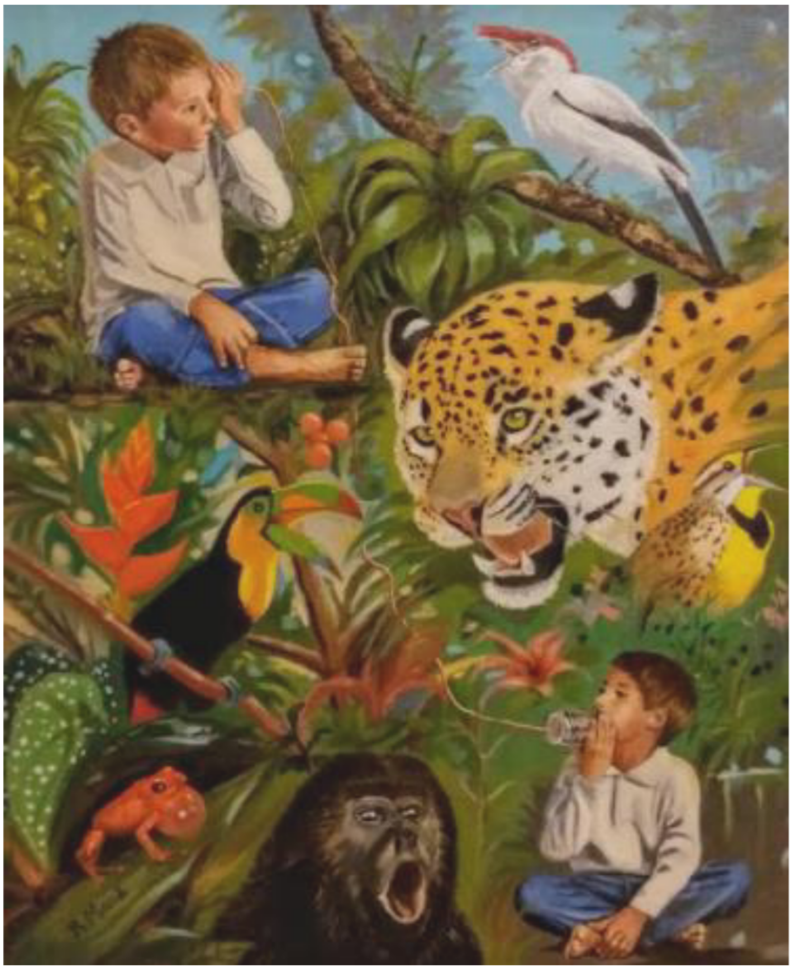
ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH



Organized by
Emma Otta & Patrícia Ferreira Monticelli
Universidade de São Paulo
Pró-Reitoria de Pesquisa da USP

DOI: 10.11606/9786587596198

The opinions in this publication are the exclusive responsibility of the authors and do not necessarily express the point of view of the Institute of Psychology of the University of São Paulo.

Organizing Committee of the ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH

Profa. Dra. Emma Otta (IPUSP)

Profa. Dra. Patrícia Ferreira Monticelli (FFCLRP)

Prof. Dr. Claudio Possani (IME-USP)

Dra. Tania Kiehl Lucci (IPUSP)

Dr. Ricardo Prist (IPUSP)

Cover Photo: Regina Macedo

Book formatting: Dra. Lilian Cristina Luchesi and Dra Aline Domingues Carneiro Gasco

English Editing Services: Michael Germain and Lisa Burger, MC TRADUÇÕES S/S LTDA

Catalogação na publicação
Serviço de Biblioteca e Documentação
Instituto de Psicologia da Universidade de São Paulo

Acoustic communication: an interdisciplinary approach / Organized por Emma Otta e Patrícia Ferreira Monticelli. -- São Paulo, Instituto de Psicologia da Universidade de São Paulo, 2021.

210 p.

E-book.

