Speaking corona? Human and machine recognition of COVID-19 from voice

Pascal Hecker, Florian B. Pokorny, Katrin D. Bartl-Pokorny, Uwe Reichel, Zhao Ren, Simone Hantke, Florian Eyben, Dagmar M. Schuller, Bert Arnrich, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Hecker, Pascal, Florian B. Pokorny, Katrin D. Bartl-Pokorny, Uwe Reichel, Zhao Ren, Simone Hantke, Florian Eyben, Dagmar M. Schuller, Bert Arnrich, and Björn W. Schuller. 2021. "Speaking corona? Human and machine recognition of COVID-19 from voice." In *Interspeech 2021, Brno, Czechia, 30 August - 3 September 2021*, edited by Hynek Heřmanský, Honza Černocký, Lukáš Burget, Lori Lamel, Odette Scharenborg, and Petr Motlicek, 1029–33. Baixas: International Speech Communication Association (ISCA). https://doi.org/10.21437/interspeech.2021-1771.

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

Speaking Corona? Human and Machine Recognition of COVID-19 from Voice

Pascal Hecker[1,2], *Florian B. Pokorny*[3,4], *Katrin D. Bartl-Pokorny*[3,4], *Uwe Reichel*[1], *Zhao Ren*[3], *Simone Hantke*[1], *Florian Eyben*[1], *Dagmar M. Schuller*[1], *Bert Arnrich*[2], *Björn W. Schuller*[1,3,5]

[1]audEERING GmbH, Gilching, Germany
[2]Digital Health – Connected Healthcare, Hasso Plattner Institute, University of Potsdam, Germany
[3]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
[4]Division of Phoniatrics, Medical University of Graz, Austria
[5]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

phecker@audeering.com

Abstract

With the COVID-19 pandemic, several research teams have reported successful advances in automated recognition of COVID-19 by voice. Resulting voice-based screening tools for COVID-19 could support large-scale testing efforts. While capabilities of machines on this task are progressing, we approach the so far unexplored aspect whether human raters can distinguish COVID-19 positive and negative tested speakers from voice samples, and compare their performance to a machine learning baseline. To account for the challenging symptom similarity between COVID-19 and other respiratory diseases, we use a carefully balanced dataset of voice samples, in which COVID-19 positive and negative tested speakers are matched by their symptoms alongside COVID-19 negative speakers without symptoms. Both human raters and the machine struggle to reliably identify COVID-19 positive speakers in our dataset. These results indicate that particular attention should be paid to the distribution of symptoms across all speakers of a dataset when assessing the capabilities of existing systems. The identification of acoustic aspects of COVID-19-related symptom manifestations might be the key for a reliable voice-based COVID-19 detection in the future by both trained human raters and machine learning models.

Index Terms: auditory disease perception, automatic disease recognition, computational paralinguistics, COVID-19, voice

1. Introduction

As of April 2021, over 125 million confirmed cases with over 2.8 million deaths due to the coronavirus disease 2019 (COVID-19) were reported by the WHO [1]. COVID-19 is caused by an infection with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which first appeared in December 2019 [2] and since then led to a pandemic.

The severity of COVID-19 is heterogeneous. While numerous patients with COVID-19 have no symptoms and are diagnosed by chance or not at all, most patients have mild-to-moderate flu-like symptoms, and others need to be hospitalised or even die from the disease. The most commonly reported symptoms of COVID-19 are respiratory and ear-nose-throat symptoms such as cough, shortness of breath, sore throat, and headache, systemic symptoms such as fever, weakness, and muscle pain, as well as loss of taste and/or smell [3]. A great proportion of the characteristic symptoms of COVID-19 affect anatomical correlates of speech production, namely the lungs, the vocal folds, the vocal tract, and/or the nasal tract [4]. A number of studies investigated voice changes in other diseases affecting those correlates. Researchers reported a range of acoustic parameters differing between healthy controls and patients with asthma [5, 6], vocal fold atypicalities [7, 8], or cleft lip and palate [9]; among these are jitter, shimmer, harmonics-to-noise-ratio (HNR), fundamental frequency (F0), first to third vowel formants, and maximum phonation time (MPT).

