Introducing the COVID-19 YouTube (COVYT) speech dataset featuring the same speakers with and without infection

Andreas Triantafyllopoulos, Anastasia Semertzidou, Meishu Song, Florian B. Pokorny, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, Anastasia Semertzidou, Meishu Song, Florian B. Pokorny, and Björn W. Schuller. 2023. "Introducing the COVID-19 YouTube (COVYT) speech dataset featuring the same speakers with and without infection." *Biomedical Signal Processing and Control* 88 (A): 105642. https://doi.org/10.1016/j.bspc.2023.105642.

Introducing the COVID-19 YouTube (COVYT) speech dataset featuring the same speakers with and without infection

Andreas Triantafyllopoulos [a,*], Anastasia Semertzidou [a], Meishu Song [a,b], Florian B. Pokorny [a,c,*], Björn W. Schuller [a,d]

[a] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany*
[b] *Educational Physiology Laboratory, University of Tokyo, Tokyo, Japan*
[c] *Division of Phoniatrics and the Division of Physiology, Medical University of Graz, Graz, Austria*
[d] *Group for Language, Audio, & Music, Imperial College, London, United Kingdom*

ARTICLE INFO

Keywords:
COVID-19
Speech dataset
Speech pathology
Computer audition
Disease detection
Machine learning

ABSTRACT

More than two years after its outbreak, the COVID-19 pandemic continues to plague medical systems around the world, putting a strain on scarce resources, and claiming human lives. From the very beginning, various AI-based COVID-19 detection and monitoring tools have been pursued in an attempt to stem the tide of infections through timely diagnosis. In particular, computer audition has been suggested as a non-invasive, cost-efficient, and eco-friendly alternative for detecting COVID-19 infections through vocal sounds. However, like all AI methods, also computer audition is heavily dependent on the quantity and quality of available data, and large-scale COVID-19 sound datasets are difficult to acquire – amongst other reasons – due to the sensitive nature of such data. To that end, we introduce the COVYT dataset – a novel COVID-19 dataset collected from public sources containing more than 8 h of speech from 65 speakers. As compared to other existing COVID-19 sound datasets, the unique feature of the COVYT dataset is that it comprises both COVID-19 positive and negative samples from all 65 speakers. We additionally provide an overview acoustic analysis and modelling baselines using different partitioning strategies. We analyse the acoustic manifestation of COVID-19 on the basis of these perfectly speaker characteristic balanced 'in-the-wild' data using interpretable audio descriptors, and investigate several classification scenarios that shed light into proper partitioning strategies for a fair speech-based COVID-19 detection.

1. Introduction

In March, 2020, the world health organisation (WHO) has categorised the novel COVID-19 as a *pandemic*, i. e., a disease characterised by worldwide spread. Following this characterisation, and the immense accompanying strain on healthcare systems, several countries have taken a series of protective measures, including testing, mask mandates, movement restrictions, and vaccination campaigns — all in an attempt to stem the devastating effects of the virus. Today, more than two years after the outbreak of COVID-19, the world is still dealing with its repercussions. As of 1 December 2022, the WHO has documented more than 650 million cases. A substantial number of cases was only recorded in the first quarter of 2022 due to the recent surge of the Omicron variant of COVID-19. While governmental responses are now transitioning towards the endemic phase in some countries, COVID-19 remains a serious health issue that warrants our continued attention.

Widespread testing is a cornerstone of the response against COVID-19. It informs public health agencies about the extent of virus spread in the community, enables the detection of new, potentially dangerous, variants, and helps citizens protect themselves and those around them by seeking timely medical assistance and self-isolating. Currently, reverse transcription polymerase chain reaction (RT-PCR) and rapid antigen tests dominate the testing strategies used to identify COVID-19 positive cases.

Recently, a plethora of artificial intelligence (AI) tools have been proposed for the automatic detection of COVID-19 [1]; the main justification in their favour are their significantly lower costs, their eco-friendly nature, and their potential to be deployed at a vastly larger scale — which have been also showcased for other respiratory track diseases [2,3]. Suggested tools analyse different types of bio-signals to make their prediction, ranging from CT-scans [4–6] and chest X-rays

* Corresponding authors.
E-mail addresses: andreas.triantafyllopoulos@uni-a.de (A. Triantafyllopoulos), florian.pokorny@medunigraz.at (F.B. Pokorny).

[7,8], to heart rate signals [9], to vocal [10–15], coughing [15], and breathing [15–17] sounds. So far, none of them has received medical certification and is, thus, not part of any official testing strategy mainly due to a lower accuracy as compared to standard test approaches. However, as those tools become increasingly more sophisticated, and the virus seems to gradually transition to an endemic stage requiring less stringent monitoring, they can nicely complement the arsenal of COVID-19 detection mechanisms at the disposal of authorities and individuals alike.

The automatic analysis of CT-scans by means of computer vision techniques has shown much promise in detecting COVID-19 infection, with accuracies reaching over 90% in some studies [18]; yet the major downside is that this approach requires the use of sophisticated medical equipment (computer tomographs) and the (suspected) patient to visit a medical facility — both aspects thus hampering a large-scale applicability. In contrast, heart rate and/or vocal signals can be easily obtained using everyday sensors, such as wristbands and/or smartphones, and, thereby, provide a useful basis for an AI-based COVID-19 detection in a large group of people without a need for them to leave their homes.

In the present work, we focus on vocal sounds, in particular speech, as the bio-signal of choice for detecting and investigating the manifestation of COVID-19. COVID-19 as a respiratory disease suggests that acoustic information can assist in its detection: coughing, shortness of breath, and sore throat are amongst the most common reported symptoms. Patients with mild-to-moderate symptoms frequently report dysphonia. Accordingly, sound researchers have investigated the effectiveness of acoustic information to differentiate between patients with COVID-19 and controls. A substantial amount of that work has concentrated on non-verbal sounds, such as coughing or breathing [11,19]. In contrast, Bartl-Pokorny et al. [20] investigated sustained vowels. Both non-verbal sound data and sustained vowels are usually obtained through standardised procedures in controlled settings. Thereby, exactly the same type of vocal sound can be compared across patients and controls with less expectable effects of language and culture as compared to speech. However, a recorded sequence of speech covers a broader range of language-inherent sounds and sound transitions and might thus contain more potentially relevant acoustic information for distinguishing between individuals with and without a COVID-19 infection. Moreover, providing a speech sample usually represents a more natural setting for people than following the instruction to cough, breath, or produce a sustained vowel. While oftentimes audio-based methods show lower sensitivity to detecting COVID-19 compared to other methods, their widespread availability and ease of use without the help of medical personnel makes them a promising means for rapid, ubiquitous early screening, which can then be complemented by more reliable tests. Naturally, this form of screening will only be useful for cases where the disease manifests in auditory symptoms; however, the same can be said about other screening methods (e.g., temperature sensors rely on patients having a fever). Therefore, audio-based monitoring can be another valuable tool in monitoring the spread of COVID-19 in the community.

AI-based digital health tools are often criticised for making unrealistic assumptions that limit their applicability in real-world applications. Recently, Coppock et al. [19] provided an overview of seven specific criticisms on audio-based COVID-19 detection — denominated as seven ‘grains of salt’:

1. Just investigating COVID-19 vs healthy condition (neglecting other diseases),
2. Presence of (confounding) background noise,
3. Subject knowledge of infection status which potentially impacts vocal expression (e. g., through emotion),
4. Questionable validity of (self-reported) COVID-19 status,
5. Lack of sound data and code availability,
6. Ignorance of demographic characteristics,
7. Lack of speaker-disjoint experiments.