ISBN: 978-65-87596-19-8

DOI: 10.11606/9786587596198

1. Animal vocalization 2. Animal communication 3. Ethology I. Title

QL765

Ficha elaborada por: Elaine Cristina Domingues CRB5984/08

Funding Acknowledgements

Chapter 13

Detecting Respiratory Insufficiency by Voice Analysis: The SPIRA Project

Marcelo Finger & Spira project group[21]

Abstract

This paper describes the initial 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. It describes 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.

Keywords: Biomarker; Convolutional Neural Networks; COVID-19; health monitoring; prosodic; Small Data approach.

The COVID-19 pandemic has forced most of the population of the world's major cities to social distance. Large gatherings of people can facilitate the spread of the virus, especially in hospitals and health centers. Monitoring potential patients remotely, frequently and automatically is the best way to combine compliance with 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 shortness of breath (Tobin, Laghi, & Jubran 2020).

[21] The SPIRA Project Group consists of Sandra M. Aluísio (ICMC-USP), Evelyn Alves Spazzapan (Fonoaudiologia-UNESP), Larissa C. Berti (Fonoaudiologia-UNESP), Augusto C. Camargo Neto (IME-USP), Arnaldo Candido Jr (CS-UTFPR-Medianeira), Edresson Casanova (ICMC-USP), Flaviane Fernandes-Svartman (FFLCH-USP), Renato Ferreira (IME-USP), Ricardo Fernandes Jr (CS-UTFPR-Medianeira), Marcelo Finger (IME-USP), Alfredo Goldman (IME-USP) , Lucas R. Gris (CS-UTFPR-Medianeira), Pedro Leyton (IME-USP), Anna S. Levin (FM-USP), Marcus Martins (FFLCH-USP), Marcelo Queiroz (IME-USP), J. Henrique Quirino (Beneficência Portuguesa-SP), Beatriz Raposo de Medeiros (FFLCH-USP), Ester C. Sabino (FM-USP), Daniel da Silva (CS-UTFPR-Medianeira).

An automatic system for early detection of respiratory insufficiency via audio analysis meets both health safety needs and medical triage burden relief. We pursue two complementary approaches to develop a detection tool. The first 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. We call this predictive task the Big Data approach. However, data-intensive approaches are notoriously unclear and do not yield satisfactory explanations of the underlying phenomena 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 Small Data 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:

- Creation of a dataset of audios containing speech records of both respiratory insufficiency patients and healthy people who do not require hospitalization. Patient audios initially originated from COVID-19 wards.
- Development of artificial intelligence algorithms and audio processing necessary for the training and execution of classifiers that will screen patient audios (Big Data approach).
- Development of a broad acoustic description (sound signal and speech and voice acoustic) and a linguistic description of respiratory insufficiency by comparing the audio signals of patients and healthy subjects (Small Data approach).
- Implementation of an automatic audio system based on a support audio classifier to assist the patient screening system.

The present chapter is organized as follows. Related work is presented in the first section, dataset construction in the second section, small data description in the third and fourth sections, and big data analysis in the fifth section 5. Brief conclusions are presented in the last section.

Related work

COVID-19 is too recent a disease to be widely covered in the speech processing literature. Even before the outbreak of the pandemic, the literature contained only a few investigations of speech as a biomarker (Botelho et al., 2019; Trancoso et al., 2019; Nevler et al., 2019; Giovanni et al., 2021). In particular, a framework that models speech production subsystems and their neuromotor coordination as a biomarker of COVID-19 has been proposed (Quatieri, Talkar, & Palmer, 2020).

With respect to detecting signs of COVID-19 in audio recordings, there are several research initiatives on the internet[22], by startups[23] or public entities[24], a few of which have already published initial results. For instance, the COVID-19 Sounds Data Collection Initiative (Tailor, Chauhan, & Mascolo, 2020) aims to detect the presence and severity of COVID-19, and the COUGHVID crowdsourcing dataset to develop a screening tool (Orlandic, Teijeiro, & Atienza, 2020). The following studies aim to diagnose COVID-19 from speech, breathing or coughing sounds.

