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

# The Phonetic Footprint of Covid-19?

*P. Klumpp*[1], *T. Bocklet*[2], *T. Arias-Vergara*[1,3,4], *J. C. Vásquez-Correa*[1,3], *P. A. Pérez-Toro*[1,3], *S. P. Bayerl*[2], *J. R. Orozco-Arroyave*[1,3], *E. Nöth*[1]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany
[2]Technische Hochschule Georg Simon Ohm, Nürnberg, Germany
[3]Faculty of Engineering, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[4]Department of Otorhinolaryngology, Head and Neck Surgery. Ludwig-Maximilians University, Munich, Germany

philipp.klumpp@fau.de

## Abstract

Against the background of the ongoing pandemic, this year's Computational Paralinguistics Challenge featured a classification problem to detect Covid-19 from speech recordings. The presented approach is based on a phonetic analysis of speech samples, thus it enabled us not only to discriminate between Covid and non-Covid samples, but also to better understand how the condition influenced an individual's speech signal.

Our deep acoustic model was trained with datasets collected exclusively from healthy speakers. It served as a tool for segmentation and feature extraction on the samples from the challenge dataset. Distinct patterns were found in the embeddings of phonetic classes that have their place of articulation deep inside the vocal tract. We observed profound differences in classification results for development and test splits, similar to the baseline method.

We concluded that, based on our phonetic findings, it was safe to assume that our classifier was able to reliably detect a pathological condition located in the respiratory tract. However, we found no evidence to claim that the system was able to discriminate between Covid-19 and other respiratory diseases.

**Index Terms**: ComParE, Covid-19 recognition, phonetic speech analysis

## 1. Introduction

Unlike any other disease of the 21st century, Covid-19 has changed the everyday life of people around the globe. The virus primarily targets the respiratory system of an individual, thus harming lungs and airways [1] and causing symptoms such as fever, dry cough and dyspnea [2]. Due to the sparse availability of drugs and vaccines, detection and isolation of infected individuals has proven to be a successful strategy to break the exponential growth of case numbers [3]. The most common form of screening are various types of medical (rapid) tests [4, 5]. However, a lack of availability, particularly during the onset of the pandemic, encouraged researchers to investigate other options to detect an infection with the severe acute respiratory syndrome coronavirus (SARS-CoV-2). Because Covid-19 symptoms strongly affect the respiratory tract, speech signals could serve as surrogates containing valuable information about the condition of an individual [6, 7]. Different studies focused on the analysis of breathing and cough sounds [8] or telephone-quality speech signals [9] to detect Covid-19 cases and reporting F1-scores or accuracy of over 90 %. There are, however, many pitfalls and challenges to overcome when designing a study to apply machine learning for automated Covid-19 screening [10].

A common flaw is the application of 0-1 loss function for all types of misclassifications. For a reliable screening solution, sensitivity is of greater importance compared to specificity, as the cost of overseeing a positive (infected) sample is considerably larger than falsely classifying a healthy individual as infected [11]. This cost inequality is not represented by a 0-1 loss function.

Another weakness potentially arises from the data itself. A successful detection of Covid-19 is challenging because initial symptoms are quite similar to those of the common flu. A successful classification between healthy individuals and Covid-19 patients does not support the assumption that a system is also able to differentiate between Covid-19 and other respiratory conditions.

This work is a contribution to this year's Computational Paralinguistics Covid-19 Speech sub-challenge (CSS) [12] and aims to provide a detailed analysis on how the condition affects the speech signals of infected patients on a phonetic level. After a brief description of both the provided Covid-19 dataset as well as three other auxiliary datasets, we introduce our deep acoustic model and describe how it was used to perform segmentation and feature extraction. After an outline of our classification system and general methodology, we provide a detailed analysis of the Covid-19 dataset, highlight important phonetic findings and discuss classification results. In the final conclusion, we summarize the most important findings.

## 2. Materials and Methods

### 2.1. Covid-19 Speech Dataset

The dataset comprised 893 speech recordings collected from 366 speakers in various languages. The language of the individual recordings as well as the mapping of the recordings to the speakers was not provided by the organizers. In every recording, a participant uttered the phrase ”I hope my data can help to manage the virus pandemic.” one or three times in their native language. Samples were distributed into training, development and test set. The training set contained 243 non-Covid (NC) and 72 Covid (COV) samples. This imbalance was not found in the development set with 153 NC and 142 COV. Ground truth of the test set was blinded. A more detailed description of the dataset can be found in the paper on the challenge [12].