The existing literature on acoustic parameters in patients with COVID-19 is still sparse. A study on Persian speakers found significant differences in voice samples of the sustained vowel /a:/ between patients with COVID-19 and healthy controls in F0 standard deviation, jitter, shimmer, HNR, difference between the first two harmonic amplitudes (H1–H2), MPT, and cepstral peak prominence [10]. To the best of our knowledge, Bartl-Pokorny et al. [11] were the first to investigate acoustic parameters in the sustained vowels /i:/, /e:/, /u:/, /o:/, and /a:/ produced by German speakers with and without COVID-19. Across all vowels, the study revealed differences in the mean voiced segment length and the number of voiced segments per second, reflecting discontinuities in the pulmonic airstream during phonation in patients with a COVID-19 infection.

Related work. Motivated by the manifestation of various diseases in the human voice, machine learning approaches have been deployed in recent years and shown effectiveness in the automatic voice-based recognition of medical conditions ranging from respiratory diseases, e. g., cold and flu [12], via psychiatric disorders, such as depression [13], to developmental disorders, such as autism spectrum disorder [14] or Rett syndrome [15], and neurodegenerative diseases, e. g., Alzheimer's disease [16]. Thus, well-founded and not surprisingly, there is also a growing body of research on the automated detection of COVID-19 from voice. Several research groups in the domain of computational paralinguistics have focused their efforts on collecting large 'crowdsourced' datasets (e. g., [17, 18, 19, 20]) using them to create machine learning models (e. g., [21, 22, 23, 24, 25, 26]). Since voice is a readily available modality and the collection of voice data is non-invasive, voice-based models for COVID-19 detection could serve as valuable screening instruments [27]. Such models could be deployed online and enable a large share of the population access to inexpensive and frequent testing.

Another, yet unexplored, aspect in context of COVID-19 is listening perception of human raters to recognise the disease. It seems intuitive that humans may be able to recognise respiratory diseases in voice as humans have substantial experience in listening to the voice of people with a cold. Empirical evidence that humans are able to recognise the presence

http://dx.doi.org/10.21437/Interspeech.2021-1771

or absence of a disease comes, e. g., from a study focusing on Rett syndrome, a rare genetic disorder associated with profound speech-language deficiencies [28]. The researchers found that professional and naive human raters were capable to differentiate between (pre-linguistic) vocalisations produced by infants with Rett syndrome and vocalisations produced by typically developing infants [28].

Contributions of this work. To the best of our knowledge, this is the very first study to assess if naive human raters can be trained to distinguish between voice samples from speakers who were tested positive or negative for COVID-19. Furthermore, we aim to compare the raters' performance to the performance of an automatic recognition approach on the same dataset.

One prominent challenge when distinguishing COVID-19 by voice is that several symptoms, such as coughing and sneezing, overlap in COVID-19 and other respiratory diseases. Therefore, in this study, we compose a carefully balanced dataset with three groups of equal sizes (all speakers tested positive (+) or negative (-) for COVID-19):

1. COVID-19+: symptomatic and asymptomatic
2. COVID-19-: matched symptoms of COVID-19+
3. COVID-19-: asymptomatic

2. Materials and Methods

2.1. Dataset

The dataset in this study, used for both the listening perception and the automatic recognition task, is composed as a subset of two existing datasets with COVID-19 positive and negative speakers from the audEERING GmbH and the University of Augsburg, Germany. The composed dataset includes only speakers from German-speaking countries (Germany, Austria, Switzerland) to minimise language-dependent influences. We select the following overlapping six prompts of both datasets: coughing deliberately 1–3 times while breathing out ('coughing'), reading aloud the first two sentences of the standard phonetic text passage "The North Wind and the Sun" in German ('read text'), as well as sustained phonations of the vowels /a:/, /e:/, /i:/, and /u:/ for 5 seconds, each. Detailed demographics of the combined and separate datasets are presented in Table 1. Of the eleven symptomatic COVID-19 positive subjects, four have respiratory and systemic symptoms, six only respiratory, and one only systemic symptoms. Those symptom patterns are closely matched in the symptomatic COVID-19 negative group.

audEERING developed a data collection platform called AI SoundLab[1], which enables the remote collection of voice samples through a web-app on any device with internet access. Users can register with their e-mail address or use pseudonymised study-tokens, which enables clear association of a unique user and their recordings from multiple sessions. Users can choose to disclose general metadata with age, gender, and mother tongue being mandatory and for every recording session, the test status for COVID-19 as well as various systemic (e. g., fever) and respiratory symptoms (e. g., coughing) are assessed.