Unlike our approach, no study aims specifically at respiratory insufficiency or patient triage, but propose to apply some form of artificial intelligence processing. It is not yet known if there is a detectable difference between enacted and spontaneous coughs, since most recordings are obtained from provoked situations. Positive identification of asymptomatic only COVID-19 infected patients was recently reported (Laguarta et al, 2020); identification uses artificial intelligence techniques over provoked cough-sound cell-phone recordings, and does not perform as well on patients with symptoms.

With the explicit goal of applying artificial intelligence to hospital triage using natural language processing, text extraction from radiology reports was developed (Hassanpour et al., 2017), as well as text processing from patient questionnaires (Spasic et al., 2019). These studies apply written language processing for patient screening and treatment selection.

Dataset

The voice samples were collected from two different sources. Initially, we collected audios from patients infected by SARS-CoV-2, in special COVID-19 wards at three different hospitals in São Paulo: two public hospitals affiliated with the

[22] "A Respiratory Sound Database for the Development of ...." 17 nov. 2017, https://link.springer.com/chapter/10.1007/978-981-10-7419-6_6. Accessed on 23 October, 2020.

[23] [1] "VoiceMed – Save lives and monitor the health of one billion people." https://www.voicemed.io/. Accessed on 23 October, 2020.

[24] "Respiratory Sound Database | Kaggle." 29 jan. 2019, https://www.kaggle.com/vbookshelf/respiratory-sound-database. Accessed on 23 October, 2020.

University of São Paulo (Hospital das Clínicas and Hospital Universitário) and a private institution (Beneficência Portuguesa). Voice samples were collected only from patients with blood oxygenation levels (SpO2) below 92%, indicating respiratory insufficiency. In the hospitals, 536 samples were collected from patients of different age groups.

The second source consisted of audio recorded via a web-based application. A system was specifically implemented to collect speech audio donations from healthy volunteers. It allowed us to form a control group. The system's URL[25] was disclosed through the local news and social networking. After blank samples were eliminated, the resulting dataset was composed of more than 6000 voice donors.

An "appendix" of special recordings was created to address the fact that a COVID-19 ward is a noisy environment: we also collected recordings consisting of pure background noise at the ward, without any voice, typically at the start of a collection session. It is important information, as ward noise is very different from the background noise found in the control group, and noise is a data bias that should be controlled during experiments. Since it is difficult to filter this kind of noise in patient audios, which risks deleting important low-intensity cues to respiratory insufficiency, we decided to gather hospital and device sounds and insert this noise in the control group. It helped address overfitting issues during model training. All collected speech audios contain three different types of utterances:

- Utterance 1, 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: *"O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa"* (*"Love of your neighbor helps strengthening the fight against Coronavirus"*);
- Utterance 2, a well-known nursery rhyme for donors with reading difficulties, due to the lack of 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*

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

small potatoes germinate, branches sprout on the ground; when the baby sleeps, its hands lay over the heart ");

- Utterance 3, a widely known song, was spoken: *"Parabéns a você"* (*"Happy birthday to you"*). The melody is the same as the song in English and the lyrics are: *"Parabéns a você, nesta data querida, muitas felicidades, muitos anos de vida"* (*"Happy birthday to you, on this cherished date, lots of happiness, many years of life"*).

Several issues with the original dataset were identified and addressed: class imbalance, consisting of fewer positive (COVID-19 patients) than negative cases (healthy individuals from the control group); sex imbalance, consisting of a greater number of healthy women than men participating in the process (there were also more men in COVID-19 wards than women); age imbalance, consisting of a higher number of older adults in hospital care than young people in our observations; utterance imbalance, as utterance 1 was more common among patients; healthy people typically recorded all the proposed utterances.

We addressed most 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 the experiments was balanced by class and sex, but not by age, to avoid drastically reducing the available data.

Other issues led to discarding samples collected from the dataset. In some patient audios, the collector's voice could be heard, mostly assisting low literacy or visually-impaired patients when reading the utterance. Some control group recordings exhibited popping and crackling noise, possibly due to the characteristics of the recording devices.

The most serious issue for bias removal 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 the following section.

