Audio MFCC-gram Transformers for respiratory insufficiency detection in COVID-19

Marcelo Matheus Gauy[1], **Marcelo Finger**[1]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)

Abstract. *This work explores speech as a biomarker and investigates the detection of respiratory insufficiency (RI) by analyzing speech samples. Previous work [Casanova et al. 2021] constructed a dataset of respiratory insufficiency COVID-19 patient utterances and analyzed it by means of a convolutional neural network achieving an accuracy of* 87.04%*, validating the hypothesis that one can detect RI through speech. Here, we study how Transformer neural network architectures can improve the performance on RI detection. This approach enables construction of an acoustic model. By choosing the correct pretraining technique, we generate a self-supervised acoustic model, leading to improved performance (*96.53%*) of Transformers for RI detection.*

1. Introduction

COVID-19 is the cause of a major pandemic that threatens to collapse the healthcare systems in many regions of the world. Respiratory insuficiency (RI) is one of COVID-19 symptoms, which often requires hospitalization and is aggravated by a common COVID-19 condition called *silent hipoxia*, low blood oxygen concentration without breath shortness [Tobin et al. 2020]. This work aims to help deal with the COVID-19 pandemic by providing an automated system, based on deep learning techniques, capable of detecting RI in COVID-19 patients. Such an automated system could, for example, support cellphone-based patient triage procedures alleviating the burden on health personnel.

We explore the view of *speech as a biomarker*, by building upon a recently shown fact: it is possible to detect respiratory insufficiency through analyzing spoken utterances in real-life conditions (typically a moderately large sentence). This hypothesis has been previously verified [Casanova et al. 2021] by using a CNN-based deep neural network. This CNN received a moderately large sentence spoken in real life conditions and had to predict whether it came from a patient with RI or from the control group. In this work, we aim to further analyze that hypothesis by studying other network architectures (namely, Transformers [Vaswani et al. 2017]), in an attempt to improve the results previously obtained in [Casanova et al. 2021], with a view of extending it in the future to RI originated from other causes, such as influenza, heart disease or mental illness.

In this work we find that Transformers can be used for detecting respiratory insufficiency with an accuracy of 96.38% up from 87.04% in [Casanova et al. 2021]. To reach that level of performance, we feed the Transformers with a sequence of Mel Frequency Cepstral Coefficients (MFCC) obtained from the patients' audios (henceforth called MFCC-gram Transformers). Like CNN-based detection from [Casanova et al. 2021], the Transformer performance drops significantly (to 82.87%) if we feed it standard spectrogram coefficients (called Spectrogram Transformers after [Gong et al. 2021]).

The Transformers [Vaswani et al. 2017] were shown to be very effective when divided in two parts [Devlin et al. 2018]. The *pretraining phase* generates a language-based acoustic model with unsupervised (or self-supervised) learning by optimizing a generic language prediction task with a large amount of generic data. Then, the acoustic model undergoes a task-specific *refinement phase* in which both the acoustic model and additional task-specific neural modules are trained on smaller-size application data. A *baseline transformer* is one in which pretraining is a random assignment of weights.

Here, we find that MFCC-gram Transformers benefit from being pretrained with large quantities of spoken Brazilian Portuguese audios, which is later refined for the target task of detecting respiratory insufficiency. For pretraining, we explore three known techniques from the literature [Liu et al. 2020b, Liu et al. 2020a] and find that they generally lead to some performance improvement over baseline transformers. Performance reaches 96.53% using the best of the available techniques.

2. Related Work

In addition to [Casanova et al. 2021] there have been other works [Pinkas et al. 2020, Laguarta et al. 2020] which study COVID-19 with deep learning using voice related data. [Pinkas et al. 2020] attempt to detect SARS-COV-2 (the virus that causes COVID-19) from voice audio data, while this work and [Casanova et al. 2021] attempt to detect RI. Furthermore, there have been previous works which support the view of speech as a biomarker [Botelho et al. 2019, Nevler et al. 2019, Robin et al. 2020].