The dataset from the University of Augsburg consists of comparable speech data of COVID-19 positive and COVID-19 negative speakers. The speakers were asked to record their voice with their smartphones and to transfer the recordings via the secure file-sharing service of the University of Augsburg. All speakers provided a copy of the result of their COVID-19 test done within the last three days before their study participation. They completed a short questionnaire including information about potential symptoms and health issues. The data collection was approved by the ethics representative of the University of Augsburg. All speakers gave their written informed consent to participate in the study.

Table 1: *Number of speakers (# Speak.), gender distribution, as well as mean ± standard deviation of speaker age rounded to full years (y) per sub-dataset, condition, class, and in total. neg = COVID-19 negative, pos = COVID-19 positive, yes = with symptoms, no = without symptoms, ♀ = female, ♂ = male*

Condition	Sub-dataset	# Speak. (♀/♂)	Age [y]
neg+no	audEERING	7 (2/5)	48 ± 11
	Uni Augsburg	6 (1/5)	42 ± 17
	Σ	13 (3/10)	45 ± 13
neg+yes	audEERING	7 (2/5)	42 ± 18
	Uni Augsburg	6 (2/4)	47 ± 21
	Σ	13 (4/9)	45 ± 19
Σ_{neg}		26 (7/19)	45 ± 16
pos+no	audEERING	2 (0/2)	41 ± 8
	Uni Augsburg	-	-
	Σ	2 (0/2)	41 ± 8
pos+yes	audEERING	5 (2/3)	37 ± 10
	Uni Augsburg	6 (1/5)	38 ± 19
	Σ	11 (3/8)	38 ± 15
Σ_{pos}		13 (3/10)	38 ± 14
$\Sigma_{\text{neg}\cup\text{pos}}$		39 (10/29)	43 ± 16

To provide enough data for human raters to learn the difference between COVID-19 negative and positive tested speakers, we partition the combined dataset into three folds for cross validation. 2/3 of the data is assigned to the train and 1/3 to the test partition. 13 COVID-19 positive speakers of both datasets are matched according to their symptoms, age, and gender with 13 COVID-19 negative ones and supplemented with 13 COVID-19 negative speakers without symptoms. Folds are created so that speakers are kept separate in the train and test partitions. Due to the odd number of speakers, the first third of speakers contains five speakers of each condition, while the other two thirds contain four.

2.2. Human raters

We conduct a binary classification study for the three data partitions in order to evaluate the capability of human raters to distinguish COVID-19 positive and COVID-19 negative speakers by only listening to voice samples. The listening study is performed using a browser-based interface provided through the gamified crowdsourcing platform *iHEARu-PLAY* [29]. Within each fold, raters listen to all train samples (2/3 of the speakers) for one prompt, while being informed about the respective ground truth labels, and rate all test samples (remaining 1/3 of the speakers) of that prompt, before advancing to the next prompt. For each test sample, raters state the confidence of their rating. In total, 15 native German-speaking raters take part (5 per fold). Their age is between 21 and 32 years, with a mean age of 25.3 years and a standard deviation of 2.4 years. The raters are employees and students from audEERING and

[1] https://aisoundlab.audeering.com/

the University of Augsburg, who have no insight into the research and study procedure. In order not to overburden the raters, the sessions are designed to last no more than one hour; raters are also allowed to take a break whenever needed. The listening study itself is a forced-choice task: a binary classification task for COVID-19 positive vs COVID-19 negative as well as a 3-class task for the confidence of their rating ('unsure', 'somewhat sure', and 'sure').

Scale merging. In order to obtain a 5-point Likert scale for later analyses of COVID-19 perception, we merge the binary answers of the raters with their judgement of how sure they are in their decision as follows: (negative+sure) is mapped to 1, (negative+somewhat sure) to 2, (negative+unsure), as well as (positive+unsure) to 3, (positive+somewhat sure) to 4, and (positive+sure) is mapped to 5.