### 2.2. Auxiliary datasets

Auxiliary datasets of three different languages were used to train a deep acoustic model. These corpora are completely unrelated to pathological speech, but they served as background information to learn what healthy speech sounds like. The largest

http://dx.doi.org/10.21437/Interspeech.2021-1488

of all three was a subset of the German Verbmobil corpus [13], containing 29 hours of dialogue speech recordings from 593 speakers (307 female, 286 male). We also incorporated the American English TIMIT corpus [14], comprising 5.3 hours of speech recordings collected from a total of 630 speakers (192 female, 438 male). Participants had to utter various phonetically rich sentences. The last dataset was the Mexican Spanish DIMEx100 corpus [15]. Each of the 100 speakers (51 female, 49 male) contributed recordings of 50 unique and 10 reference (identical for all speakers) phrases, totaling a bit more than 6 hours of speech.

### 2.3. Multilingual phonetic concept

The term *phoneme* is often used incorrectly [16]. In our setup, the ground truth of all three auxiliary datasets consisted of phoneme sequences of the respective language. Because phonemes are strictly bound to a certain language, we had to adapt the target space to work for multiple languages at the same time. We selected 35 target PHONE classes, such that each PHONE would be the union of elementary phones that could serve as a valid realization of the respective PHONE in any of the three languages.

### 2.4. Audio processing

Whenever necessary, an audio recording was first resampled to 16 kHz, followed by a root mean square normalization to a level of $-10$ dB and removal of DC offset. We then computed magnitude and phase spectrograms (2048 FFT points) over a window of 25 ms, shifting by 5 ms per frame. Both spectrograms were converted to logarithmic scale of base 10 and afterwards filtered with a triangular Mel-bank with 128 frequency bands to resemble the human perception of speech. The resulting dual-channel spectrograms served as input to our deep acoustic model.

### 2.5. Deep acoustic model

The deep acoustic model was comprised of two major components, a convolutional part for feature extraction and a recurrent part for sequential analysis. The neural network was trained in two separate steps to improve the final sequence prediction. Initially, the model operated as a framewise PHONE classifier. At every time step $t$, the network would distribute a probability mass over all PHONE targets.

After pre-training a framewise PHONE classifier, we slightly increased the number of hidden units of the recurrent cells and retrained the model. During this second training stage, we omitted any alignment information. The network now had to learn the alignment of PHONE sequences itself by using connectionist temporal classification (CTC) loss [17]. Through a beam search, we could ultimately predict the most probable sequence of output tokens, in our case PHONE classes, from the outputs of the acoustic model.

The convolutional feature extraction part was constructed with two major building blocks inspired by the Inception architectures presented for image classification [18]. The core idea was to apply multiple convolution kernels of varying sizes in parallel such that the network itself would learn which kernel worked best for what task. Within an Inception block, initial 1x1 convolutions perform a channel reduction to make the following convolution operations less parameter-intensive. Subsequent 1x1, 3x3 and 5x5 convolutions operated in parallel and (for 3x3 and 5x5) were implemented with separated kernels (3x3 = 3x1 $\times$ 1x3) to further reduce their complexity. A final 1x1 convolution projected the concatenated results from all three convolutions back to the original channel configuration. Before adding the result to the input of the Inception block to realize a residual connection [19], we applied activation scaling ($s = 0.3$) as proposed in [18] to stabilize training. The second important building block of the network was the reduction inception block. The major purpose of that component was to reduce the remaining number of frequency bins while leaving the time dimension unchanged. Reduction was performed with parallel max-pooling and strided convolution layers and their outputs were concatenated.

Table 1: *Outline of the CTC* PHONE *recognition model. Output size depended on the length of the sample (T). #c indicates number of channels. #x# denotes kernel size in temporal (first) and frequency (second) domain. [#, #] denotes the stride in the respective domain.*

| Output size | Layer |
|---|---|
| 2Tx64, 60 | 60c 1x4 Conv [1, 2] |
| Tx64, 120 | 120c 5x1 Conv [2, 1] |
| Tx32, 160 | 160c 1x4 Conv [1, 2] |
| Tx16, 200 | 200c 1x4 Conv [1, 2] |
| Tx16, 200 | 2 x Residual Inception Block<br>ch. reduced: 70 |
| Tx8, 340 | Reduction Inception Block<br>ch. reduced: 70 |
| Tx8, 340 | 2 x Residual Inception Block<br>ch. reduced: 120 |
| Tx4, 580 | Reduction Inception Block<br>ch. reduced: 120 |
| Tx4, 580 | 2 x Residual Inception Block<br>ch. reduced: 200 |
| Tx300 | Depthwise separable convolution |
| Tx480 | 2 x BiLSTM 240 hidden units |
| Tx35 | Linear projection to 35 targets PHONES |