In the present work, we introduce the COVYT speech dataset, which attempts to mitigate several of those issues. Using an easily scalable collection and pre-processing protocol allowing to make data from social media channels scientifically exploitable, we provide a unique multilingual dataset for investigating COVID-19 detection from free speech that features COVID-19 positive and negative speech samples from exactly the same speakers. In this way, we control for bias potentially introduced by imbalances in intrinsic speaker characteristics, such as gender, age, and language. In addition, the presence of both positive and negative samples from the same speakers allows us to explore personalisation approaches which disentangle the effects of infection from individual voice characteristics that can confound analysis [3,14]. We present baselines with respect to (i) the manifestation of a COVID-19 infection in different acoustic descriptors and (ii) automatic COVID-19 detection in various scenarios addressing several factors that influence model performance. The COVYT dataset as well as the code for our machine learning experiments are publicly available to facilitate reproducibility and to motivate further research comparable to the provided baselines [21,22].

The remainder of this work is organised as follows. In Section 2, we introduce our COVYT speech dataset and present related data collection and pre-processing protocols, statistics, and partitioning. In Section 3, we compare the COVYT speech dataset to other relevant currently existing vocal sound datasets for COVID-19 detection. Section 4.3 then reveals our dataset baselines, while Section 5 positions the COVYT dataset and our findings with respect to previous research and discusses the strengths and limitations of our work, connecting it to the above mentioned seven ‘grains of salt’. We conclude our work in Section 7.

2. COVYT dataset

2.1. Data collection

As the pandemic is a global phenomenon that has dominated the public’s attention since its very beginning, people – and in particular celebrities – have often ‘announced’ their positive results on social media. Some of those cases, like that of the former US president Donald Trump Jr., receive considerable media attention due to the nature and position of the person affected. It became common for news outlets to run features using footage of well-known people discussing their experiences on having (had) a COVID-19 infection; or even for celebrities spreading footage by themselves through their private channels. Such footage is typically recorded in the days following a positive COVID-19 test, when subjects are required to stay in quarantine. Hence, the COVID-19 status labels are self-reported and cannot be officially verified.

The data collection phase for the COVYT speech dataset took place between November 2020 and November 2021. During that time, two popular media platforms are combed for appropriate material, namely YouTube and TikTok. Different data collection protocols for positive cases are utilised for each platform. A targeted search is performed on YouTube, where high-profile cases, e. g., actors, politicians, celebrities, etc. that come to the authors’ attention through the media, are intentionally looked for. In contrast, a global search is performed on TikTok by using keywords like “COVID” or “symptoms”. In both scenarios, when finding an eligible COVID-19 positive example, we search for uploads of the same speakers preceding the date of infection to serve as negative examples. In case we easily find more than one positive examples of a speaker, all clips are included. Different protocols serve to mitigate potential biases in our data collection process by incorporating diverse recording scenarios, ranging from high-quality, professional interviews to homemade smartphone videos. In all identified clips, the speakers were audio-video recorded or recorded themselves while talking, e. g., during an interview, a public speech, a narrative, or a ‘story’. As most clips were released by speakers to explicitly inform the community about their infection status, we are able to harness them for

Table 1
Overview of dataset statistics. We show the language-wise and total number of (#) speakers as well as # utterances and utterance duration for the respective points in time at which the speakers had a COVID-19 infection (T+) and the respective points in time at which the speakers did not have a COVID-19 infection (T−). Gender-wise numbers and durations are given in parentheses in order male/female.

		T+		T−	
Language	# Speakers	# Utterances	Duration (hh:mm:ss)	# Utterances	Duration (hh:mm:ss)
Chinese	2 (0/2)	112 (0/112)	00:03:43 (00:00:00/00:03:43)	28 (0/28)	00:00:53 (00:00:00/00:00:53)
English	40 (28/12)	2454 (1988/466)	01:59:45 (01:36:00/00:23:45)	4686 (4068/618)	03:45:54 (03:12:52/00:33:03)
French	2 (2/0)	84 (84/0)	00:04:10 (00:04:10/00:00:00)	96 (96/0)	00:03:04 (00:03:04/00:00:00)
German	2 (1/1)	27 (16/11)	00:01:26 (00:00:48/00:00:38)	85 (65/20)	00:04:28 (00:02:59/00:01:29)
Greek	14 (5/9)	544 (187/357)	00:29:28 (00:08:33/00:20:55)	869 (179/690)	00:43:27 (00:08:46/00:34:41)
Polish[a]	1 (1/0)	23 (23/0)	00:02:01 (00:02:01/00:00:00)	254 (254/0)	00:11:59 (00:11:59/00:00:00)
Portuguese	1 (1/0)	21 (21/0)	00:00:50 (00:00:50/00:00:00)	118 (118/0)	00:05:33 (00:05:33/00:00:00)
Slovakian	1 (1/0)	54 (54/0)	00:02:54 (00:02:54/00:00:00)	273 (273/0)	00:12:41 (00:12:41/00:00:00)
Spanish	2 (1/1)	35 (18/17)	00:01:02 (00:00:29/00:00:33)	650 (632/18)	00:23:35 (00:22:32/00:01:03)
Total	**65 (40/25)**	**3354 (2391/963)**	**02:45:18 (01:55:45/00:49:33)**	**7059 (5685/1374)**	**05:31:35 (04:20:26/01:11:09)**

[a] A portion of the negative utterances of this speaker is actually in English; however, we consider this to have a negligible effect on our analysis.

acoustic analysis and model training. Recordings during the COVID-19 positive state were taken mostly indoors – at home, a hospital, a TV/Radio studio, or a press conference room – as subjects had to be under quarantine. To avoid undesired interference, only videos with minimal background noise or quiet music, e. g., from news reports and media covers, are accepted for this work. In total, we include 185 videos — 89 positive and 96 negative examples. Of these, 83 just contain a single speaker, whereas the rest contain multiple ones — requiring additional processing to extract the utterances of the target speaker.

2.2. Data preparation

First, we download all identified videos in .MP4 format using freely available tools. We then manually annotate the acoustic environment of each video according to recording location — 'indoor' vs 'outdoor' — and recording setting. For recording setting, we distinguish between: 1. speeches (or press-releases), which are longer recordings, where the target speaker releases a statement in front of staged, professional-grade cameras; 2. interviews, where the target speaker is part of a (live or online) conversation; and 3. self-recordings, where the target speaker uses his or her own smartphone to make a short video (usually a social-media-style 'story'). Following this annotation, we extract the audio streams and store them as .WAV files in format 16 kHz, 16 bit, single-channel, PCM. The clips range in duration from 10 s to 1 h 11 m — some clips contain long monologues of the target speaker. Therefore, we segment all clips into single utterances for further processing. We choose a semi-automatic segmentation approach conservative in the amount of utterances it keeps. We employ a recurrent neural network (RNN)-based voice activity detection (VAD) model [23] as the initial stage, followed by a manual verification stage using ELAN[1] [24]. Utterances that do not exclusively contain speech of the target speaker or samples that contain music or background noise are excluded.