Statistical analyses. Separately for each prompt, we first test by means of two-tailed binomial tests whether human raters' performance is significantly better than chance level. For this assessment, performance is measured in terms of the proportion of correct answers and compared against the chance level 0.5 (raters are not aware of the class distribution). We repeat this analysis for aggregated binary judgements, which are derived the following way: one aggregated judgement is obtained per rater and speaker by calculating the proportion of COVID-19 positive judgements of the rater over all prompts of the speaker. If this proportion is greater than 0.5, the aggregated judgement is set to *positive*, else to *negative*.

Further, we assess for each prompt whether COVID-19 positive speakers and/or symptomatic speakers are scored higher on the 5-point Likert scale than COVID-19 negative and/or asymptomatic speakers. These questions are addressed by means of two-tailed Mann-Whitney U tests.

Finally, for each prompt we record the proportion of Likert-level 3 ('unsure') answers as a proxy for perceived task difficulty, and we collect for each prompt the unweighted average recall (UAR) value over all raters to compare it with the machine learning results.

2.3. Machine Learning

We use the baseline feature set of the annual COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) [30] to generate $6,373$ features for each audio sample by means of the open-source feature extraction toolkit openSMILE [31]. Furthermore, we employ the COMPARE baseline machine learning model, i. e., linear-kernel support vector machines (SVMs), to classify the audio samples into *COVID-19 negative* vs *COVID-19 positive*.

Based on the same three data partitions already used for the listening study, we train and evaluate the SVM models separately for each prompt following a three-fold cross-validation scheme. The predictions for the test set of each fold are connected to form predictions for all samples of the respective prompt. The SVM complexity parameter C is optimised $\in [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$ in order to obtain the best UAR across all folds. To mitigate the issue of class imbalances, SVM models are trained with class weights being inversely proportional to the class frequencies. In addition to the prompt-wise performance evaluation scenarios, we combine the predictions for all samples of a speaker to derive a final decision for him or her as a whole based on majority voting (in at least four of six prompts, the speaker has to be classified as *positive* for being assigned to the *COVID-19 positive* class as a whole).

3. Results

Human performance. Only for sustained /i:/ and for the aggregated judgements, human raters perform weakly significantly better than chance level (two-tailed binomial tests, $p = 0.09$ and 0.06, respectively). Accuracies amount in both cases to $57\,\%$. Judgement performances for all other prompts are close to $50\,\%$ and thus do not differ from chance level.

Five-point Likert scale. Only for the read text, we find significantly higher scores for COVID-19 positive and for asymptomatic speakers (two-tailed Mann-Whitney U tests; COVID-test: $U = 3131.0, p = 0.04$; symptoms: $U = 3332.5$, $p = 0.04$). Effect sizes however are small, Cohen-d values amount to 0.28 for COVID-test and 0.23 for symptoms.

Perceived task difficulty. Perceived task difficulty is measured in terms of the overall proportion of level 3 ('unsure') scores. Perceived difficulty is overall high and amounts to 0.54 for coughing, 0.68 for read text and for sustained /e:/, 0.71 for sustained /i:/, 0.72 for sustained /a:/, and 0.76 for sustained /u:/. Thus, human raters are most confident when judging coughing and least confident when judging sustained /u:/.

Machine judgement. For most prompts, the SVM models achieve an UAR around or slightly above chance level, i. e., 0.5. Aggregating the predictions per speaker across all prompts leads to an UAR of 0.58. The best machine judgement, i. e., an UAR of 0.63, is obtained for the read text.

Comparison of human and machine judgements. For the purpose of performance comparison, we present one confusion matrix each for human and machine judgements for the best performing prompts, which are /i:/ and read text, respectively. Further, we compare the UAR scores separately for all prompts. Figure 1 shows the confusion matrix of human COVID-19 judgements for sustained /i:/. Overall, there is a bias towards COVID-19 negative. A systematic influence of symptoms towards COVID-19 positive judgements cannot be observed.

Figure 2 shows the confusion matrix of machine COVID-19 judgements for read text. Recall turns out to be high for the negative, but low for the positive class. Again, symptoms do not show a systematic influence on classification.