A depthwise separable convolution [20] was used to project any remaining frequency bands and their channels after the last inception block down to 300 values per time step. The sequential analysis was realized through a stack of two bidirectional Long Short-Term Memories (BiLSTM) [21]. Each BiLSTM was configured with 240 hidden units per direction in the case of CTC, 200 units for pre-training to warm up the convolutional part. The final linear layer projected the output of the recurrent part to the 35 PHONE targets.

We used hard swish activation function [22] and performed batch normalization [23] before the non-linearity. After each normalization layer, we applied a dropout of 10 % to prevent overfitting. The model for framewise pre-training comprised around 6.6 million parameters, the final CTC model was slightly larger with roughly 7.2 million parameters. The architecture of the final CTC PHONE recognition model is shown in Table 1.

### 2.6. Methodology

For every sample of the CSS dataset, the PHONE sequence was predicted with the acoustic model. With every PHONE prediction, we also collected its posterior probability as well as a 480-dimensional feature vector. It represented the output hidden state of the last RNN layer at that time-step where the PHONE was predicted. The individual features required no further scaling, because hidden states of an LSTM are computed as the product of a logistic and a hyperbolic tangent function, thus ranging between [-1,1]. The technique of using a phonetic reference model trained on vast amounts of healthy speech for segmentation and feature extraction had already been successfully applied before [24].

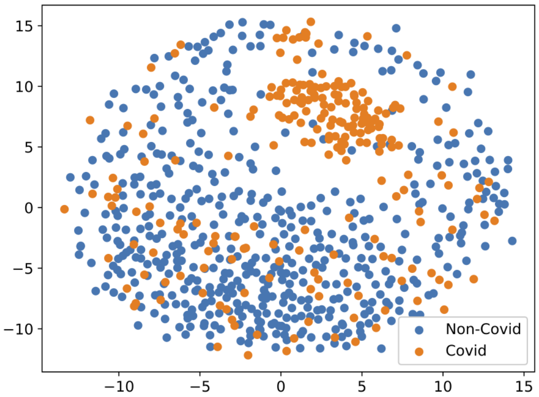

Figure 1: *Results of a t-SNE embedding of velar fricative [x] realizations for non-Covid and Covid samples.*

For an initial phonetic analysis of the challenge data, we grouped all feature vectors according to their associated PHONE prediction and applied t-Distributed Stochastic Neighbor Embedding (t-SNE) embedding [25] to identify if certain PHONES or PHONE groups showed distinct differences between NC and COV productions.

Because we had no information about (1) the number of different languages in the CSS dataset, (2) their relative frequency and (3) their distribution among training, development and test split, we computed a phonetic profile for every sample. This was done by computing the relative frequency of every PHONE in an individual sample. The result was a distinct profile, mainly dependent on the language in which the reference phrase was uttered. We computed t-SNE embedding on these profiles to identify clusters of phonetically similar samples. We then fitted a Gaussian Mixture Model (GMM) on the embedded space to group each sample into one of ten clusters. This strategy allowed us in the following steps to compare only samples of similar phonetic profiles, which corresponded to the same language.

With a better understanding of how Covid-19 affected the speech of individuals on a phonetic level, we designed our classification pipeline. We trained a linear Support Vector Machine (SVM) classifier for every PHONE and every phonetic cluster, using the 480-dimensional feature vectors of every PHONE prediction as input and performing 5-fold cross-validation (CV) to optimize regularization hyperparameter C ($10^x$ for $x \in [-4 .. 0]$). Note that the folds were created with respect to samples, not speakers, because the mapping of speakers to samples was unknown. Afterwards, the classifiers were used to compute the signed distance to the decision boundary for every feature vector. The sum of distances for every PHONE were stored in a 35-dimensional vector for each sample. The final prediction was then computed with a logistic regression classifier.