Transformers were designed for NLP [Vaswani et al. 2017, Devlin et al. 2018], and were also later used in audio processing tasks [Liu et al. 2020b, Liu et al. 2020a, Schneider et al. 2019, Baevski et al. 2020, Baevski et al. 2019, Song et al. 2019]. In Mockingjay and Tera [Liu et al. 2020b, Liu et al. 2020a], it was used in phoneme classification and speaker recognition tasks. There it was shown that variants of the Cloze task [Taylor 1953, Devlin et al. 2018] for audio could be used for unsupervised pretraining of Transformers. In Wav2Vec and its variants [Schneider et al. 2019, Baevski et al. 2020, Baevski et al. 2019], a contrastive loss is used to enable unsupervised pretraining, which is later finetuned to speech and phoneme recognition tasks. In Speech-XLNet [Song et al. 2019], a speech based version of the XLNet [Yang et al. 2019] was proposed. The XLNet is a network that maximizes the expected log likelihood of a sequence of words with respect to all possible autoregressive factorization orders.

3. Methodology

3.1. Datasets

For the task of respiratory insufficiency detection, the data used in the refinement phase is the same one used in [Casanova et al. 2021]. There, COVID patient utterances were collected by medical students at COVID wards from patients with blood oxygenation level below 92%, as an indication of RI. Control data was collected by voice donations over the internet without any access to blood oxygenation measurements and were therefore assumed healthy. As COVID wards are noisy locations, an extra collection was made consisting of samples of pure background noise (no voice). This is a crucial step in preventing the network to overfit to the background noise differences in data collection.

The gathered audios contained 3 utterances:

- A long sentence with 31 syllables. It was designed by linguists to be long enough to have reading pauses while being simple for even low literacy donors to speak.
- A widely known nursery rhyme for readers with reading impediments.
- A well known song along the lines of 'Happy birthday to you'.

As suggested in [Casanova et al. 2021], we select only audios from the first utterance and sample balance the dataset by class and sex. The presence of ward background noise in the patient audios is treated in a similar way: we insert noise to the control group as that is easier than removing it from the patients' signal. This prevents that we eliminate from the signal, audio that is relevant to the network's classification.

We employ the same division in training, validation and test as done in [Casanova et al. 2021]. The best signal-noise ratio audios are included in the test set. The second best audios are in the validation set. This is done to detect training overfitting. Table 1 contains information on the number of audio files for each class.

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

Table 1. Filtered dataset information.

For the pretraining phase, we use datasets containing Brazilian Portuguese speech. These datasets are NURC-Recife [Oliviera Jr et al. 2016], ALIP [Gonçalves 2019], C-Oral Brasil [Raso and Mello 2012] and SP2010 [Mendes 2013]. Together, they contain more than 200 hours of Brazilian Portuguese speech.

3.2. Preprocessing

As we face similar audio processing issues as [Casanova et al. 2021], we employ similar preprocessing steps. In the dataset, the majority of audios were sampled at $48kHz$. We preprocess the files using Torchaudio $0.9.0$. We extract either the mel-spectrogram (for Spectrogram Transformers) or the MFCCs of the audios with default Torchaudio parameters and retain 128 coefficients. Torchaudio, by default, employs a Fast Fourier Transform [Brigham and Morrow 1967] with a $400ms$ window and hop length 200.

As the dataset has an inherent imbalance in the audio lengths from patients and control we do not use the full audios of the first utterance. Instead, we break each audio into 4 seconds chunks, with a windowing of 1 second steps. Such a windowing method was observed in [Casanova et al. 2021] to be more effective than, for example, padding the audios with zeros to make all the audios have the same length. The windowing technique solves the problem of the imbalance between audio lengths and guarantees the network will not pay too much attention to the audio lengths and instead focuses on the content. The windowing technique also serves as a kind of data augmentation as, for example, an audio with 8 seconds becomes 5 audios with 4 seconds. We observe that the windowing should be done before the spectrogram or MFCCs feature extraction.