The UAR values for all prompts and both humans and machine are presented in Figure 3. Scores are low for both humans and machine, with a slight advantage for the latter.

4. Discussion

In our study, naive human raters and machine learning models struggle to detect COVID-19 from voice better than chance level. For human raters, we only find slight indications of a perceptual sensitivity to COVID-19 for vowel /i:/ and read text. For machine learning models, the best performance is achieved for read text. Our findings suggest that distinguishing COVID-19 positive and negative speakers with similar respiratory symptoms is a challenge both for human raters and machine learning methods. Here, we prominently address the symptom similarity aspect by matching every COVID-19 positive speaker with a COVID-19 negative speaker with similar respiratory or systemic symptoms. Most crowdsourcing efforts do not consider potential recurring users and that multiple samples of one and the same user could be partitioned both into the train and test sets. To account for this issue, our data collection approach ensures clear association of all recording sessions to unique users. This procedure results in a rather small, but to date uniquely carefully balanced dataset.

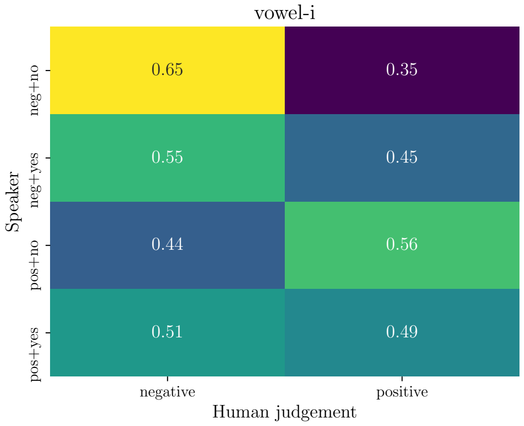

Figure 1: *Confusion matrix for human judgements for sustained /i:/. COVID-19 positive (pos) and negative (neg) ground truth is further split into symptomatic* (yes) *and asymptomatic* (no)*. Absolute counts (row-wise): 38, 27, 33, 32, 6, 4, 29, 26*

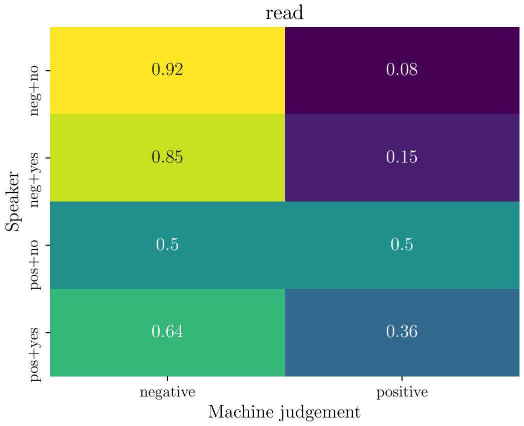

Figure 2: *Confusion matrix for machine judgements for read text. COVID-19 positive (pos) and negative (neg) ground truth is further split into symptomatic* (yes) *and asymptomatic* (no)*. Absolute counts (row-wise): 12, 1, 11, 2, 1, 1, 7, 4.*

For automated recognition of COVID-19, our results are in line with findings of previous studies. Coppock et al. [21] reported lower UAR and area under the receiver operating characteristic curve (AUC) scores when distinguishing COVID-19 positive speakers with a cough from COVID-19 negative speakers with a cough. Stasak et al. [25] similarly obtained chance-level results when regarding COVID-19 positive and negative speakers with moderate COVID-19-like symptoms. Conversely, Han et al. [22] reported a high rate of asymptomatic patients getting misclassified as healthy speakers.

With regard to human perception of voice samples, it is worth emphasising that the findings of our study clearly cannot be extrapolated to the performance of trained experts like medical doctors or speech-language therapists, who may perform better on this task.

When aggregating a rater's judgements over a speaker, performance slightly improves. This might indicate that the raters need a larger amount of speech material from a speaker in order to succeed. Similarly, sufficient amounts of data are also required for machine learning methods to reach robustness and to generalise well. In this regard, it is beneficial that more and more publicly available COVID-19 voice datasets are provided to the community (such as [19, 20]).