## 3. Results

The phonetic analysis of NC and COV samples showed distinct patterns in certain PHONE groups, while others remained unaffected. Our analysis of vowel t-SNE embeddings showed no identifiable clusters. The same was the case for other PHONES that were produced in the upper parts of the vocal tract, like the lips (e. g. [p] and [b]). We found that PHONES that have their place of articulation deeper in the vocal tract were more severely affected by Covid-19. An outstanding example of this was the velar fricative [x]. Figure 1 depicts the embedding result for NC and COV samples. We observed a dense cluster of COV samples that clearly separated from almost any NC sample. We found similar but less pronounced patterns for other velar PHONES, such as [g] and [k], as well as non-velar PHONES [r] or [h] for example.

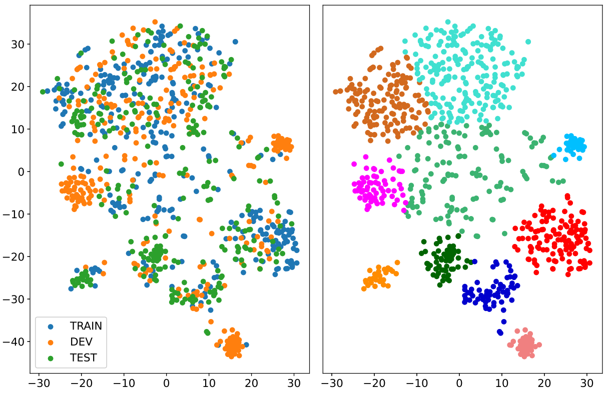

Figure 2: *Clustering of a t-SNE embedding of phonetic profiles for all samples. Left: Distribution of phonetic profiles among training, development and test split. Right: Same embedding, clustered with a Gaussian Mixture Model into 10 different components.*

To better understand how languages were distributed in the CSS dataset, we created t-SNE embeddings of the phonetic profile of every sample as shown in Figure 2. With a GMM, individual clusters could be separated from each other. Looking at the distribution of phonetic profiles among train, development and test split, we found that there were multiple clusters which were dominated by only one of the three splits. Particularly the three dense orange clusters of samples from the development set caught our attention. Similarly, we found regions where test samples were stronger represented than the other splits. Certain clusters from the GMM were found to have very few representative samples from the training or development split. In general, we found that samples within certain clusters of phonetic profiles were not evenly distributed among the three data splits.

Our final classifier achieved an unweighted average recall (UAR) on the development set of 73.8 % (baseline: 57.9 %)[1]. On the test set, we achieved 64.2 % (baseline: 72.1 %). Figure 3 shows the confusion matrix of our classification results on the development set and allows a further interpretation of the results. With a false negative rate (FNR) of 29.6 %, our solution still missed a good amount of COV patients. On the other side, the FNR of the baseline method was profoundly larger with 63.4 %. While our system missed almost 30 % of COV cases, it falsely predicted 22.9 % of NC as positive.

The receiver operating characteristic (ROC) curve is illustrated in Figure 4. Area under the curve was 0.752 on the development set. For further interpretation of the results, we decided to mark a few notable points on the curve. First, the equal error rate of our system was 27.5 %. After shifting the decision boundary to 10 % FNR, the true negative rate (TNR) was 22.2 %. For an even lower FNR of 1 %, the remaining TNR was 2 %. These numbers were interesting because screening solutions should be very sensitive to ensure that only a very small fraction of infected individuals would be misclassified. This comes at the cost of specificity, thus classifying more healthy samples incorrectly. Under the assumption that costs for a false

[1] We refer to the openSMILE system as baseline, since it performed best on test and no confusion table for the End2You system was provided in [12]

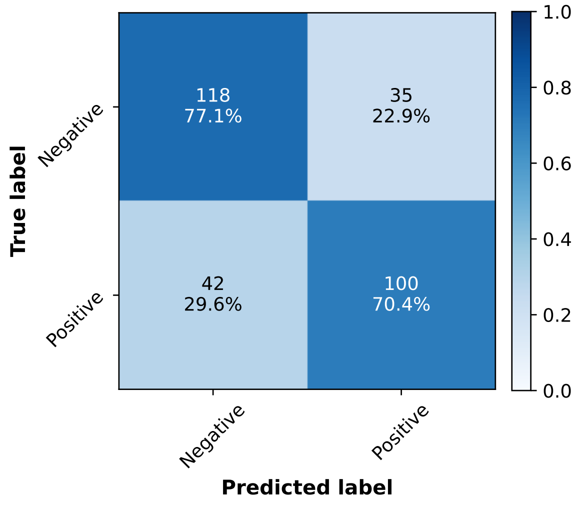

Figure 3: *Confusion matrix of Covid-19 classification results on the development set.*

negative prediction are 25 times larger than those of a false positive, and 5 % of tested individuals actually suffer from Covid-19, the green curve in Figure 4 highlights how the total relative costs changed while shifting the decision boundary of our system. In this scenario, the optimal point of our classifier almost halved the cost of the baseline system.