3.3. Noise insertion

The noise in COVID wards is a serious bias source. This can be seen in our experiments and in the original work with the dataset by [Casanova et al. 2021]. One potential way of dealing with this bias source is to filter the noise and eliminate it. However, this has the risk that we eliminate important low-energy information from the data, information which would have been useful in detecting whether a patient had RI. Moreover, eliminating the noise could also create extra biases, as different procedures for eliminating patient and control noises would be required. Thus, instead of eliminating the noise, we consider it much easier to insert the noise present in the COVID wards into all the audio samples.

The original dataset contained 16 samples of 1 minute each containing just the background noise present in COVID-19 wards. These noise samples are added to all the training, validation and test audios, similarly to what was done in [Casanova et al. 2021]. We experiment with the amount of noise we add to each of the audio files. During training, audio samples are injected with one or more noise samples. These are selected randomly from the pool of noise samples each time an audio is used for training. The starting point of each noise sample is also selected randomly. Lastly, a factor to change the intensity of the sample is drawn. This factor is limited by a maximum amplitude value which depends on the patient audio noises. This process is similar to the one in [Casanova et al. 2021] and the goal is inserting noise as similar to the pre-existing noise as possible.

3.4. Transformers

We consider two types of Transformers: MFCC-gram Transformers and Spectrogram Transformers. They are equivalent except in the data features that are fed to them: MFCC-gram Transformers receives MFCC audio features and Spectrogram Transformers receive mel spectrogram audio features. Our Transformers are equivalent to the Transformer Encoder units described in [Vaswani et al. 2017]. Namely, we use a multi-layer Transformer encoder (3 layers) with multi-head self-attention. Each encoder layer has two sub-layers, the first being a multi-head self-attention network and the second being a fully connected feed-forward layer. Each sub-layer has a residual connection followed by layer normalization [Ba et al. 2016]. Every encoder layer and sub-layers produce outputs of dimension 512. In addition to the attention sub-layers, each encoder layer contains a fully connected feed-forward network with an inner layer of dimension 2048.

In order to generate the sequence of tokens that is sent to the Transformers the MFCC and/or Spectrogram is split into its frames. Each frame of the MFCC or spectrogram corresponds to one token fed to the sequence. We also attempted joining multiple frames into one token but this typically produced worse results than the one to one framework. We use sinusoidal positional encoding [Vaswani et al. 2017, Liu et al. 2020b, Pham et al. 2019] to make our model position aware. As suggested by [Liu et al. 2020b], each frame is first projected linearly to a hidden state of dimension 512.

Our Transformers are trained in two phases: pretraining and refinement. In the pretraining phase, we leverage the unsupervised training techniques described in Section 3.5 to build an acoustic model over generic audio data. In the refinement phase, the pretrained Transformers is refined over COVID related audio data. For some experiments, we bypass the pretraining phase by initializing the Transformers with a random assignment of weights and refining that over the COVID data. This is done to get a baseline performance and we call these Transformers the baseline Transformers. We will name our Transformers types baseline MFCC-gram Transformers and baseline Spectrogram Transformers when we consider Transformers which bypass the pretraining phase.

Our code is based on the guide “The annotated Transformer”[1]. While our Transformers are small in comparison to the ones used, e.g. in BERT [Devlin et al. 2018], the amount of available data for respiratory insufficiency detection is also rather small so we do not expect that larger Transformers would yield significantly improved results. Once more data is available, it is recommended to also increase our Transformers.

3.5. Unsupervised pretraining: acoustic model construction

We describe three techniques to pretrain acoustic models in a self-supervised way. They are based off Masked acoustic modelling [Liu et al. 2020b]. This erases a fraction of the input and tries to reconstruct the erased parts from the remaining frames. They are bidirectional methods and the reconstruction depends on both left and right contexts.

Time Alteration: also called Masked acoustic modelling [Liu et al. 2020b]. Start by selecting frames up to 15% of the input[2], 1) mask them all to zero 80% of the time, 2) replace all with a random frame 10% of the time or 3) leave the frames be in the remaining 10% of the time. The goal of this process (as opposed to always masking the frames) is to alleviate the mismatch between training and inference.

