DETECTING RESPIRATORY INSUFFICIENCY VIA VOICE ANALYSIS: THE SPIRA PROJECT

Sandra M. Aluísio
Augusto C. Camargo Neto
Edresson Casanova
Flaviane Fernandes-Svartman
Renato Ferreira
Marcelo Finger
Alfredo Goldman
Pedro Leyton
Anna S. Levin
Marcelo M. Gauy
Marcus Martins
Marcelo Queiroz
Beatriz Raposo de Medeiros
Ester C. Sabino
University of São Paulo (USP)

Arnaldo Candido Jr
Ricardo Fernandes Jr
Lucas R. S. Gris
Daniel da Silva
Federal University of Technology, Paraná (UTFPR)

Evelyn Alves Spazzapan
Larissa C. Berti
São Paulo State University (UNESP)

J. Henrique Quirino,
Beneficiência Portuguesa

ABSTRACT

This paper describes the first stage activities of the SPIRA Project, a COVID-19 motivated research effort to design a system for the early prediction of respiratory insufficiency via audio analysis. We describe the research motivation, its organization in research lines, the initial results obtained in those lines and a preview of the future steps in this research project.

1 INTRODUCTION

The COVID-19 pandemic has stressed the need to develop simple, cheap and widely available biomarkers. Monitoring potential patients remotely, frequently and automatically is the best way to combine respect for social distancing and patient safety. According to specialists, one of the most important symptoms of COVID-19 that leads to hospitalization is *respiratory insufficiency*, a condition that is amplified in the case of the current pandemic due to the frequent occurrence of *silent hypoxia*, that is, low blood oxygenation without noticeable short breath (Tobin et al., 2020).

This project aims at investigating voice signals as a biomarker, investigating the feasibility of early detection of respiratory insufficiency. We propose both using artificial intelligence techniques, which operate mostly as black boxes, as well as interpretable, more traditional voice analysis tools. Before the eruption of the pandemic, the literature already mentioned speech as a biomarker, a point of view which this work supports (Botelho et al., 2019; Trancoso et al., 2019; Nevler et al., 2019; Giovanni et al., 2021). An early model for speech production subsystems and their neuromotor coordination as a biomarker of COVID-19 has been proposed (Quatieri et al., 2020).

With respect to detecting signs of COVID-19 in audio recordings, there are several research initiatives on the Internet[1], by startups[2] or public challenges[3], a few of which have already published initial results (Tailor et al., 2020; Orlandic et al., 2021; Laguarta et al., 2020). Our work was the among the first to aim specifically at respiratory insufficiency for patient triage. However, there are several initiatives using language processing and artificial intelligence tools for patient screening and treatment selection, such as patient selection by text extraction from radiology report (Hassanpour et al., 2017) and also text processing from patient questionnaires (Spasić et al., 2019). Those works, however, employ written language processing.

[1] https://link.springer.com/chapter/10.1007/978-981-10-7419-6_6

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

[3] https://www.kaggle.com/vbookshelf/respiratory-sound-database

We pursue two complementary approaches to develop a detection tool. The first approach collects large amounts of data from respiratory insufficiency patients and healthy people, and applies artificial intelligence and machine learning techniques to obtain a speech classification system. This predictive task we call the *Big Data* approach. However, data-intensive approaches are notoriously opaque and do not yield satisfactory explanations on the underlying phenomena that are present in the audio signals. The descriptive task of providing a detailed description of signal properties pertaining to respiratory insufficiency in voice and speech signals is our second approach, called the *White Box approach*.

Our general approach subscribes to the view of speech and voice as biomarkers (Botelho et al., 2019). In this respect, the goals of the SPIRA Project are as follows: (a) creation of the dataset; (b) development of audio preprocessing methods and artificial intelligence algorithms and audio processing necessary to analyze the audios; (c) development of a broad acoustic description (sound signal and speech and voice acoustic) and a linguistic description of the audios; (d) implementation of an automatic audio classifier.

The main results of the first stage of the project are:

(a) Constructing a dataset of audios for COVID-19 related respiratory insufficiency;

(b) Showing the feasibility of detecting respiratory insufficiency with several neural net methods, obtaining more than 90% accuracy.