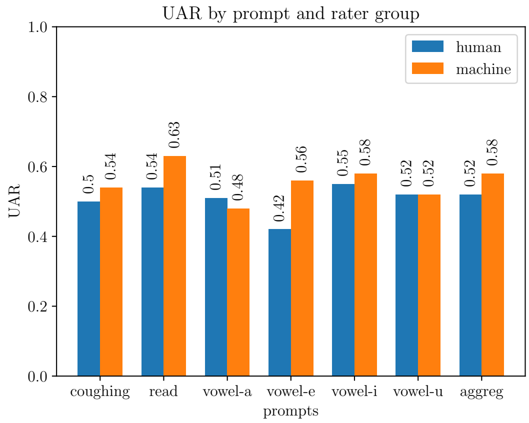

Figure 3: *Unweighted average recall (UAR) comparison between human and machine judgements for each prompt and prompt aggregation ('aggreg').*

The issue of class imbalances was approached in [23, 24] by using data augmentation techniques. It is essential to report information on class distributions and performance on those more fine-grained targets to get a fundamental understanding of proposed systems and their potential shortcomings, as interpretation is otherwise challenging (e. g., [26]).

Finally, fusion of modalities showed promising results to increase the robustness of automated COVID-19 detection. Han et al. [22] gained their best results when regarding self-reported symptoms of the speakers in addition to voice-based analyses.

5. Conclusion and Outlook

To the best of our knowledge, this study is the first of its kind to evaluate how well naive human raters can distinguish voice samples from COVID-19 positive and negative speakers, compared to a machine learning baseline. It represents an initial exploration based on a small, but carefully balanced dataset, pointing out difficulties for both human raters and the machine in reliably recognising COVID-19 due to the potential presence of symptoms in both COVID-19 positive and negative speakers. Future work shall expand our approach by additionally including expert human raters, such as healthcare professionals, and by using larger amounts of data, allowing for an application of more sophisticated machine learning methodology.

6. Acknowledgements

We thank all speakers, who donated their voice samples to our dataset and the *iHEARu-PLAY* users for rating those samples. Furthermore, we thank Anastasia Semertzidou for assistance in data preparation, Hesam Sagha and his team for the development of AI SoundLab, and Lin Zhou for proofreading.

7. References

[1] "WHO Coronavirus (COVID-19) Dashboard," Geneva: World Health Organization, accessed 2021-04-02. [Online]. Available: https://covid19.who.int/

[2] "Novel coronavirus – China," Geneva: World Health Organization, disease outbreak news: Update. Accessed 2021-04-02. [Online]. Available: http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/

[3] H. Esakandari, M. Nabi-Afjadi, J. Fakkari-Afjadi, N. Farahmandian, S.-M. Miresmaeili, and E. Bahreini, "A comprehensive review of COVID-19 characteristics," *Biological Procedures Online*, vol. 22, pp. 1–10, 2020.

[4] Z. Zhang, "Mechanics of human voice production and control," *Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2614–2635, 2016.

[5] M. Dogan, E. Eryuksel, I. Kocak, T. Celikel, and M. A. Sehitoglu, "Subjective and objective evaluation of voice quality in patients with asthma," *Journal of Voice*, vol. 21, no. 2, pp. 224–230, 2007.

[6] B. T. Balamurali, H. I. Hee, O. H. Teoh, K. P. Lee, S. Kapoor, D. Herremans, and J.-M. Chen, "Asthmatic versus healthy child classification based on cough and vocalised /a:/ sounds," *Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. EL253–EL259, 2020.

[7] M. Petrović-Lazić, S. Babac, M. Vuković, R. Kosanović, and Z. Ivanković, "Acoustic voice analysis of patients with vocal fold polyp," *Journal of Voice*, vol. 25, no. 1, pp. 94–97, 2011.

[8] L. M. Jesus, J. Martinez, A. Hall, and A. Ferreira, "Acoustic correlates of compensatory adjustments to the glottic and supraglottic structures in patients with unilateral vocal fold paralysis," *BioMed Research International*, vol. 2015, pp. 1–9, 2015.

[9] M. Segura-Hernández, V. M. Valadez-Jiménez, P. A. Ysunza, A. P. Sánchez-Valerio, E. Arch-Tirado, A. L. Lino-González, and X. Hernández-López, "Acoustic analysis of voice in children with cleft lip and palate following vocal rehabilitation. Preliminary report," *International Journal of Pediatric Otorhinolaryngology*, vol. 126, p. 109618, 2019.