Channel Alteration: this techinique is from [Liu et al. 2020a]. Randomly mask a block of consecutive quefrency channels to zero for all time steps of the input sequence. First, the width W_C of the block is selected uniformly from $\{0, 1, \ldots, W\}$ where W is a 10% fraction of the total number of channels. Second, sample a channel index I_C from $\{0, 1, \ldots, H - W_C - 1\}$ where H is the total number of channels in the input. Then, channels from I_C to $I_C + W_C - 1$ are masked to zero. Observe that (as with time alteration), a fraction of the time none of the channels will be masked.

Noise Alteration this technique is from [Liu et al. 2020a]. Apply sampled Gaussian noise to change the magnitude of the inputs with a probability of 10%. For that end, we sample a random magnitude matrix with the same size as the input. Each element in the matrix is sampled from a normal distribution with mean zero and 0.2 variance. The matrix is then added to the real input frames.

4. Results and Discussion

Here we show the results obtained by the two experiments performed: the first where we compare baseline MFCC-gram Transformers, baseline Spectrogram Transformers and the CNN from [Casanova et al. 2021], and the second where we try different unsupervised pretraining techniques to improve baseline Transformers by building an acoustic model.

First, we note that when no ward noise is added to either the patient or control files, baseline MFCC-gram Transformers performs very well (98.89 ± 0.38) in the test files. However, this performance drops dramatically (to 70.07 ± 3.15) if we add noise to the test files and this is a strong sign the model is biased by the noise. This bias is less

[1] `http://nlp.seas.harvard.edu/2018/04/03/attention.html`

[2] More precisely, we select a fraction of the frames in chunks of a certain size so that the total number of frames masked amounts to 15%. In the experiments, the chunk size was 7.

extreme than what was observed at the MFCC-gram CNN in [Casanova et al. 2021] but is still present. Therefore, in our experiments, noise is added to the training and test files.

In the first experiment, we consider baseline Transformers and bypass the pre-training phase. We vary the amount of ward noise we add to the training and test files. We add to the audio files between 0 and 3 noise files, including either the same amount of noise files to the patient and control audio files or one more file to the control files. This is comparable to the Experiments $3.x$ from [Casanova et al. 2021] and we can directly compare baseline MFCC-gram Transformers, baseline Spectrogram Transformers with the CNN from [Casanova et al. 2021]. We perform each experiment for 20 epochs and repeat the experiments 10 times. The batch size is set to 16. The results are in Table 2. We show both the performance when including noise as well as the performance without including noise in the test samples. Figure 1 shows the same data as Table 2.

Model	Noise Samples		Accuracy (with noise in test samples)	Accuracy (without noise in test samples)
	Patient	Control		
Baseline MFCC-gram Transformers	0	1	$\mathbf{96.38 \pm 0.72}$	96.85 ± 0.84
	1	1	96.30 ± 1.12	$\mathbf{97.36 \pm 1.89}$
	1	2	95.39 ± 1.26	96.44 ± 1.72
	2	2	95.68 ± 0.48	97.35 ± 1.01
	2	3	94.33 ± 1.48	96.53 ± 1.13
	3	3	94.86 ± 0.75	96.63 ± 1.09
MFCC-gram CNN	0	1	74.07 ± 1.93	61.11 ± 8.40
	1	1	86.11 ± 2.98	66.67 ± 3.74
	1	2	83.33 ± 3.34	84.26 ± 6.17
	2	2	85.19 ± 0.93	88.89 ± 0.53
	2	3	85.19 ± 1.85	74.07 ± 5.10
	3	3	87.04 ± 0.93	91.67 ± 2.98
Baseline Spectrogram Transformers	0	1	82.87 ± 1.48	68.73 ± 3.88
	1	1	82.84 ± 1.82	82.65 ± 2.25
	1	2	80.75 ± 2.22	77.18 ± 1.56
	2	2	79.02 ± 3.19	82.05 ± 2.57
	2	3	78.07 ± 2.78	74.67 ± 1.99
	3	3	78.73 ± 2.33	81.96 ± 1.97

Table 2. The performance of the Transformers and the CNN is shown in Table 2. The different lines show performance of the network according to the number of noise files added to the test files, both for patients and control.