(c) White box description of pauses in speech as COVID-19 biomarkers.

The paper is organized as follows. Dataset construction is presented in Section 2, which is followed by white box description in Sections 3 and 4, and big data analysis in Section 5. Brief conclusions are presented in Section 6.

2 DATASET

The collected voice samples from two different sources. Initially we started collecting audios from patients infected by SARS-CoV-2, in special COVID-19 wards in three different hospitals of São Paulo: two public hospitals linked to the University of São Paulo (Hospital das Clínicas and Hospital Universitário) and a private one (Beneficência Portuguesa). Voice samples were collected only from patients with blood oxygenation level (SpO2) inferior to 92%, as an indication of respiratory insufficiency. In the hospitals, 536 samples were collected from patients in different age groups.

A second source consisted of audios recorded via a web-based application. A system was specifically implemented to collect speech audio donations from healthy volunteers. This allowed us to build a control group. The system URL[4] was disclosed through local news and social networking, which allowed the collection of more than 6 thousands voices. A third source of audios was recordings of noise ward, created to deal with the fact that this kind of noise is present only patient's audios and not on the control group.

Different utterances were captured, but predominantly, the dataset consists of the sentence "*O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa*" ("Love of your neighbor helps in strengthening the fight against Coronavirus"). This work focuses on this utterance, a moderately long sentence containing 31 syllables and syntactic/prosodic branching constituents, designed to allow for possible breathing breaks in major syntactic boundaries (e.g. the syntactic boundary between the branching subject and the predicate), while being relatively simple to be spoken, even by low literacy voice donors.

Several issues with the original dataset were identified and treated: class imbalance (fewer patient instances); sex imbalance (more males in patients and more females in the control group); age imbalance (more elderly patients). We addressed most of the dataset issues by sample balancing, taking advantage of the greater number of control group samples. The number of samples used in experiments was balanced by class and sex, but not by age, to avoid drastically reducing the available data. Other issues actually led to discarding collected samples from the dataset (second voices in the audio, popping and crackling noise, among others). The most serious issue for bias removal,

[4] https://spira.ime.usp.br

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.

3 Speech pause as a COVID-19 Biomarker (White Box)

It is widely known that human voice is multidimensional as it involves a coordinated action of respiration, phonation and resonating systems (Kent, 1997; Patel et al., 2018; Asiaee et al., 2020). Any clinical or health condition that interferes with these systems may affect vocal production and vocal quality, voice aspects that can be designated as dysphonia. The literature has reported that 28.6% of those infected with COVID-19 showed symptoms of dysphonia (Lechien et al., 2020), as COVID-19 patients may present decreased or lack of energy for vocal production Asiaee et al. (2020).

The following hypotheses guided this work. Admitting that respiratory insufficiency caused by COVID-19 will impact the production of patient speech which, in comparison with the healthy subjects, may present: H1) a greater number of pauses during speech production; H2) longer pauses; and H3) longer utterances. It is also expected that there are influences of sex and age on predicted variables, such as number of pauses, duration of pauses and duration of utterances.

The goals of this part of the study are to compare the characteristics of the speech pause (number and duration) in the speech of individuals with and without COVID-19; to check if there are differences in these variables as a function of the sex and age of the participants.

From the collected dataset, this study selected 193 speech samples, 94 of which came from the speech production of COVID-19 patients (46 men and 48 women) who constituted the Patient Group (PG) and 99 samples from healthy subjects (50 men and 49 women) constituting the Control Group (CG). Patients were hospitalized and had blood oxygen saturation level below 92%, indicating respiratory insufficiency.

Samples were selected using visual and auditory acoustic inspection of audio files, from oscillogram and spectrogram analyses. Recordings were divided into 5 age groups (ranging from ≥ 30 years to ≤ 60 years of age). Samples of speech consisted in the production of the dataset recorded sentence. That sentence was designed in such a way that, if pauses are made, they occur conditioned by prosody restrictions; it is moderately long, containing 31 syllables, designed to allow for possible boundary breaks syntax that correspond to higher level boundaries of prosody constituents. An automated speech segmentation tool was developed to analyze pauses (number and duration) and extraction of the total duration of the utterance. Three variables were analyzed, namely: total duration of utterance; number of pauses; and average pause duration.