## 4. Discussion

Results of the phonetic analysis of COV speech compared to NC samples showed very distinct deviations for a small group of PHONES. Most of them had their place of articulation deep inside the vocal tract, which could be an indicator of, for example, an infection of the respiratory system. Our deep acoustic model trained only with healthy speech was able to recognize pathological changes in PHONE realizations. However, because Covid-19 is not the only condition that affects the respiratory tract, it is impossible to exclusively attribute the observed patterns to that disease. They could also be the result of an influenza, or a simple cold.

The clusters of phonetic profiles observed for the CSS dataset led us to the assumption that certain languages corresponding to said profiles were unevenly distributed among the splits. This might also explain the large deviation of UAR results on development and test sets. The absolute difference for our system was 9.6 %, the one of the baseline was even larger with 14.2 %. These differences could well be caused by an uneven distribution of languages.

The confusion matrix reveals that our classifier was able to detect around 70 % of all COV samples correctly. Compared to the baseline approach, this was a major improvement. However, the cost analysis on the ROC curve showed that for a Covid-19 screening scenario, the most important factor is sensitivity. The costs $R$ for classification decisions can be generalized as

$$R_{\mathrm{HH}} \leq R_{\mathrm{CC}} < R_{\mathrm{HC}} < R_{\mathrm{CH}} \qquad (1)$$

where the first subscript indicates the ground truth and the second indicates the prediction (H = Healthy, C = Covid). A classifier trained with the 0-1 loss for both $R_{\mathrm{CH}}$ and $R_{\mathrm{HC}}$ does not account for $R_{\mathrm{CH}} > R_{\mathrm{HC}}$, therefore it is indispensable to incorporate knowledge about the cost structure already in the optimization. This is also the reason why the costs at FNR values of 10 % and 1 % were rather high again. The system was simply not optimized to perform well at high sensitivity levels.

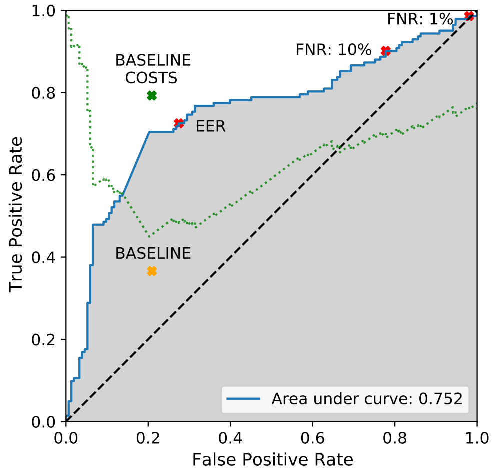

Figure 4: *ROC curve for the classification results on the development set. BASELINE: Performance of the baseline method. EER: Point of equal error rate. FNR: 10 / 1%: Points on ROC curve with 10 / 1% false negative rate, respectively. Green curve shows relative hypothetical costs of misclassification. BASELINE COSTS shows relative costs of the baseline system.*

## 5. Conclusions

Whilst we found pronounced phonetic patterns in the CSS dataset that could help to better understand how Covid-19 affected an individual's speech, we don't want to attribute any such patterns exclusively to Covid-19. Without additional meta-information about language, age, gender, other respiratory conditions or risk factors (such as smoking), an interpretation of our findings with respect to Covid-19 is impossible. Hence, we conclude that any such classifier should predict whether further testing is required rather than making a hard decision whether an individual suffers from Covid-19 or not. This is also desirable to better account for asymptomatic Covid-19 cases.

Understanding how the disease affected an individuals articulation of certain PHONE groups could help to improve first-level screening solutions in general. We recommend that individuals are asked to produce sentences with many velar phonemes of their native language, because these encode valuable information about the condition of the respiratory tract.