Signal Description (small data)

The description of speech and voice has been a challenging task for this project's scope, since data collection took place in different environments and different sound capture configurations and equipment. Thus, the aspects selected for the vocal and phonetic/phonological analysis could produce more reliable measures, which will be discussed in the section below titled Proposed measures: temporal and spectral.

Respiratory sounds and linguistic utterances are both viewed as important to the signal described in this study. As described in the previous section, creating a target sentence to utter was necessary. The length and syntactic/prosodic branching of constituents were controlled in the creation of this target sentence. A nursery rhyme and the spoken version of “Parabéns a você” were also part of the dataset, but were not recorded for both groups (patients and control) and therefore not considered for the analysis.

It is widely known that the human voice is multidimensional, since 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 known as dysphonia. The literature reports that 28.6% of individuals infected with COVID-19 showed symptoms of dysphonia (Lechien et al., 2020). Asiaee et al. (2020) explain that a patient with COVID-19 may exhibit decreased or lack of energy for vocal production, leading to an interruption or change in speech production.

In order to carry out voice and linguistic analysis, we built a reliable subdataset (n≅200) that would guarantee suitable answers to our questions. In the next subsections, we present the steps to create our questions, outline the analysis model, provide some details about measures and exhibit preliminary results.

Analysis: general proposal

At the beginning of this study, we expected that the two different groups of speakers (patient versus control) would display significant differences, mainly related to the presence of noisy breathing in the patient’s utterance as opposed to its absence in control group participants. However, based on the advice of the medical doctors of our project, we had to rethink our expectations, after we realized that a severe respiratory condition would not be manifested until a very advanced stage of COVID-19. However, even before listening to and visualizing the acoustic signals of voice and speech, we raised two more general hypotheses to explore: (i) presence of more pauses and (ii) vocal deviation in the patient’s speech.

From these hypotheses, we were able to design a study model that would allow us to treat and analyze the data to answer the following specific questions: (1) Are patient utterances longer than those of the control participants? (2) Are there more pauses and are they longer in patient utterance? (3) Are these pauses in the same grammatical locations in patient and control group utterances? (4) Is the speech rate (for example, syllables per second) of patients lower than that of control subjects? (5) Is the patients' mean fundamental frequency (F0) significantly different from that of the control subjects? (6) Do patients exhibit vocal deviation when compared to the control subjects?

Analysis model

The following analysis model was outlined:

- Two groups of speakers: control group and patient group.
- One target sentence.
- Measures of voice and speech aspects: duration, F0 contour, F0 mean and voice harmonicity.
- Voice aspects will be described by sex.

The target utterance was utterance 1 previously mentioned: *"O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa"* (*"Love of your neighbor helps strengthen the fight against Coronavirus"*).

For the proposed model, analyses were carried out in three domains: temporal, prosodic and spectral. For each of these domains, measures were determined using Praat software, version 6.1.20 (Boersma & Weenink, 2020). In the temporal domain, we measured duration to obtain target sentence length and speech rate. In the prosodic domain, since the target sentence was isolated, we were able to describe the F0 contour and relate it to mean F0. In the spectral domain, in addition to mean F0 per participant, voice harmonicity was determined using the CPPS measure (Cepstral Peak Prominence Smoothed). For the spectral measures, sex was considered.

Proposed measures: temporal and spectral

Using Praat's textgrid annotation resource, target sentence boundaries were obtained and visually isolated from the remaining audio file portions. The criteria used to mark sentence boundaries were the waveform first pulse of the first vowel and the last pulse of the last uttered vowel. This boundary labeling was necessary for both linguistic (phonetic and phonological) and spectral analysis (voice).

Two criteria were used to measure the syllables: (i) a phonological criterion, taking into account the ideal number of syllables (31 syllables) and absence of pauses and (ii) a phonetic criterion, marking actually produced syllables and pauses. It is important to underscore that even for this criterion, in order to avoid infinite detailed segmentation, we used the expected realization in which there is resyllabification, as in / aen / of the *"a enfrentar"* portion (translated: *to face*). Thus, we proposed at least two levels of speech rate: one with 31 syllables and the other with around 26. The speech rate was calculated using the syllable/sentence duration ratio (sr=syllable/sentence duration).

