH*.

K. Matsuoka. 1992. Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):436–440.

Diganta Misra. 2019. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*.

Nauman Munir, Hak-Joon Kim, Jinhyun Park, Sung-Jin Song, and Sung-Sik Kang. 2019. Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions. *Ultrasonics*, 94:74 – 81.

Naomi Nevler, Sharon Ash, David J Irwin, Mark Liberman, and Murray Grossman. 2019. Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology*, 6(1):4–14.

J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, and M.J. Yancheva. 2020. Evaluation of speech-based digital biomarkers: Review and recommendations. *Digital Biomarkers*, 4(3):99–108.

I. Spasić, D. Owen, A. Smith, and K. Button. 2019. Klosure: Closing in on open–ended patient questionnaires with text mining. *Journal of Biomedical Semantics*, 10.

Martin J Tobin, Franco Laghi, and Amal Jubran. 2020. Why covid-19 silent hypoxemia is baffling to physicians. *American Journal of Respiratory and Critical Care Medicine*, 202(3):356–360.

Isabel Trancoso, Maria Joana Ribeiro Folgado Correia, Francisco Teixeira, Alberto Abad, Maria Catarina Tavares Botelho, and Bhiksha Raj. 2019.

Speech as a (private?) biomarker for speech affecting diseases. In *In ICIEA 2019 - The 14th IEEE Conference on Industrial Electronics and Applications*, Xi'an, China. IEEE. Keynote paper.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Wei Zhang, Chuanhao Li, Gaoliang Peng, Yuanhang Chen, and Zhujun Zhang. 2018. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing*, 100:439 – 453.

Richard M Zur, Yulei Jiang, Lorenzo L Pesce, and Karen Drukker. 2009. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical physics*, 36(10):4810–4818.