2.3. Facts and figures

An overview of dataset statistics is given in Table 1. The COVYT dataset contains 10 413 utterances with a total duration of 8 h 15 m of speech from 65 speakers – 25 females and 40 males – at ages ranging from 23 to 74 years at time of infection (mean = 46 years ± 13 years standard deviation). Each speaker has an average number of 161 utterances — 52 COVID-19 positive and 109 COVID-19 negative examples. Henceforth, the respective point in time at which the speakers had a COVID-19 infection is referred to as T+; the point in time at which they did not have a COVID-19 infection is referred to as T−. Speakers are celebrities of various domains: actors, athletes, journalists, models, musicians, politicians, presenters, reporters, singers, and writers. Moreover, the dataset covers 9 different languages, namely Chinese (CN), English (EN), French (FR), German (DE), Greek (GR), Polish (PL), Portuguese (PT), Slovakian (SK), and Spanish (ES), with a majority of English speakers (40) followed by Greek speakers (14). The number of speakers and utterances is given in detail in Table 1. With regard to location and setting, most clips were recorded indoors rather than outdoors (T−: 92 vs 4; T+: 83 vs 6); most clips were recorded in an interview setting (T−: 55; T+: 27), followed by a speech/press-release setting (T−: 29; T+: 6), and, lastly, by a self-recording setting (T−: 12; T+: 56). Naturally, the chance to give a speech or an interview decreases following a COVID-19 diagnosis, as speakers have to self-quarantine or are hospitalised for a certain time. Interviews at T+ took place through online teleconferencing, and speeches were made in front of staged cameras (presumably without the presence of reporters or assistants due to quarantine restrictions).

2.4. Partitioning

Proper partitioning is a crucial aspect of any dataset if used for machine learning (ML) purposes, as the data must be split in a way that allows for building well-performing models, but also enables a fair evaluation. Taking into account the relatively small size of the COVYT dataset, with a total of 10 413 samples, we opt for a cross-validation scheme, for which we provide training/development/testing folds. Given that we also aim to investigate different scenarios of COVID-19 detection, we introduce four different partitioning strategies, each targeted to a different aspect of interest. Some of our partitioning strategies are intentionally not speaker-independent to test the influence of having the same speakers in the training and evaluation folds. The number of speakers and samples for each fold in each partition is shown in Appendix.

1. **Speaker-disjoint partitioning:** As discussed in Coppock et al. [19], a major limitation of several existing COVID-19 datasets is their lack of subject-disjoint evaluation. If data from the same speaker is involved in both training and testing, performance will be most probably higher because the model might re-identify the speaker identity instead of performing the actual target task. Individual voice characteristics often negatively impact generalisation; having the same speaker in the training and testing partitions results in obtaining over-optimistic performance scores. This particular partitioning scheme is implemented by randomly splitting the speakers into 3 disjoint groups (G_1, G_2, G_3), and subsequently taking all possible permutations of this set, resulting in 6 folds (the permutations are obtained by training on G_1, validating on G_2, testing on G_3, then training on G_2, validating on G_1, testing on G_3 — so each fold is train/validated/tested on twice).

[1] https://archive.mpi.nl/tla/elan

Table 2
Comparison of the COVYT dataset with other datasets for audio-based COVID-19 detection.

Dataset	# Speakers (+)	Sound type	# Languages	Data availability	COVID-19 labels	± Same speaker	Baseline
COVIDTelephone [25]	19 (10)	free speech	N/A	✓	level-1	✗	AUC: .92
AI4COVID [26]	543 (70)	coughing	N/A	✗	unknown[a]	✗	Acc: .93
Coswara [27]/DiCOVA [28][b]	941 (104)/990 (60)	breathing coughing vowels digit counting	multiple	✓	level-1	✗	AUC: .74
COUGHVID [29]	14787 (410	coughing	N/A	✓	level-1	✗	N/A
YourVoiceCounts [20]	22 (11)	coughing vowels read speech	1	✗	level-2	✗	N/A
CoughAgaistCovid [30]	3621 (2001)	coughing	N/A	✗	level-2	✗	AUC: .68
COVID19 [31]	78 (29)	coughing vowel/voiced consonant counting digits	1[c]	✗	level-2	✗	AUC: .81
YourVoiceCounts audEERING [10]	39 (19)	coughing vowels read speech	1	✗	level-1/2	✗	UAR: .63
Cambridge Longitudinal [14]	212 (106)	breathing coughing read speech	8	🔒	level-1	✓	AUC: .79
COVID-19 Sounds [32]	36116 (2496)	breathing coughing read speech	8	🔒	level-1	✓	AUC: .71
COVYT	65 (65)	free speech	9	✓	level-1	✓	

= number of; + = COVID-19 positive; ± = COVID-19 positive and COVID-19 negative (samples)
level-1 self-assessment/no test required; *level-2* Medically certified COVID-19 test result required.
[a] Not specified; however, most likely level-2 as the study was conducted in medical facilities.
[b] DiCOVA is derived from Coswara; Coswara is constantly growing so DiCOVA is bigger than the 1st version.
[c] Not mentioned; presumably Hebrew as the study was conducted in Israel.

2. **Speaker-inclusive partitioning:** This partitioning is intentionally introduced to quantify the effect of having negative and positive data of the same speakers in training and testing. The data of all files from each speaker is randomly split into three partitions; then, the partitions are permuted to obtain 6 folds in a similar fashion as before.
3. **File-disjoint partitioning:** With this partitioning strategy, we want to investigate whether a speaker-inclusive, but file-disjoint partitioning would also yield a disproportionately good performance. In particular, we ensure that the same speaker is present in the test and either the training or development partitions, but the same file is not. This enables us to quantify whether it is indeed the individual speaker characteristics that cause models to overperform in the speaker-inclusive scenario, or whether it is simply the side-effect of having an identical recording (which contains data from the same acoustic conditions). Given that for most speakers the COVYT dataset contains only one original clip per state (COVID-19 negative or COVID-19 positive), we randomly split speakers into two groups, G_1 and G_2. We then create two test sets, T_1 and T_2, with T_1 containing all T+ instances of G_1 and all T− instances of G_2, and T_2 in contrast with all T− instances from G_1 and all T+ instances of G_2. This ensures that each test set contains files from all speakers but in only one of the two conditions (COVID-19 negative or COVID-19 positive), thus, ensuring that files are disjoint (as each recording contains the speaker in only one condition). For each fold, we additionally create 2 variants where the training set is randomly split into 2 training/development sets (this time in speaker-disjoint fashion, as most speakers have only one clip per condition), resulting in a total of 4 folds when taking all permutations.
4. **Language-disjoint partitioning:** For this final scheme, we split the data by taking language information into account. As most of our data is either English or Greek, we bundle all other languages in a separate *'other'* group, thus resulting in 3 folds. Once again, we utilise all permutations to create 6 folds.

We note that the information regarding partitioning is included in the metadata released together with the dataset, thus ensuring reproducibility of our results and a fair comparison of different scenarios. However, not all partitioning schemes are relevant for potential real-world applications. We encourage authors to primarily use our *speaker-disjoint* partitioning for future work.

3. COVYT vs other datasets

Table 2 provides a representative overview of currently existing COVID-19 sound datasets and specifies aspects that the COVYT dataset is meant to improve on. In particular, we focus on the type of audio content that each dataset contains. The COVYT dataset contains free (multilingual) speech; this is in contrast to most other listed datasets that primarily focus on breathing and coughing sounds, on sustained vowels, or on reading standard texts. Besides some advantages of free speech (please also see Section 1; coverage of several language-inherent sounds, natural recording setting), our motivation to use this type of sound data is a pragmatic one: Free speech is most easily available on the data sources we examined. The amount of different languages contained in a dataset and their distribution are mostly relevant for those including linguistic sound types, such as speech or sustained vowels; however, it also serves as a proxy for demographic diversity which is an important consideration for making datasets fair and audio-based COVID-19 detection applicable across different countries. This is why we make the COVYT dataset as diverse as possible and provide different partitioning strategies including a language-disjoint one. Data availability is in turn crucial for ensuring study reproducibility and transparency of research findings, and also for fostering new advances from researchers without access to own datasets. Here, the COVYT dataset enjoys an advantage over other datasets, as it is sourced entirely from the public domain.