In order to extract F0 and harmonicity voice parameters, the audio recordings were edited to minimize external interference at the time of recording and consider only the participants' continuous speech. Thus, we excluded portions of pauses between one vocalization and another as well as portions with device noises or the voice of a health professional, given that they could interfere with the extraction of the acoustic measures of voice. Our choice to obtain a CPPS measure relies on the fact that it shows the extent to which F0 harmonics are individualized and stand out regarding the noise level present in the acoustic signal (Asiaee et al., 2020). It is worth highlighting that in relation to vocal parameters (F0 and CPPS), the two groups were divided by sex, since male and female voices are different when it comes to the mechanics of vocal fold vibration and signal energy.

In order to determine whether the proposed questions and measures were promising, a very small data subset was analyzed in this initial stage of our study. Syllable and pause duration have yet to be extracted, implying we still do not have precise data on speech rate. However, some preliminary results were obtained, since we were able to use what we call first level speech rate. It consisted of dividing the fixed number of syllables (31) by the actual sentence duration. With respect to voice parameters, some initial observations indicated that F0 values seem to differentiate the two male groups and CCP values the two female groups.

Preliminary results

In the temporal domain, results obtained from a data subset (n=100) indicate a slight difference between groups, in both total duration (I) and speech rate (II) - syllables per second. The patient group (PG n=50) has a tendency to produce longer utterances (average=7.87s) and fewer syllables per second (average = 4.23 syl/s), which may be related to the duration and number of pauses in patient speech, caused by respiratory insufficiency. On the other hand, for the control group (CG, n=50), the average total duration was 5.40s and the speech rate 5.88 syl/s. In addition, the control group had a smaller standard deviation in both conditions (CGsd I= 0.86, II= 0.92; PGsd I = 2.22; II=0.92).

In the spectral domain, preliminary results indicate a slight difference between female groups in F0 standard deviation (SD) and CPPS measures. The female patient group (PG) showed more unstable emission. This might be due to poorer control in sustaining F0 (PG, F0 SD=36.66Hz) compared to the control group (CG, F0 SD=22.35Hz). In addition, the female PG emitted more noise in the vocal signal, when we compared harmonic behavior (PG, CPPS=7.93dB) between groups, considering the sex variable (CG, CPPS=10.127dB). Females also exhibited more vocal deviation than their male counterparts. Males in the patient group had a higher F0 in relation to the controls (PG = 116.5 Hz; CG=133.02 Hz).

These results indicate that differences in temporal, prosodic and spectral domains may be found between groups. For the large dataset to be analyzed, measures need to be automatically extracted (see the next section of this chapter), which may account for temporal and spectral analysis. An intense dialogue with signal processing colleagues is ongoing to solve problems related to the acoustic signal and corresponding linguistic units.

Signal Processing

The signal processing team at the SPIRA project is involved in two tasks, namely the extraction of features that are important for linguistic and vocal signal descriptions (as outlined in the previous section), and the production of alternative representations that are relevant for machine learning and input signal classification (as described in the next section).

Segmentation

Audio signal segmentation is a preliminary step for many subsequent signal processing tasks. The first segmentation level consists of identifying speech utterances and background noise, which may include sound-producing electrical appliances and other voices (an occurrence that affects mainly the recordings of hospitalized patients).

A straightforward, albeit not perfect, approach is to use energy thresholding to identify the segments containing the main speaker. In this method, the level (in dB) of noise floor is estimated for each audio signal from the minimum values of the energy curve, and a threshold above this noise floor is used for the binary classification of each audio frame; subsequent smoothing (e.g., majority vote) can be applied to avoid rapid alternation between speech and noise segments.

Finer levels of segmentation (e.g., phonetic segmentation) can be obtained using other classification strategies, such as voiced/unvoiced classification, phoneme detection, etc.