Speech	*Control*		*Patient*	
parameter	*Average*	*Std Dev*	*Average*	*Std Dev*
Utterance duration (s)	5.34	0.85	7.95	2.59
Pause duration (s)	0.13	0.16	0.53	0.19
Number of pauses	0.85	0.94	3.16	2.02

Table 1: Average and standard deviation values for the study variables

Initial results are shown in Table 1, according to which the group of patients has a longer average duration of the utterance (5.34s and 7.95s) and duration of pauses (0.13s and 0.53s), as well as a greater number of pauses (0.85 and 3.16 pauses per utterance). It is noteworthy that the maximum number of pauses for the control group is 4, while for the patients this value is 11.

Analyzing sex and age influence, we noted that patients tend to have approximately the same duration of their utterances, regardless of the sex and age range, with the exception of the stratum >60. The control group has a gradual increase in duration, so that the elderly utterances are longer than those of young people.

The results fully corroborate the hypotheses. The symptoms of COVID-19 are fundamentally related to the respiratory system which, in turn, critically influences speech production process. Changes in speech fluency caused by COVID-19 were similar both for men and women, although women may present differences in the inflammatory process due to the high expression of ACE2, the COVID-19 receptor enzyme. Furthermore, patients tend to produce utterances with relatively similar duration, with the exception of the last age group. This means that the longer duration of utterances in the elderly may reflect the degree of sarcopenia (more accentuated in patients) and/or the decrease in neuromuscular control that can affect speech motor control.

To conclude, patients with COVID-19 show changes in speech fluency and speech pauses can be considered a biomarker and be used to identify respiratory insufficiency caused by COVID-19.

4 Signal Processing

Signal processing involves extraction of features that are relevant for linguistic and vocal signal descriptions as well as the production of alternative representations that are relevant for machine learning and classification of the input signals. Four main operations were defined to process the signal: segmentation; noise reduction; feature extraction and annotation alignment.

Segmentation of the audio signal is a preliminary step for many subsequent signal processing tasks. A first level of segmentation consists in identifying speech utterances and background noise. The second operation is noise reduction, which improves both features extraction and machine learning. Noise gating is a well-known technique for noise reduction which is based on a gaussian representation of the noise spectrum. Using speech/noise segmentation to define such a gaussian model of the noise allows the training of an adaptive non-linear filter which selectively suppress or attenuate specific time-frequency components within the signal's spectrogram, which may then be resynthesized as a new noise-reduced audio signal. Feature extraction allows to gather additional metadata relevant for linguistic and vocal signal description (Mitrović et al., 2010), producing augmented representations for machine learning, including statistics related to the number and duration of continuous speech utterances and interruptions from speech/noise. Pitch and timbre related descriptions, easily obtained from different signal representations (Mauch & Dixon, 2014; Hibare & Vibhute, 2014; Gómez & Herrera, 2004), provide several useful statistics of f0 and cepstral peak prominence as well as characterization of formants. Regarding annotation alignment, automated techniques are under development.

5 Signal Classification (big data)

For machine learning purposes, the dataset was divided into training (292 audios), validation (292) and test (108). We selected audios with the best signal-noise ratio to use in the test set, and the second best audios were used for validation. We used Convolutional Neural Networks classify the MFCCs of the input signal (Casanova et al., 2021b;a).

The first step in the training is to preprocess the obtained audios. 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, audios were resampled at 16kHz for dimensionality reduction.. Second, we extracted the MFCCs using a 400ms window employing Fast Fourier Transform (FFT) (Brigham & 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. The difference in audio duration were adressed using windowing, to obtain 4 seconds audios (1 second hop).