This leads to the issue of label reliability. Relying on self-reported labels and crowdsourced data holds lots of potential for scaling the size of a dataset. However, it comes with the danger of erroneous labels. In contrast, manually collected datasets with strict medical protocols,

where the label is verified through proper medical examination and further audited by a medical practitioner is the gold standard for digital health applications, but is harder to scale due to the amount of resources it requires. Our approach lies somewhere in between. While, technically, we still rely on self-reported labels, we make use of the scrutiny that celebrities are subjected to from the press, which renders a fake COVID-19 report rather unlikely (though not impossible). To roughly distinguish between different quality types of labelling protocols, we adopt a 2-level system, which ranks datasets according to whether they solely rely on self-reported labels (*level-1*), or whether a medically certified test result was required (*level-2*). The COVYT dataset falls under the *level-1* category.

Additionally, we include another indicator in our comparison, namely whether a dataset contains samples of the same speaker with and without infection. We consider this as an important aspect towards minimising potential biases caused by imbalances in specific speaker characteristics, such as gender, age, or any other intrinsic anatomical/voice-physiological properties. Investigating exactly the same voice with vs without infection across many speakers, respectively, increases the chance to effectively identify disease-related phenomena. Moreover, recent findings from the *Cambridge Longitudinal* dataset [14] suggest that there are individual effects in the manifestation of COVID-19 in the voice (in their case, non-linguistic vocalisations). To the best of our knowledge, the COVYT dataset is the only dataset alongside *Cambridge Longitudinal* and *COVID-19 Sounds* (both variants of the same data) to fulfil the criterion of having COVID-19 positive and negative samples from the same speakers, and, in particular, the only one to fulfil this criterion and to contain free speech samples, which enables the use of personalised ML algorithms, which have been shown to improve performance in other speech-based tasks [33]. However, it has to be considered, that the recording setting (recording equipment, location, situation, speaker mood, etc.) might differ between the respective positive and negative sample of one and the same speaker, which potentially introduces systematic acoustic bias.

Finally, we included the baseline model performance reported with the introduction of each dataset (when available). Different authors computed different metrics; most used area under the curve (AUC), whereas one work used accuracy (ACC) and another unweighted average recall (UAR). This allows us to compare with our work, where we use fairly straightforward baseline methods. Performance ranges from a lowest 63% UAR to a highest 93% accuracy. This illustrates how performance might vary depending on dataset, which can be explained by various reasons, such as recording conditions or patient demographics.

4. Baseline evaluations

In the following, we provide standard solutions for answering two main research questions (RQs) on the basis of COVYT data: (4.2) Which speech parameters differ most between speakers at T− vs T+? (4.3) Can T− and T+ be automatically differentiated from speech? To this end, we derive different audio representations from the available utterances. The generated results shall serve as a benchmark for future analyses carried out using the COVYT dataset.

4.1. Audio representations

Speech processing applications typically rely on hand-crafted sets of less than 100 until several 1000 features, which attempt to holistically describe an input utterance [34,35]; importantly, certain features allow for signal interpretation from a voice-physiological perspective. In recent years, however, learnt representations have shown superior performance and robustness in several tasks [36,37], driving their adaptation from the community. In the present study, we carry out experiments using three different audio representations:

We begin with the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [34] — a rather small set of 88 acoustic parameters that has previously been shown to contain relevant information for the manifestation of COVID-19 in sustained vowels [20]. In contrast, the Interspeech Computational Paralinguistics ChallengE (ComParE) set is a large-scale feature set of 6 373 acoustic parameters. As the official baseline feature set of the ComParE series from 2013 until 2021 [35,38], it has been successfully used for several computer audition tasks over the last decade. Both the eGeMAPS and the ComParE set are extracted using the open-source toolkit openSMILE [39].

Finally, we use learnt representations from w2v2-large-xlsr [37], a multilingual variant of wav2vec2.0 [36]. This model was pre-trained in self-supervised fashion on a large corpus containing 53 languages. We thus expect this network to generalise better to our multilingual data than the vanilla wav2vec2.0. The architecture consists of 7 convolutional neural network (CNN) feature extraction layers followed by 24 transformer (self-attention) layers. Here, we use the intermediate features which are extracted by the CNN layers. These roughly correspond to 25 ms of audio with a stride of 20 ms, which we subsequently average over time to obtain the final embeddings. We also experiment with the contextualised representations, which are extracted after the self-attention layers. However, we get consistently worse performance. As our intention here was to obtain competitive baselines, we do not fine-tune w2v2-large-xlsr on COVYT, even though this is known to lead to substantially better performance [40].

4.2. Acoustic analysis (RQ 1)

The analysis of acoustic differences between speakers at T+ vs T− is done on the basis of the extracted eGeMAPS representation, as eGeMAPS features generally offer interpretability from a clinical/voice-physiological perspective. To ensure equal speaker and COVID-19 status weighting, we average feature values across all utterances of a single speaker at T+ and T−, respectively. Thus, for each of the 88 eGeMAPS features, we produce exactly one value per speaker at T+ and one value per speaker at T−. We find that the values of the single feature at T+ and T− are not normally distributed. Thus, we feature-wisely apply the Mann–Whitney U test and derive the effect size r, i. e., the absolute value of the correlation coefficient calculated as the z-value divided by the square root of the number of samples [41]. We finally rank the eGeMAPS features according to the effect size and define top features to have an $r > 0.3$ (fair correlation at minimum). In addition, we report two-sided p-values; however, we do not consider them to accept or reject a null-hypothesis and, therefore, do not adjust them for multiple comparisons.

We identify three top features, namely (I) the coefficient of variation of the spectral flux, i. e., spectral change between consecutive time frames [34], in voiced regions, (II) the coefficient of variation of local shimmer, i. e., change in amplitude between consecutive fundamental frequency periods [34], and (III) the coefficient of variation of the spectral flux in the entire speech segment. Fig. 1 reveals the respective boxplots for T− vs T+ alongside the effect sizes and p-values. In all three top features, T+ is characterised by a lower coefficient of variation as compared to T−. This means that there is restricted variation with regard to spectral and amplitude change within an utterance in COVID-19-related speech. Fig. 2 exemplarily shows the spectrograms of an utterance produced at T− and an utterance produced at T+. Both utterances originate from the same speaker, namely the speaker with the highest average top feature (coefficient of variation of spectral flux in voiced regions) difference between T+ and T−. Obviously, the utterance produced at T+ exhibits more inharmonic overtones in voiced sounds as compared to the utterance produced at T−, which is associated with more vocal coarseness as typical for a respiratory disease. Moreover, the presented spectrogram related to T+ indeed suggests less variation of spectral change over time in voiced regions. However, this finding has to be interpreted with caution, as not only

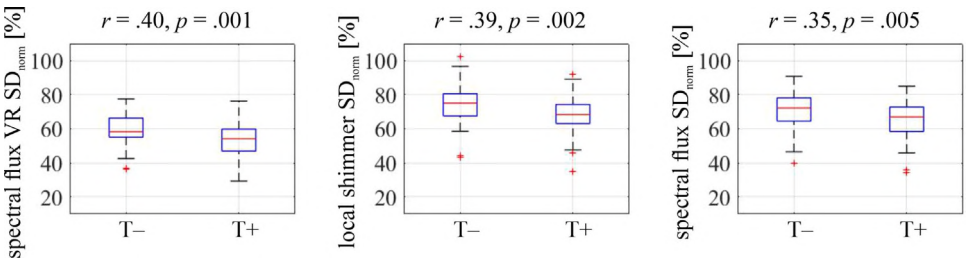

Fig. 1. Comparison of speakers at their respective point in time without COVID-19 infection (T–) vs the point in time with COVID-19 infection (T+) by means of boxplots for the three identified acoustic features with a differentiation effect $r > 0.3$. Boxplots are ordered according to a decreasing r from left to right. Effect size r and p-value of the Mann–Whitney U difference test are given above each boxplot. r is rounded to two decimal places. p is rounded to three decimal places. Outliers (red plus symbols) are defined as values more than 1.5 times the interquartile range away from the bottom or top of the respective box. SD_{norm} = standard deviation normalised by arithmetic mean (= coefficient of variation), VR = voiced regions.