Noise reduction

Reducing background noise is important for both improving feature extraction and creating alternative representations for machine learning. Concentrating spectral information on the parts that most likely belong to the speech utterance improves the signal-to-noise ratio, stabilizes the estimation of fundamental frequency (F0) and improves peak-to-valley measurements in both spectrum and cepstrum (e.g., CPP), among other benefits.

Noise gating is a well-known technique for noise reduction based on a gaussian representation of the noise spectrum. Using speech/noise segmentation to define a gaussian model of noise allows the training of an adaptive non-linear filter that selectively suppresses or attenuates specific time-frequency components within the signal's spectrogram, which may then be resynthesized as a new noise-reduced audio signal.

Feature extraction and augmented representations

Many audio features are easily obtained from the original signal and metadata such as speech/noise segmentation (Mitrović, Zeppelzauer, & Breiteneder, 2010). These additional metadata are relevant for linguistic and vocal signal description and investigation of discriminating parameters that would allow the identification of speech utterances affected by respiratory insufficiency, as well as for producing augmented representations in the context of machine learning, since metadata that are known to facilitate the classification of affected and healthy individuals would also probably ease the convergence of hyperparameters during training of automatic learning models.

Speech/noise segmentation provides the first source of many relevant features that may be useful to both description and classification. From this simple on-off description of the signal, one can obtain information on the number and duration of continuous speech utterances and interruptions, such as number of noise segments (a rough proxy for respiratory rate), the ratio between continuous speech duration and the signal's total duration, mean and variance of the duration of both continuous speech utterances and interruptions, among others.

Pitch and timbre-related descriptions may be easily obtained from F0 profiles, such as using pYIN (Mauch & Dixon, 2014), spectrograms, cepstral representations including MFCC (Hibare & Vibhute, 2014)) and harmonic representations such as HPCP (Gomez & Herrera, 2004), among others, obtained for the entire signal or for segments with continuous speech utterances. Several statistical measures can be derived from these representations, such as mean/median/std/min/max of F0 and cepstral peak prominence, which are already under investigation, as well as the characterization of voice formants and voiced/unvoiced segmentation.

Annotation transfer between signals

Automatic labeling of signals according to phonetic information, such as syllable or phoneme transcription, is a difficult task prone to a number of errors even with state-of-the-art techniques. An easier alternative is to transfer labels from signals that already received these manual annotations (which is also difficult and time-consuming for humans). This can be done by exploiting the fact that several recordings have the same spoken sentence, and by aligning recordings that receive manual annotations to those that did not.

Dynamic Time Warping is an algorithm for time-aligning two symbolic sequences that use dynamic programming to build a map of timestamp correspondences between the two sequences. It depends heavily on the choice of a representation that produces similar symbolic sequences for two speech utterances of the same sentence, regardless of the speaker's individual timbre characteristics. This requires representations such as cepstral coefficients, which are less sensitive to pitch or energy variations (related to prosody and thus varying significantly between speakers), and more sensitive to the differences between the spectral structure of the phonemes.

Signal Classification (big data)

For machine learning purposes, the dataset was divided into training (292 audios), validation (292) and tests (108), as is usual in statistical learning. We selected audios with the best signal-to-noise ratio for the test set, and the second best audios were used for validation. The aim of this partitioning is to detect training overfitting. The method chosen to classify the input signal was based on artificial neural networks. The MFCC representation is extracted from the audio and then presented to a convolutional neural network. The first step in this process is to preprocess the audios obtained.

In general, the majority of the audios in the dataset were sampled at 48kHz. We pre-processed these files using Torch Audio 0.5.0 as follows: 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 & Morrow, 1967), with hop length 160 and 1,200 FFT components, retaining only 40 coefficients. However, before applying the MFCC feature extraction process, we need to address the duration difference observed in our data.

Ward noise is a serious source of bias. In this scenario, a neural network can be biased during training by focusing only on background noise. To address this issue, we injected pure background noise samples obtained from COVID wards into patient and control group audios. A total of 16 1-minute samples were recorded. To avoid bias in this process, we decided to inject noise into all training and validating samples for both patients and the control group. We also injected noise into some testing samples in order to check for model bias. The test and validation sets were created in such a way as to allow overfitting detection, since they are composed mostly of audios with a very limited amount of noise.