We observe a significant improvement in performance for baseline MFCC-gram Transformers when compared to the MFCC-gram CNN. When including noise in test samples, the best performance is attained by baseline MFCC-gram Transformers where we add a single noise file to the control files and keep the patient files unchanged. When we compare without noise being added the best performance is attained by baseline MFCC-gram Transformers where we add a single noise file to both the patient and control files. We would like to point out though that the differences are rather small and baseline MFCC-gram Transformers performs well as long as some noise is added.

For the second experiment, we fix the amount of ward noise we insert to the train-

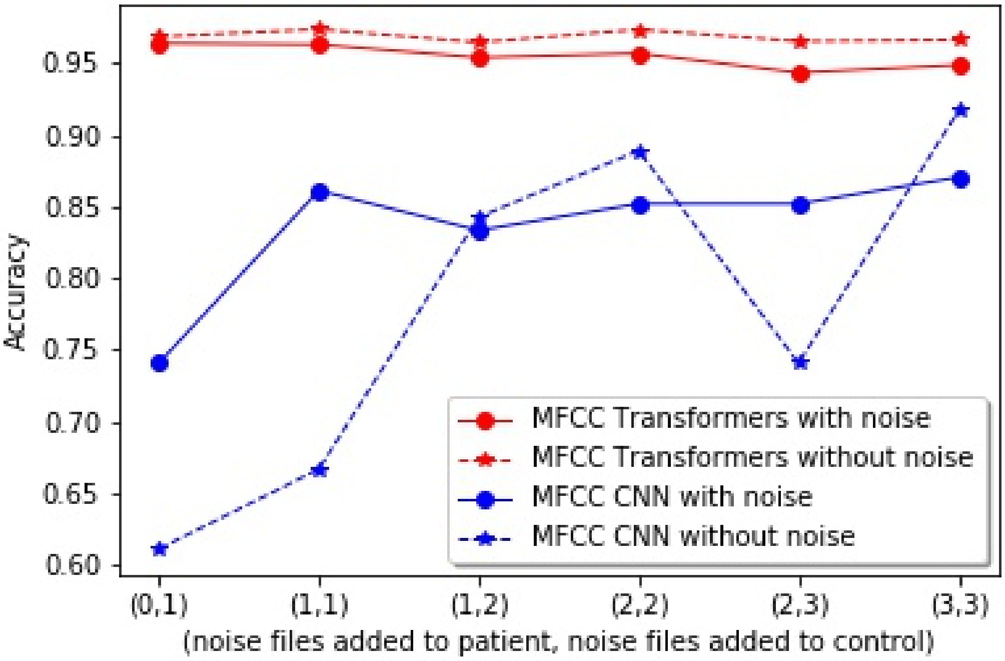

Figure 1. This has the same data as Table 2. The y axis shows accuracy and the x axis shows the number of noise files added to patient and control files.

ing and test files to be a single noise file for both patient and control audio files. We vary the technique employed for unsupervised pretraining, attempting time alteration, channel alteration and noise alteration techniques as described in Section 3.5. We pretrained on the corpuses of NURC-Recife, C-Oral Brasil, SP 2010 and ALIP. Pretraining consisted of 5 epochs on the data of all those corpuses, splitting each file into 4 seconds audio with a 1 second window step. Finetuning on the respiratory insufficiency data was performed in 20 epochs and repeated 10 times so the results are averaged. We show the performance of each for both MFCC-gram Transformers and Spectrogram Transformers in Table 3.

We observe a small improvement (over the baseline) using time alteration when we test MFCC-gram Transformers including noise in the test files. We also observe an improvement using noise alteration when we test MFCC-gram Transformers without including noise in the test files. In principle, one could combine these techniques as they are independent ways of masking the input. We have done that by performing all three techniques at the same time as shown in the table. Note that the performance of Spectrogram Transformers increases even more robustly than that of MFCC-gram Transformers.