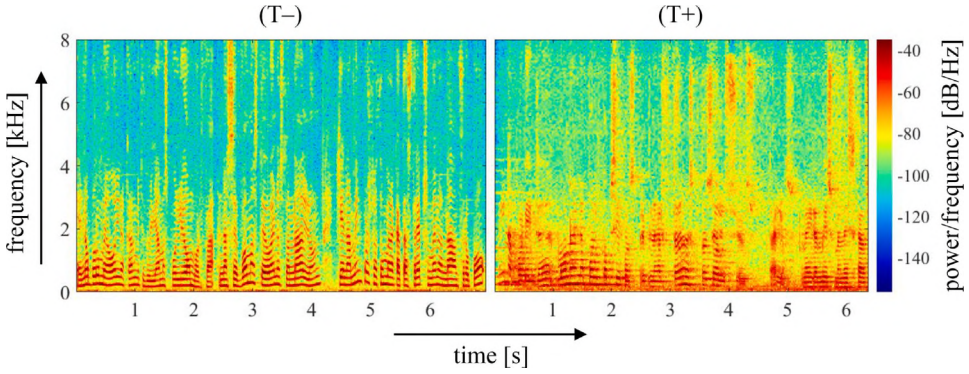

Fig. 2. Spectrograms of utterances produced by one and the same speaker (gender: female, nationality: Greek, age at T+: 43) at a point in time without COVID-19 infection (T–) and at a point in time with COVID-19 infection (T+).

mechanisms of voice production but also room acoustics affect the recorded audio signal and, thereby, a wide range of derived acoustic features. An auditory inspection of the utterances presented in Fig. 2 yields that the utterance produced at T+ was obviously recorded in a room with a longer reverberation time as compared to the utterance recorded at T– (please also see limitations discussed in Section 6).

4.3. Automatic COVID-19 detection (RQ 2)

In this subsection, we present our automatic COVID-19 classification performance results on all four partitioning strategies described in Section 2.4. In all cases, we perform z-score normalisation on each feature separately, computing the statistics on the training set of each fold in our cross-validation setup separately, and consequently applying them on the development and test partitions. As a typical baseline classifier, we utilise support vector machines (SVMs), where we optimise the complexity parameter C in [.0001, .0005, .001, .005, .01, .05, .1, .5, 1] as well as the kernel function across the types linear, polynomial, and radial basis function (RBF). Optimisation is always done on the development partition of each fold. We additionally utilise two distinct evaluation protocols, each highlighting a different aspect of the underlying problem. As our metric of choice we use the AUC for both.

The first protocol is adhering to standard ML practice. As discussed in Section 2.2, the downloaded clips are segmented into utterances, which we use as training/validation/testing instances in the context of a ML pipeline. These instances inherit the label (presence or absence of a COVID-19 infection) from the file they originate from. In each fold, the model is trained on the training instances and evaluated on the test instances, resulting in one prediction value per instance. These predictions are then evaluated against the ground truth labels over the entire test set to provide a general, segment-level performance score. This procedure is repeated for each fold in each partitioning scheme. Within the standard evaluation protocol, this procedure is affected by the fact that utterances coming from the same file are not independent instances. This biases our results and in particular confidence measures and subsequent analysis. Therefore, we only report it for the sake of completeness and focus on the following, second protocol.

The second protocol is motivated by the clinical evaluation setting for which our application is intended. Under this perspective, the different utterances resulting from the segmentation process can be seen as repeated measurements of the same underlying variable, namely the manifestation of COVID-19 in the speaker's voice. In this setting, segment-level decisions are aggregated to provide a final, holistic evaluation, which takes all utterances into account and provides a single label for each file in the test partition. As our present focus is on providing a set of competitive baselines, we adopt the simplest possible aggregation process, namely *max voting*, where the label corresponding to each file is defined as the most often-predicted label across all its utterances. As this results in only one prediction per file, we have now independent measures of each speaker's COVID-19 status for each file.

We present AUC results for all partitioning scheme and audio representation combinations in Table 4, as well as the accompanying ROC curve for the speaker disjoint partitioning in Fig. 3. For all partitioning strategies, we compute average AUC and 95% bootstrap confidence intervals (CIs), using 1000 bootstrap samples. For file-level results, we compute the CIs by sampling instances with replacement from the file-level predictions (this is possible because they are independent). For segment-level results, where predictions per file are *not* independent, we compute the CIs as follows: (a) We first sample speakers randomly with replacement. This accounts for the lack of independence caused by the same speaker having multiple utterances in the same state from the same file. (b) For each sampled speaker, we randomly sample utterances (also with replacement) from their pool of available utterances. This accounts for the randomness caused by the different utterances as some might have better predictive power than others.

Table 3
Average Brier score and 95% CIs for all partitioning strategies and audio representations using support vector machines (SVMs) with different features.

Partitioning	Speaker-disjoint	Speaker-inclusive	File-disjoint	Language-disjoint
eGeMAPS	.245 (.181–.317)	.121 (.095–.152)	.258 (.205–.318)	.297 (.221–.374)
ComParE	.206 (.157–.258)	.120 (.092–.151)	.206 (.157–.258)	.246 (.176–.326)
wav2vec2.0	.187 (.115–.269)	.038 (.018–.061)	.202 (.144–.272)	.237 (.159–.324)

Table 4
Cross-validation area under the curve (AUC) results for all partitioning strategies and audio representations using support vector machines (SVMs) with different features. We report mean AUC and 95% confidence intervals over all folds.

Partitioning	Speaker-disjoint		Speaker-inclusive		File-disjoint		Language-disjoint	
Audio representation	Segment-level AUC	File-level AUC	Segment-level AUC	File-level AUC	Segment-level AUC	File-level AUC	Segment-level AUC	File-level AUC
eGeMAPS	.629 (.453–.774)	.687 (.544–.816)	.952 (.929–.967)	.919 (.873–.958)	.658 (.529–.781)	.685 (.569–.794)	.534 (.357–.690)	.586 (.415–.760)
ComParE	.761 (.599–.883)	.779 (.649–.891)	.960 (.935–.977)	.916 (.872–.954)	.772 (.635–.865)	.769 (.668–.861)	.643 (.490–.781)	.722 (.558–.860)
wav2vec2.0	.758 (.586–.902)	**.818 (.693–.920)**	.997 (.994–.999)	**.984 (.966–.997)**	.816 (.695–.915)	**.798 (.698–.884)**	.706 (.536–.860)	**.736 (.577–.876)**