Proposed Model

Several models were tested in preliminary experiments and we describe the one that provided the best results. Regarding topology and model parameters, preliminary experimental results showed that CNNs (Convolutional Neural Networks) applied to MFCCs are useful in analyzing this type of problem. Figure 13.1 presents the selected 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 & Koltun, 2015) by D; and fully connected layers by FC. The input size varies according to the experiment. We investigated the use of Mish activation function (Misra, 2019) due to its regularization effects during training, which helps prevent overfitting.

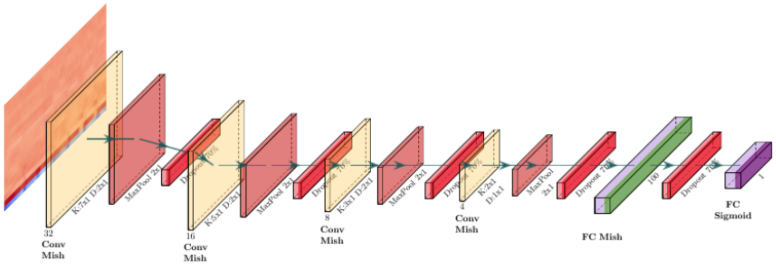


Figure 13.1. CNN topology proposed with four convolutional and two fully connected layers.

We used Binary Cross-Entropy as loss, and Adam optimizer (Kingma & Jimmy, 2014). The initial learning rate was set at 10^{-3}, and the Noam's decay scheme (Vaswani, 2017) was applied every 1,000 steps. For each proposed experiment, we trained the model for 1,000 epochs using a batch size of 30. With respect to regularization, overfitting mitigation is a major concern, given our dataset noise characteristics. As such, several approaches for regularization were applied. In addition to 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 the output layer. Finally, after each convolutional layer, we applied group normalization (Wu & He, 2018) to pairs of convolution filters. Thus, the number of groups is half the number of filters.

Our models were implemented using PyTorch 1.5.1. We ran the experiments on a NVIDIA Titan V GPU with 12GB RAM on a server with an Intel Core i7-8700 CPU and 16GB of RAM.

Experiments and Results

Experiments were projected to determine the optimal amount of noise insertion. Note that better results were sometimes obtained without noise in test samples and vice versa. In general, bias is greatly reduced by inserting at least one noise sample into the negative instances. As expected, inserting too much noise decreases model performance. The best overall accuracy was obtained in 3 noise samples, which reached 91% accuracy in the task. The accuracy of each experiment is presented in Figure 13.2, both with and without artificial insertion of ward noise into the test samples.

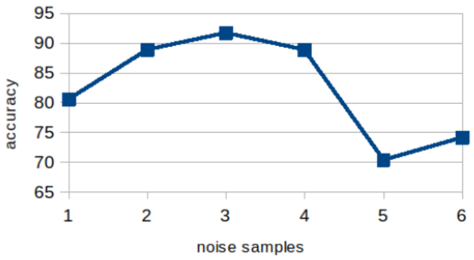


Figure 13.2. Accuracy obtained per number of noise samples inserted in training data

Concluding remarks

The initial results obtained for the SPIRA project seem to validate the original assumption of the project that respiratory insufficiency can be detected to an acceptable level of accuracy from audio signals obtained by remote recordings. Thus, we are encouraged to develop a pre-diagnostic assistance tool to help health professionals in patient triage.

Future work will involve detailed descriptions of the signal properties of patients and non-patients, as well as an extension of the current study to address 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

Asiaee, M., Vahedian-Azimi, A., Atashi, S. S., Keramatfar, A., & Nourbakhsh, M. (2020). Voice quality evaluation in patients with COVID-19: An acoustic analysis. *Journal of Voice*, https://doi.org/10.1016/j.jvoice.2020.09.024

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.26 from http://www.praat.org/

Botelho, M. C., Trancoso, I., Abad, A., & Paiva, T. (2019, May). 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.