[10] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh, "Voice quality evaluation in patients with COVID-19: An acoustic analysis," *Journal of Voice*, 2020.

[11] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. W. Schuller, "The voice of COVID-19: Acoustic correlates of infection in sustained vowels," *Journal of the Acoustical Society of America*, 2021 (in press).

[12] M. Albes, Z. Ren, B. Schuller, and N. Cummins, "Squeeze for sneeze: Compact neural networks for cold and flu recognition," in *Proc. Interspeech*. Shanghai, China: ISCA, October 2020, pp. 4546–4550.

[13] Z. Zhao, Q. Li, N. Cummins, B. Liu, H. Wang, J. Tao, and B. W. Schuller, "Hybrid network feature extraction for depression assessment from speech," in *Proc. Interspeech*. Shanghai, China: ISCA, October 2020, pp. 4956–4960.

[14] F. B. Pokorny, B. W. Schuller, P. B. Marschik, R. Brueckner, P. Nyström, N. Cummins, S. Bölte, C. Einspieler, and T. Falck-Ytter, "Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach," in *Proc. Interspeech*. Stockholm, Sweden: ISCA, August 2017, pp. 309–313.

[15] F. B. Pokorny, P. B. Marschik, C. Einspieler, and B. W. Schuller, "Does she speak RTT? Towards an earlier identification of Rett syndrome through intelligent pre-linguistic vocalisation analysis," in *Proc. Interspeech*. San Francisco, CA, USA: ISCA, September 2016, pp. 1953–1957.

[16] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, B. Daniel, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *Proc. Interspeech*. Shanghai, China: ISCA, October 2020, pp. 2182–2186.

[17] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. San Diego, CA, USA: ACM, Aug. 2020, pp. 3474–3484.

[18] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[19] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *arXiv preprint arXiv:2009.11644*, Sep. 2020.

[20] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara – a database of breathing, cough, and voice sounds for COVID-19 diagnosis," in *Proc. Interspeech*. Shanghai, China: ISCA, October 2020, pp. 4811–4815.

[21] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. W. Schuller, "End-2-end COVID-19 detection from breath & cough audio," *arXiv preprint arXiv:2102.08359*, 2021.

[22] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," *arXiv preprint arXiv:2102.05225*, 2021.

[23] M. E. Chowdhury, N. Ibtehaz, T. Rahman, Y. M. S. Mekki, Y. Qibalwey, S. Mahmud, M. Ezeddin, S. Zughaier, and S. A. S. Al-Maadeed, "QUCoughScope: An artificially intelligent mobile application to detect asymptomatic COVID-19 patients using cough and breathing sounds," *arXiv preprint arXiv:2103.12063*, 2021.

[24] A. Fakhry, X. Jiang, J. Xiao, G. Chaudhari, A. Han, and A. Khanzada, "Virufy: A multi-branch deep learning network for automated detection of COVID-19," *arXiv preprint arXiv:2103.01806*, 2021.

[25] B. Stasak, Z. Huang, S. Razavi, D. Joachim, and J. Epps, "Automatic detection of COVID-19 based on short-duration acoustic smartphone speech analysis," *Journal of Healthcare Informatics Research*, pp. 1–17, 2021.

[26] K. K. Lella and P. Alphonse, "Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice," *AIMS Public Health*, vol. 8, no. 2, pp. 240–264, 2021.

[27] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis," *arXiv preprint arXiv:2003.11117*, March 2020.

[28] P. B. Marschik, C. Einspieler, and J. Sigafoos, "Contributing to the early detection of Rett syndrome: The potential role of auditory gestalt perception," *Research in Developmental Disabilities*, vol. 33, no. 2, pp. 461–466, March 2012.

[29] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. Int. Workshop on Automatic Sentiment Analysis in the Wild held in conjunction with the biannual Conference on Affective Computing and Intelligent Interaction*. Xi'an, China: IEEE, 2015, pp. 891–897.

[30] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*. Lyon, France: ISCA, August 2013, pp. 148–152.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.