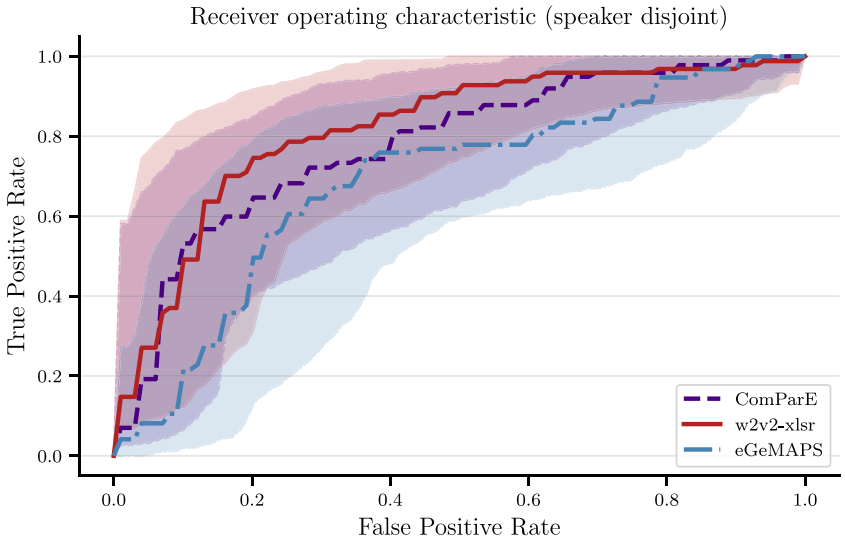

Fig. 3. ROC curve with mean (lines) and 95% CIs (shaded areas) of true positive rates over all folds for the speaker disjoint partitioning.

(c) We accumulate all bootstrap AUC estimates and compute 95% CIs for each fold. (d) We compute the average CI ranges over all folds for each partition scheme. At first glance, wav2vec2.0 results in the highest performance across all schemes with an AUC of .818 on the speaker-disjoint partition, followed by ComParE (AUC: .779), while eGeMAPS is lagging further behind (AUC .687). This demonstrates the power of learnt representations for COVID-19 detection in free speech samples.

We additionally report Brier scores (average and 95% CIs) for file-level probabilities in Table 3. Brier score is the mean squared difference between classifier probabilities and classes. It takes values in the range of ($\{0, 1\}$; 0: perfect accuracy, 1: perfect inaccuracy), with values $\leq .25$ being better than a classifier assigning an equal random probability (0.5) to both outcomes. It thus quantifies how well the probabilities are *calibrated*, beyond the accuracy of the predictions. We note that for SVMs classifiers, probabilities are derived using Platt scaling, which does not always correspond to the model decision; nevertheless, we consider it an acceptable proxy. File-level probabilities are computed by taking the average over segment-level probabilities. For the speaker-disjoint experiments, we observe lower scores for wav2vec2.0 features (.187) indicating a better calibration than ComParE (.206) and eGeMAPS (.247).

The different partitioning schemes allow us to find the answers of two important questions:

1. *How important is the use of speaker-disjoint sets?* The vastly superior performance of all audio representations on speaker-inclusive experiments, where wav2vec2.0 reaches an average AUC of .997 on the instance- and .985 on the file-level, compared to the alternative partitioning strategies, where wav2vec2.0 reaches a maximum average AUC of .758 and .818 on instance- and file-level evaluations, respectively, shows that this partitioning can result in a large overestimation of real-world performance and should, therefore, be avoided. However, the drop of performance observed when switching to file-, but not speaker-disjoint evaluations also indicates that it is not necessarily the presence of the same speaker in training and testing partitions per se that causes this overestimation, but potentially (also) the splitting of the same recording (whose utterances form repeated measurements of the same speaker in the same acoustic environment). Actually, the difference between our speaker-disjoint and file-disjoint partitioning schemes is negligible, showing that the speaker effect is small when taking care to avoid other confounders (e. g., recording conditions, background noise, etc.). Still, to be on the safe side, future

work on the COVYT dataset and comparable datasets (where applicable) should avoid the use of speaker-inclusive partitions in order to obtain accurate generalisation estimates.

2. *Is a speech-based detection of COVID-19 universally possible, i. e., across languages?* The evidence here is inconclusive. Our language-disjoint partitioning turns out to be the most challenging for the audio representations tested here, with performance dropping to .706/.736 even for the best-performing wav2vec2.0, while resulting in random-chance performance for eGeMAPS (.534/.586). This is expected, as several speech-based computer audition tasks struggle with cross-language generalisation. Even so, the fact that the learnt representations of wav2vec2.0, which has been pre-trained on several languages, still shows the highest performance is a promising sign that future 'multilingual' foundation models might bridge this gap and form the basis of a universal COVID-19 detection model. Though, until that time, it is highly recommended that models are applied with special care to different languages/cultures, and (if possible) evaluated on a representative test set beforehand.

5. Discussion

Our findings confirm that COVID-19 detection from free speech samples is feasible. Using standard procedures, we are able to obtain detection rates in the range of 70%–80%AUC for the most realistic testing scenarios — which are comparable to those achieved for other datasets with free speech. For example, Schuller et al. [38] report the best result for the COVID-19 speech sub-task of the 2021 ComParE (the baseline) at 72% unweighted average recall (UAR), whereas Dang et al. [14] report an AUC of .66 on *CambridgeLongitudinal*, which increases to .79 when utilising the longitudinal nature of the data. This showcases that speech is suitable for the detection of COVID-19. Free speech is easier to obtain than read texts/scripted speech, as it can be unobtrusively collected by means of passive recordings instead of requiring the speaker to actively record data and follow any instructions; we therefore expect this type of data to prove highly relevant for future COVID-19 speech research.

An undesirable side-effect of the 'wildness' of our data is that it makes the task harder to solve with interpretable features. This is reflected both in the lower detection performance obtained with eGeMAPS and the fact that our acoustic analysis did not yield that clear trends as seen, e. g., in studies using data acquired in a more controlled setting, such as Bartl-Pokorny et al. [20]. Nevertheless, there is some consistency between our present findings and those of Bartl-Pokorny et al. [20]: Both studies report shimmer and spectral flux to be relevant for the identification of speakers with COVID-19. The fact that other features, such as variations in the fundamental frequency or the harmonics-to-noise ratio are not found to be relevant in our study might be related to the different sound type used — free speech in the present study vs sustained vowels in Bartl-Pokorny et al. [20]. Moreover, it needs to be considered that the COVYT dataset used in the present study includes multiple languages and, thus, a huge range of phonemes and phoneme transitions, making a feature comparison with monolingual studies difficult.

6. Contributions and limitations

The COVYT dataset is the first of its kind COVID-19 speech dataset sourced from public multimedia platforms — a heretofore untapped resource for such data. Moreover, the presence of the same speaker with and without infection makes it a natural candidate for personalisation approaches, which are expected to improve COVID-19 detection performance as previous work found its symptom manifestation to have an individualised component [14]. These two aspects, combined with the open-source nature of the data, make the COVYT dataset a prime basis for further research in voice-based COVID-19 detection, which has a huge potential to massively increase future testing capacities while saving waste at the same time. However, the COVYT dataset inherently comes with a number of limitations, especially as the exploited video clips were not recorded with the intention to generate data for later scientific analyses. We thus return to the 'seven grains of salt' of Coppock et al. [19], and attempt to position the **contributions** and **limitations** of our work accordingly.