Brigham, E. O., & Morrow, R. E. (1967). The fast Fourier transform. IEEE spectrum, 4(12), 63-70. https://doi.org/10.1109/MSPEC.1967.5217220

Giovanni, A., Radulesco, T., Bouchet, G., A Mattei, A. J Révis, J., Bogdanski, E., Michel, J. (2021). 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*. https://doi.org/10.1007/s00405-020-06200-7

Gómez, E., & Herrera, P. (2004, October). Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies. In *Proceedings of the International Symposium for Music Information Retrieval (ISMIR*), pp. 92–95.

Hassanpour, S., Langlotz, C. P., Amrhein, T. J., Befera, N. T., & Lungren, M. P. (2017). 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.

Hibare, R., & Vibhute, A. (2014). Feature extraction techniques in speech processing: A survey. *International Journal of Computer Applications, 107(5)*, 1-8. https://doi.org/10.5120/18744-9997

Kent, R. D. (1997). *The speech sciences*. San Diego: Singular Publishing Group, .

Kingma, D. P. and Ba, J. L. (2014). Adam: A method for stochastic optimization, arXiv:1412.6980, 201

Laguarta, J., Hueto, F., & Subirana, B. (2020). COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 275-281. https://doi.org/10.1109/OJEMB.2020.3026928

Lechien, J. R., Chiesa-Estomba, C. M., Cabaraux, P., Mat, Q., Huet, K., Harmegnies, B., ... & Saussez, S. (2020). Features of mild-to-moderate COVID-19 patients with dysphonia. *Journal of Voice*.https://doi.org/10.1016/j.jvoice.2020.05.012.

Mauch, M., & Dixon, S. (2014, May). 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.

Misra, D. (2019). Mish: A self-regularized non-monotonic neural activation function. arXiv preprint arXiv:1908.08681, 4. https://arxiv.org/pdf/1908.08681.pdf

Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Features for content-based audio retrieval. In Marvin V. Zelkowitz (ed.) Advances in computers: Improving the Web (Vol. 78, pp. 71-150). London, UK: Academic Press

Nevler, N., Ash, S., Irwin, D. J., Liberman, M., & Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology*, *6*(1), 4-14.

Orlandic, L., Teijeiro, T., & Atienza, D. (2020). The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms. *arXiv preprint arXiv:2009.11644*.

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., ... & Hillman, R. (2018). 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. https://doi.org/10.1044/2018_AJSLP-17-0009

Quatieri, T. F., Talkar, T., & Palmer, J. S. (2020). A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 203-206.10.1109/OJEMB.2020.2998051. Retrieved May, 2021 from: https://ieeexplore.ieee.org/abstract/document/9103574

Spasić, I., Owen, D., Smith, A., & Button, K. (2019). KLOSURE: Closing in on open–ended patient questionnaires with text mining. *Journal of Biomedical Semantics*, *10*(1), 1-11. https://doi.org/10.1186/s13326-019-0215-3

Tailor, S. A., Chauhan, J., & Mascolo, C. (2020, September). 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).https://doi.org/10.1145/3410530.3414440

Tobin, M. J., Laghi, F., & Jubran, A. (2020). Why COVID-19 silent hypoxemia is baffling to physicians. *American Journal of Respiratory and Critical Care Medicine*, *202*(3), 356-360. https://doi.org/10.1164/rccm.202006-2157CP

Trancoso, I., Correia, M. J. R., Teixeira, F., Abad, A., Botelho, M. C. T. , and Raj, B. (2019).. Speech as a (private?) biomarker for speech affecting diseases. In ICIEA 2019 -- The 14th IEEE Conference on Industrial Electronics and Applications. Keynote paper ed. Xi'an, China: IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N, Polosukhi. (2017). Attention is all you need. In: Advances in neural information processing systems. pp. 5998-6008.

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Wu, Y., & He, K. (2018). Group normalization. In: *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19). https://link.springer.com/chapter/10.1007/978-3-030-01261-8_1