To address ward noise only in patients audios, we injected pure background noise samples obtained from COVID wards both for patients and control group. For this, we recorded 16 samples with approximately 1 minute each. We decided to inject noise in all training and validation samples for both patients and control group, to prevent bias in the process. We also inject noise in some testing samples. Several neural network models were tested in preliminary experiments and we describe the one that led to the best results. Figure 1 presents the chosen CNN model main features including layers, filters, kernels, number of neurons and activation functions. The following conventions is adopted: K is the kernel size ; D is convolutional dilation size (Yu & Koltun, 2015); FC represents fully connected layers. The input size varies according to the experiment. We investigated the use of Mish activation function (Misra, 2019) due to its regularization effects, which helps prevent overfitting.

We used Binary Cross-Entropy as loss, and Adam optimizer (Kingma & Ba, 2014). The initial learning rate was set to 10^{-3}, and the Noam's decay scheme (Vaswani et al., 2017) was applied on

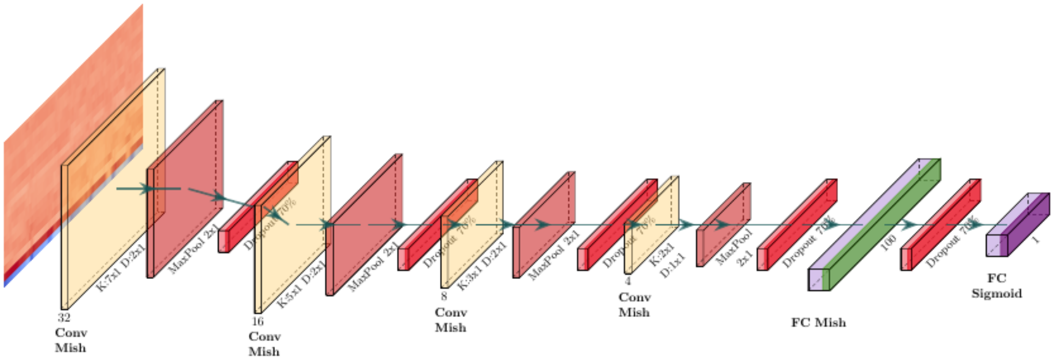

Figure 1: CNN topology proposed with four convolutional layers and two fully connected layers

each 1,000 steps. For each proposed experiment, we trained the model for 1,000 epochs using a batch size of 30 instances. Regarding regularization, overfitting mitigation is a major concern given the dataset noise characteristics. 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, as an overfit reduction strategy. Last, we applied group normalization (Wu & 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. Our models were implemented using Pytorch 1.5.1 and trained on a NVIDIA Titan V GPU with 12GB RAM.

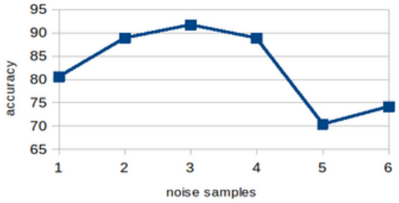

Figure 2: Accuracy obtained per number of noise sampled inserted in training data

Experiments were projected to determine the optimal amount of noise samples inserted in training and validation instances. Some experiments included noise injection also in the test set. In general, the bias is greatly reduced by inserting at least one noise sample on the negative instances. The best overall accuracy was obtained inserting 3 noise samples in both patients and control group, reaching 91% accuracy in the task. Figure 2 presents experiment regarding noise injection. Recent work is exploring new methods to obtain even better accuracy Gauy & Finger (2021).

6 Conclusion

The paper presented the SPIRA Project, an ongoing project. The results obtained so far, seems to validated the original project assumption that respiratory insufficiency can be detected to an acceptable level of accuracy from audio signals obtained by remote recordings. Thus we are encouraged to proceed to develop a pre-diagnostic assistant tool to help health professionals in patient triage activities. Future work involved detailed descriptions of signal properties of patients and non patients, as well as an extension of the current work to deal with respiratory insufficiency originating from causes other than COVID-19.

Acknowledgments

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

REFERENCES

Maral Asiaee, Amir Vahedian-Azimi, Seyed Shahab Atashi, Abdalsamad Keramatfar, and Mandana Nourbakhsh. Voice quality evaluation in patients with covid-19: An acoustic analysis. *Journal of Voice*, 2020.