1. *COVID-19 vs other diseases*. The COVYT dataset does not fulfil this requirement as it primarily contains healthy or COVID-19 positive samples. Information about potential chronic diseases, such as asthma, are not reliably available. However, this also holds true for most previous works. → **Minor limitation**.
2. *Background noise*. Due to our strict data collection and preparation protocols, we expect only a limited amount of background noise (if at all) to be present in our processed utterances. Furthermore, by collecting data 'in-the-wild' from several sources, we cover a wide range of recording conditions relevant for potential future test applications while still ensuring a higher data quality as compared to fully crowd-sourced datasets. On the downside, most COVID-19 positive samples were recorded in isolated environments without other people present, as subjects were under quarantine, whereas COVID-19 negative samples encompass a wider gamut of recording environments; see data statistics in Section 2.3. Thus, we are aware of potential systematic differences in background noise and room acoustics/reverberation as well as recording setting, which models may be able to exploit. → **Balanced**.
3. *Subject knowledge of infection status*. Subjects were not only aware of their infection status, but in many cases created the recordings ad hoc to convey this status to a wider public. Given that our dataset consists of celebrities who rely on building emotional ties with their audience, it is highly possible that some of them modulated their voice accordingly. Thus, (intended) emotion could be a potential confounder when building COVID-19 detection models on the basis of the COVYT dataset. Nevertheless, (negative) affect is also a potential disease indicator [42]. Furthermore, it is possible that the vocabulary speakers use when infected (T+) and are making a press-release/interview contains explicit mentions to their health status which can act as 'shortcuts' for the models to learn instead of the actual task. This is particularly relevant for large, pre-trained models like wav2vec2.0 which are known to rely on linguistic information (when available) [40,43,44]. However, this effect is stronger in the deeper transformer layers [43,44], not the earlier convolution ones whence we extracted our embeddings here — thus we expect this effect to be absent from our study. → **Major limitation**.
4. *Validity of labels*. Although the labels used here are essentially self-reported, the high level of scrutiny which celebrities are being subject to (especially w. r. t. a positive diagnosis in the early days of the pandemic) strengthens our confidence in label validity. More problematic than the fact of just knowing a speaker's COVID-19 status in terms of negative vs positive is the missing knowledge about (i) the period between a positive COVID-19 test and the time of recording, (ii) the type of the used COVID-19 test, (iii) the cycle threshold (CT) value at the time of recording in case of a PCR test, (iv) the specific COVID-19 variant, (v) the range and severity of the speaker's symptoms at the time of recording, (vi) potential diagnoses of other (chronic) diseases, (vii) the speaker's vaccination status, etc. → **Minor limitation**.
5. *Data and code availability*. The COVYT dataset as well as the code for all experiments presented in this work are publicly released (see Section 1). → **Major contribution**.

Table A.5
Number of speakers and utterances (in parentheses) for each fold and partitioning strategy.

Partitioning	Speaker-disjoint			Speaker-inclusive			File-disjoint			Language-disjoint		
Fold	Train	Dev	Test	Train	Dev	Test	Train	Dev	textbfTest	Train	Dev	Test
Fold 1	19 (3361)	23 (3804)	23 (3248)	65 (3463)	65 (3574)	65 (3376)	28 (2650)	28 (2651)	37 (5112)	40 (7080)	14 (1413)	13 (1920)
Fold 2	19 (3361)	23 (3248)	23 (3804)	65 (3463)	65 (3376)	65 (3574)	28 (2651)	28 (2650)	37 (5112)	40 (7080)	13 (1920)	14 (1413)
Fold 3	23 (3804)	19 (3361)	23 (3248)	65 (3376)	65 (3574)	65 (3463)	37 (2556)	37 (2556)	28 (5301)	13 (1920)	14 (1413)	40 (7080)
Fold 4	23 (3804)	23 (3248)	19 (3361)	65 (3376)	65 (3463)	65 (3574)	37 (2556)	37 (2556)	28 (5301)	13 (1920)	40 (7080)	14 (1413)
Fold 5	23 (3248)	19 (3361)	23 (3804)	65 (3574)	65 (3463)	65 (3376)	N/A	N/A	N/A	14 (1413)	40 (7080)	13 (1920)
Fold 6	23 (3248)	23 (3804)	19 (3361)	65 (3574)	65 (3376)	65 (3463)	N/A	N/A	N/A	14 (1413)	13 (1920)	40 (7080)

6. *Demographic variability*. The COVYT dataset does not cover subjects from a wide range of socioeconomic backgrounds; celebrities typically come from the upper echelons. Nevertheless, we provide samples in 9 different languages, produced by speakers from different ethnicities (not all English speakers were natives) and age groups. → **Minor contribution**.
7. *Speaker-disjoint experiments*. Speaker identity is available for all utterances. Thus, speaker-disjoint partitions can be created. We provide a baseline partitioning scheme to allow standardised evaluation protocols. Furthermore, the COVYT dataset is the only dataset alongside the *Cambridge Longitudinal* dataset [14], which contains data of the same speakers with and without infection, which enables future research in personalised approaches that account for individual differences in the manifestation of COVID-19 in human voices. → **Major contribution**.

7. Conclusion

We introduced the COVYT speech dataset for the investigation of (i) the acoustic manifestation of a COVID-19 infection as well as (ii) the audio-based automatic detection of COVID-19 in free speech samples. The dataset contains 8+ hours of publicly available audio material and, in contrast to most other datasets in this research field, it features both COVID-19 positive and negative speech samples of all 65 included speakers. In our baseline experiments, we identified three acoustic features – two related to spectral flux and one related to local shimmer – to differ between the COVID-19 positive and negative samples. Moreover, we obtained a AUC over 0.7 for the automatic classification of speech samples according to COVID-19 status by using pre-trained speech models.

The COVYT dataset together with the provided benchmarks shall boost further research in the field of speech-based COVID-19 detection while ensuring reproducibility and comparability of results. Furthermore, as the dataset contains samples of the same speakers with and without COVID-19 infection, we expect it to prove a valuable conduit for future efforts in personalisation approaches that can adapt to the characteristics of individual speakers and, thus, improve performance and reliability.

CRediT authorship contribution statement

Andreas Triantafyllopoulos: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Anastasia Semertzidou:** Conceptualization, Investigation, Data curation, Writing – review & editing. **Meishu Song:** Data curation, Writing – review & editing. **Florian B. Pokorny:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Björn W. Schuller:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Bjoern Schuller reports a relationship with audEERING GmbH that includes: employment.

Data availability

Data is publicly available.

Acknowledgements

This work has received funding from the DFG, Germany's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS) and from the EU's Horizon 2020 grant agreement No. 826506 (sustAGE).

Appendix. Partitions

See Table A.5 for number of speakers and utterances in each fold.