We think that in future challenges, participants should be given additional meta-information, not only to improve classification results, but more importantly, to allow for a better interpretation of findings. However, we assume that this was simply not possible for the CSS dataset due to data privacy obligations.

## 6. Acknowledgements

This study has been funded by the Federal Ministry of Education and Research of Germany in the ASA-KI project, grant No. 16SV8469. The authors also acknowledge to the Training Network on Automatic Processing of PAthological Speech (TAPAS), grant agreement No. 766287 funded by the Horizon 2020 program of the European Commission. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS. This work was also financed by CODI at UdeA grant No. PRG2017-15530.

# 7. References

[1] X. Cao, "COVID-19: immunopathology and its implications for therapy," *Nature Reviews Immunology*, vol. 20, no. 5, pp. 269–270, may 2020.

[2] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," *Journal of Autoimmunity*, vol. 109, p. 102433, may 2020.

[3] T. A. Treibel, C. Manisty, M. Burton, Á. McKnight, J. Lambourne, J. B. Augusto, X. Couto-Parada, T. Cutino-Moguel, M. Noursadeghi, and J. C. Moon, "COVID-19: PCR screening of asymptomatic health-care workers at London hospital," *The Lancet*, vol. 395, no. 10237, pp. 1608–1610, may 2020.

[4] R. Weissleder, H. Lee, J. Ko, and M. J. Pittet, "COVID-19 diagnostics in context," Sci. Transl. Med, Tech. Rep., 2020. [Online]. Available: https://stm.sciencemag.org/content/12/546/eabc1931/tab-pdf

[5] R. W. Peeling, P. L. Olliaro, D. I. Boeras, and N. Fongwen, "Scaling up COVID-19 rapid antigen tests: promises and challenges," *The Lancet Infectious Diseases*, 2021. [Online]. Available: www.thelancet.com/infectionPublishedonline

[6] G. Deshpande and B. W. Schuller, "An Overview on Audio, Signal, Speech, & Language Processing for COVID-19," arXiv:2005.08579 [cs.CY], Tech. Rep., 2020.

[7] T. F. Quatieri, T. Talkar, J. S. Palmer, and S. Member, "Technology Letter A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, 2020.

[8] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 Detection System using Recurrent Neural Networks," in *Proceedings of the 2020 IEEE International Conference on Communications, Computing, Cybersecurity, and Informatics, CCCI 2020*, 2020.

[9] K. V. S. Ritwik, S. B. Kalluri, and D. Vijayasenan, "COVID-19 patient detection from telephone quality speech data," 2020. [Online]. Available: https://github.com/shareefbabu/covid

[10] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, and C.-B. Schönlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, 2020. [Online]. Available: https://doi.org/10.1038/s42256-021-00307-0 http://arxiv.org/abs/2008.06388

[11] J. Wilson and G. Jungner, "Principles and practice of screening," *WHO: Geneva*, vol. 69, no. 5, p. 1085, nov 1968. [Online]. Available: http://annals.org/article.aspx?doi=10.7326/0003-4819-69-5-1085_2 http://libdoc.who.int/php/WHO_PHP_34_rus.pdf

[12] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, M. R. Leon J. J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *Proceedings INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*. Brno, Czechia: ISCA, September 2021, to appear.

[13] W. Wahlster, *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993. [Online]. Available: papers://e7d065ae-9998-4287-8af0-c9fa85af8e96/Paper/p44370

[15] L. A. Pineda, L. V. Pineda, J. Cuétara, H. Castellanos, and I. López, "DIMEx100: A new phonetic and speech corpus for Mexican Spanish," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 3315. Springer, 2004, pp. 974–983.

[16] R. K. Moore and L. Skidmore, "On the use/misuse of the term 'phoneme'," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, no. October, pp. 2340–2344, 2019.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ACM International Conference Proceeding Series*, vol. 148, 2006, pp. 369–376.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, pp. 4278–4284.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016, pp. 770–778.

[20] L. Sifre, "PhD thesis Rigid-Motion Scattering For Image Classification," *Ph. D. thesis*, 2014.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *arXiv*, 2019, pp. 1314–1324.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, 2015.

[24] P. Klumpp, T. Arias-Vergara, J. C. Vásquez-Correa, P. A. Pérez-Toro, F. Hönig, E. Nöth, and J. R. Orozco-Arroyave, "Surgical mask detection with deep recurrent phonetic models," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-Octob, pp. 2057–2061, 2020.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.