5. Conclusion and Future work

By employing a Transformers network to the dataset of respiratory insufficiency from COVID-19 detection created in the paper [Casanova et al. 2021], we improved the performance of their CNN network from 87.04% to 96.38%. Moreover, we found that MFCC and Spectrogram based Transformers improve their performance through unsupervised pretraining on a large amount of unlabeled data.

Model	Pretraining type	Accuracy (with noise in test samples)	Accuracy (without noise in test samples)
MFCC Transformers	Baseline	96.30 ± 1.12	97.36 ± 1.89
	Time Alteration	$\mathbf{96.53 \pm 0.71}$	97.00 ± 1.55
	Channel Alteration	96.15 ± 0.84	97.04 ± 1.52
	Noise Alteration	95.93 ± 0.66	98.21 ± 0.89
	Time + Channel + Noise	96.38 ± 1.24	$\mathbf{98.54 \pm 1.56}$
Spectrogram Transformers	Baseline	82.84 ± 1.82	82.65 ± 2.25
	Time Alteration	80.99 ± 3.49	87.90 ± 2.75
	Channel Alteration	82.41 ± 1.75	87.53 ± 2.25
	Noise Alteration	80.61 ± 1.32	86.08 ± 2.70
	Time + Channel + Noise	81.67 ± 1.51	86.93 ± 2.61

Table 3. The performance of the Transformers network is compared when unsupervised pretraining is done. The different pretraining techniques are compared for MFCC-gram and Spectrogram Transformers. We fix the amount of noise insertion to be one noise file inserted at patient and control files.

Future work could involve augmenting the dataset with audios from patients of many more respiratory illnesses besides COVID-19. Moreover, we could ideally get audio from patients and control under similar conditions. Furthermore, one could attempt improving the performance of Spectrogram Transformers so that they match the performance of MFCC-gram Transformers. Moreover, we currently train our acoustic model in the single task of respiratory insufficiency detection. It would be interesting to extend our model for other tasks, creating the first acoustic model of spoken Brazilian Portuguese.

6. Acknowledgement

We would like to thank the LNCC for providing us with the computational resources required to do this work. All experiments were run in the LNCC servers. This work was supported by FAPESP grant number 2020/16543-7.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baevski, A., Schneider, S., and Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Botelho, M. C., Trancoso, I., Abad, A., and Paiva, T. (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.

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

Casanova, E., Gris, L., Camargo, A., Silva, D., Gazzola, M., Sabino, E., Levin, A., Candido Jr, A., Aluisio, S., and Finger, M. (2021). Deep learning against covid-19: Respiratory insufficiency detection in brazilian portuguese speech. *To appear in ACL2021.*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Gonçalves, S. C. L. (2019). Projeto alip (amostra linguística do interior paulista) e banco de dados iboruna: 10 anos de contribuição com a descrição do português brasileiro. *Estudos Linguísticos (São Paulo. 1978)*, 48(1):276–297.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778.*

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

Liu, A. T., Li, S.-W., and Lee, H.-y. (2020a). Tera: Self-supervised learning of transformer encoder representation for speech. *arXiv preprint arXiv:2007.06028.*

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020b). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

Mendes, R. B. (2013). Projeto sp2010: Amostra da fala paulistana. *http://projetosp2010. fflch. usp. br¿. Acesso em*, 1(12):2013.

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

Oliviera Jr, M. et al. (2016). Nurc digital um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nurc). *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 3(2):149–174.

Pham, N.-Q., Nguyen, T.-S., Niehues, J., Müller, M., Stüker, S., and Waibel, A. (2019). Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377.*

Pinkas, G., Karny, Y., Malachi, A., Barkai, G., Bachar, G., and Aharonson, V. (2020). Sars-cov-2 detection from voice. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:268–274.

Raso, T. and Mello, H. (2012). The c-oral-brasil i: reference corpus for informal spoken brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 362–367. Springer.

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

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Song, X., Wang, G., Wu, Z., Huang, Y., Su, D., Yu, D., and Meng, H. (2019). Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks. *arXiv preprint arXiv:1910.10387*.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.