References

[1] K. Alyafei, R. Ahmed, F.F. Abir, M.E.H. Chowdhury, K.K. Naji, A comprehensive review of COVID-19 detection techniques: From laboratory systems to wearable devices, Comput. Biol. Med. 149 (2022) 106070.
[2] R.V. Sharan, U.R. Abeyratne, V.R. Swarnkar, P. Porter, Automatic croup diagnosis using cough sound recognition, IEEE Trans. Biomed. Eng. 66 (2019) 485–495.
[3] A. Triantafyllopoulos, M. Fendler, A. Batliner, M. Gerczuk, S. Amiriparian, T. Berghaus, B.W. Schuller, Distinguishing between pre- and post-treatment in the speech of patients with chronic obstructive pulmonary disease, in: Proc. INTERSPEECH, Incheon, South Korea, 2022, pp. 3623–3627.
[4] L. Wang, Z.Q. Lin, A. Wong, COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-Ray images, Sci. Rep. 10 (2020) 1–12.
[5] V. Shah, R. Keniya, A. Shridharani, M. Punjabi, J. Shah, N. Mehendale, Diagnosis of COVID-19 using CT scan images and deep learning techniques, Emerg. Radiol. 28 (2021) 497–505.
[6] K.T. Rajamani, H. Siebert, M.P. Heinrich, Dynamic deformable attention network (DDANet) for COVID-19 lesions semantic segmentation, J. Biomed. Inform. 119 (2021) 103816.
[7] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks, Pattern Anal. Appl. 24 (2021) 1207–1220.
[8] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19, IEEE Rev. Biomed. Eng. 14 (2020) 4–15.
[9] S. Liu, J. Han, E.L. Puyal, S. Kontaxis, S. Sun, P. Locatelli, J. Dineley, F.B. Pokorny, G. Dalla Costa, L. Leocani, et al., Fitbeat: COVID-19 estimation based on wristband heart rate using a contrastive convolutional auto-encoder, Pattern Recognit. 123 (2022) 108403.
[10] P. Hecker, F.B. Pokorny, K.D. Bartl-Pokorny, U. Reichel, Z. Ren, S. Hantke, F. Eyben, D.M. Schuller, B. Arnrich, B.W. Schuller, Speaking Corona? Human and machine recognition of COVID-19 from voice, in: Proceedings INTERSPEECH, 2021, pp. 701–705.
[11] M.A. Nessiem, M.M. Mohamed, H. Coppock, A. Gaskell, B.W. Schuller, Detecting COVID-19 from breathing and coughing sounds using deep neural networks, in: Proceedings International Symposium on Computer-Based Medical Systems, IEEE, 2021, pp. 183–188.
[12] B.W. Schuller, D.M. Schuller, K. Qian, J. Liu, H. Zheng, X. Li, COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis, Front. Digit. Health 3 (2021) 14.
[13] G. Deshpande, A. Batliner, B.W. Schuller, AI-based human audio processing for COVID-19: A comprehensive overview, Pattern Recognit. 122 (2022) 108289.
[14] T. Dang, J. Han, T. Xia, D. Spathis, E. Bondareva, C. Siegele-Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, R.A. Floto, et al., Exploring longitudinal cough, breath, and voice data for COVID-19 progression prediction via sequential deep learning: Model development and validation, J. Med. Internet Res. 24 (2022) e37004.
[15] V. Despotovic, M. Ismael, M. Cornil, R.M. Call, G. Fagherazzi, Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results, Comput. Biol. Med. 138 (2021) 104944.

[16] T. Nguyen, F. Pernkopf, Lung sound classification using Co-tuning and stochastic normalization, IEEE Trans. Biomed. Eng. 69 (2022) 2872–2882.
[17] Z. Chen, M. Li, R. Wang, W. Sun, J. Liu, H. Li, T. Wang, Y. Lian, J. Zhang, X. Wang, Diagnosis of COVID-19 via acoustic analysis and artificial intelligence by monitoring breath sounds on smartphones, J. Biomed. Inform. 130 (2022) 104078.
[18] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19, IEEE Rev. Biomed. Eng. 14 (2020) 4–15.
[19] H. Coppock, L. Jones, I. Kiskin, B. Schuller, COVID-19 detection from audio: Seven grains of salt, The Lancet Dig. Health 3 (2021) e537–e538.
[20] K.D. Bartl-Pokorny, F.B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, et al., The voice of COVID-19: Acoustic correlates of infection in sustained vowels, J. Acoust. Soc. Am. 149 (2021) 4377–4383.
[21] A. Triantafyllopoulos, A. Semertzidou, M. Song, F.B. Pokorny, B.W. Schuller, COVYT: Introducing the Coronavirus YouTube speech dataset featuring the same speakers with and without infection, 2022, http://dx.doi.org/10.5281/zenodo.6962930.
[22] A. Triantafyllopoulos, A. Semertzidou, S. Meishu, F. Pokorny, B. Schuller, COVYT baseline code, 2022, http://dx.doi.org/10.5281/zenodo.7041183, URL https://github.com/ATriantafyllopoulos/covyt-baseline-code.
[23] G. Hagerer, V. Pandit, F. Eyben, B. Schuller, Enhancing lstm rnn-based speech overlap detection by artificially mixed data, in: Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio, Audio Engineering Society, 2017.
[24] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, ELAN: A professional framework for multimodality research, in: Proceedings LREC, 2006, pp. 1556–1559.
[25] K.V.S. Ritwik, S.B. Kalluri, D. Vijayasenan, COVID-19 patient detection from telephone quality speech data, 2020, arXiv preprint arXiv:2011.04299.
[26] A. Imran, I. Posokhova, H.N. Qureshi, U. Masood, M.S. Riaz, K. Ali, C.N. John, M.I. Hussain, M. Nabeel, AI4covid-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app, Inform. Med. Unlocked 20 (2020) 100378.
[27] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. Chetupalli, R. Nirmala, P. Kumar Ghosh, S. Ganapathy, Coswara-a database of breathing, cough, and voice sounds for COVID-19 diagnosis, in: Proceedings INTERSPEECH, 2020, pp. 4811–4815.
[28] A. Muguli, L. Pinto, N. R, N. Sharma, P. Krishnan, P.K. Ghosh, R. Kumar, S. Bhat, S.R. Chetupalli, S. Ganapathy, S. Ramoji, V. Nanda, DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics, in: Proceedings INTERSPEECH, 2021, pp. 901–905.
[29] L. Orlandic, T. Teijeiro, D. Atienza, The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms, Sci. Data 8 (2021) 1–10.
[30] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, R. Panicker, Cough against COVID: Evidence of COVID-19 signature in cough sounds, 2020, arXiv preprint arXiv:2009.08790.
[31] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, V. Aharonson, SARS-CoV-2 detection from voice, IEEE Open J. Eng. Med. Biol. 1 (2020) 268–274.
[32] T. Xia, D. Spathis, J. Ch, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, P. Cicuta, et al., COVID-19 sounds: A large-scale audio dataset for digital respiratory screening, in: Advances in Neural Information Processing Systems: Datasets and Benchmarks Track (Round 2), 2021.
[33] A. Triantafyllopoulos, S. Liu, B.W. Schuller, Deep speaker conditioning for speech emotion recognition, in: Proceedings ICME, Shenzhen, China, 2021, pp. 1–6.
[34] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, et al., The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2015) 190–202.
[35] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, in: Proceedings INTERSPEECH, 2013.
[36] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Adv. Neural Inf. Process. Syst. 33 (2020) 12449–12460.
[37] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, in: Proceedings INTERSPEECH, 2021, pp. 2426–2430.
[38] 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, J. Rothkrantz, J. Treep, C. Kaandorp, The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates, in: Proceedings INTERSPEECH, 2021.
[39] F. Eyben, M. Wöllmer, B. Schuller, openSMILE: The Munich versatile and fast open-source audio feature extractor, in: Proceedings ACM Multimedia, 2010, pp. 1459–1462.
[40] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, B.W. Schuller, Dawn of the transformer era in speech emotion recognition: Closing the valence gap, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[41] R. Rosenthal, Parametric measures of effect size, in: H. Cooper, L.V. Hedges (Eds.), The Handbook of Research Synthesis, Russell Sage Foundation, New York, 1994, pp. 231–244.
[42] R.J. Larsen, M. Kasimatis, Day-to-day physical symptoms: Individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses, J. Pers. 59 (1991) 387–423.
[43] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, B.W. Schuller, Probing speech emotion recognition transformers for linguistic knowledge, in: Proceedings INTERSPEECH, 2022, pp. 146–150, http://dx.doi.org/10.21437/Interspeech.2022-10371.
[44] J. Shah, Y.K. Singla, C. Chen, R.R. Shah, What all do audio transformer models hear? Probing acoustic representations for language delivery and its structure, 2021, arXiv preprint arXiv:2101.00387.