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

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

Edresson Casanova, A Candido, RC Fernandes, Marcelo Finger, Lucas Rafael Stefanel Gris, Moacir A Ponti, Daniel Peixoto Pinto Da Silva, et al. Transfer learning and data augmentation techniques to the covid-19 identification tasks in compare 2021. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pp. 4301–4305, 2021a.

Edresson Casanova, Lucas Gris, Augusto Camargo, Daniel da Silva, Murilo Gazzola, Ester Sabino, Anna Levin, Arnaldo Candido Jr, Sandra Aluisio, and Marcelo Finger. Deep learning against covid-19: Respiratory insufficiency detection in brazilian portuguese speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 625–633, 2021b.

Marcelo Gauy and Marcelo Finger. Audio mfcc-gram transformers for respiratory insufficiency detection in covid-19. In *Proceedings of the XIII Brazilian Symposium on Information Technology and Human Language (STIL21)*, pp. 143–152, Porto Alegre, RS, Brazil, 2021. SBC. doi: 10.5753/stil.2021.17793. URL `https://sol.sbc.org.br/index.php/stil/article/view/17793`.

Antoine Giovanni, Thomas Radulesco, Gilles Bouchet, Alexia Mattei, Joana Révis, Estelle Bogdanski, and Justin Michel. Transmission of droplet-conveyed infectious agents such as sars-cov-2 by speech and vocal exercises during speech therapy: preliminary experiment concerning airflow velocity. *European Archives of Oto-Rhino-Laryngology*, 278(5):1687–1692, 2021.

Emilia Gómez and Perfecto Herrera. Automatic extraction of tonal metadata from polyphonic audio recordings. In *AES*, 2004.

Saeed Hassanpour, Curtis P Langlotz, Timothy J Amrhein, Nicholas T Befera, and Matthew P Lungren. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. *American Journal of Roentgenology*, 208 (4):750–753, 2017.

Rekha Hibare and Anup Vibhute. Feature extraction techniques in speech processing: a survey. *International Journal of Computer Applications*, 107(5), 2014.

R.D. Kent. *The Speech Sciences*. Singular Publishing Group, 1997. ISBN 9781565936898. URL `https://books.google.com.br/books?id=YtVqAAAAMAAJ`.

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

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

Jerome R Lechien, Carlos M Chiesa-Estomba, Pierre Cabaraux, Quentin Mat, Kathy Huet, Bernard Harmegnies, Mihaela Horoi, Serge Daniel Le Bon, Alexandra Rodriguez, Didier Dequanter, et al. Features of mild-to-moderate covid-19 patients with dysphonia. *Journal of Voice*, 2020.

Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 659–663. IEEE, 2014.

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

Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. Features for content-based audio retrieval. In *Advances in computers*, volume 78, pp. 71–150. Elsevier, 2010.

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

Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):1–10, 2021.

Rita R Patel, Shaheen N Awan, Julie Barkmeier-Kraemer, Mark Courey, Dimitar Deliyski, Tanya Eadie, Diane Paul, Jan G Švec, and Robert Hillman. Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function. *American journal of speech-language pathology*, 27(3):887–905, 2018.

Thomas F Quatieri, Tanya Talkar, and Jeffrey S Palmer. A framework for biomarkers of covid-19 based on coordination of speech-production subsystems. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:203–206, 2020.

Irena Spasić, David Owen, Andrew Smith, and Kate Button. Klosure: Closing in on open–ended patient questionnaires with text mining. *Journal of Biomedical Semantics*, 10(1):1–11, 2019.

Shyam A Tailor, Jagmohan Chauhan, and Cecilia Mascolo. A first step towards on-device monitoring of body sounds in the wild. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pp. 708–712, 2020.

Martin J Tobin, Franco Laghi, and Amal Jubran. Why covid-19 silent hypoxemia is baffling to physicians. *American journal of respiratory and critical care medicine*, 202(3):356–360, 2020.

Isabel Trancoso, Maria Joana Ribeiro Folgado Correia, Francisco Teixeira, Alberto Abad, Maria Catarina Tavares Botelho, and Bhiksha Raj. Speech as a (private?) biomarker for speech affecting diseases. *In ICIEA*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

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