Journal of Healthcare Informatics Research (2021) 5:201–217
https://doi.org/10.1007/s41666-020-00090-4

RESEARCH ARTICLE

Automatic Detection of COVID-19 Based on Short-Duration Acoustic Smartphone Speech Analysis

Brian Stasak[1] · Zhaocheng Huang[1] · Sabah Razavi[2] · Dale Joachim[2] · Julien Epps[1]

Received: 11 November 2020 / Revised: 11 November 2020 / Accepted: 3 December 2020 /
Published online: 11 March 2021

Abstract

Currently, there is an increasing global need for COVID-19 screening to help reduce the rate of infection and at-risk patient workload at hospitals. Smartphone-based screening for COVID-19 along with other respiratory illnesses offers excellent potential due to its rapid-rollout remote platform, user convenience, symptom tracking, comparatively low cost, and prompt result processing timeframe. In particular, speech-based analysis embedded in smartphone app technology can measure physiological effects relevant to COVID-19 screening that are not yet digitally available at scale in the healthcare field. Using a selection of the Sonde Health COVID-19 2020 dataset, this study examines the speech of COVID-19-negative participants exhibiting *mild* and *moderate* COVID-19-like symptoms as well as that of COVID-19-positive participants with *mild* to *moderate* symptoms. Our study investigates the classification potential of acoustic features (e.g., glottal, prosodic, spectral) from short-duration speech segments (e.g., held vowel, pataka phrase, nasal phrase) for automatic COVID-19 classification using machine learning. Experimental results indicate that certain feature-task combinations can produce COVID-19 classification accuracy of up to 80% as compared with using the all-acoustic feature baseline (68%). Further, with brute-forced *n*-best feature selection and speech task fusion, automatic COVID-19 classification accuracy of upwards of 82–86% was achieved, depending on whether the COVID-19-negative participant had *mild* or *moderate* COVID-19-like symptom severity.

Keywords Digital medicine · Machine learning · Respiratory illness · Remote sensing

1 Introduction

In less than a year, the new virulent respiratory disease COVID-19 has quickly risen into a pandemic—with well over 28 million globally confirmed cases as of September

✉ Brian Stasak
b.stasak@unsw.edu.au

Extended author information available on the last page of the article

2020 [1]. To date, COVID-19 symptoms include fever, dry cough, fatigue, muscle aches, sore throat, diarrhea, conjunctivitis, headache, loss of taste/smell, and in more severe cases, shortness of breath, chest pain, and loss of speech or movement [2]. However, COVID-19 symptoms are also commonly found in many other types of illnesses. For instance, a cough is the most common reason for visiting a primary care physician because it is exhibited as a symptom in dozens of other types of both respiratory and non-respiratory illnesses (e.g., asthma, bronchitis, gastroesophageal reflux disease, tuberculosis) [3, 4]. Due to this symptomology overlap, new methods for specifically identifying COVID-19 versus other common illnesses are needed to help monitor and identify individuals who are potentially infected.

Over the last decade, smartphone technology has shown promise as a convenient biosensor to help track many different categories of illnesses (e.g., cardiovascular, mental, neurological, respiratory) [5–10]. Further, using audio recordings collected via smartphone devices, biomedical studies [11–15] have investigated a variety of acoustic feature types and machine learning techniques to help automatically detect respiratory illnesses. For example, glottal speech features (e.g., glottal-to-noise excitation) have been explored in respiratory disease detection studies [11] to measure differences in fundamental frequency (F0), excitation cycle, and vocal tract airflow. Suprasegmental prosodic-based speech features, such as pitch, formant frequencies, and loudness, have also been investigated for COPD/asthma-related illness detection [13, 14]. Perhaps, the most commonly used speech features for automatic respiratory disease detection are spectral (e.g., cepstral) derived from the short-term power spectrum of the speech signal [5, 14, 15].

Only very recently has audio processing via smartphone devices been proposed to assist in studying individuals who tested for COVID-19 and, further, enabled preliminary analysis of the effects that this disease has on vocal respiratory function [4, 16–20]. For example, in [20], a smartphone device app was used to collect recordings of participants' breathing and coughs for automatic COVID-19 screening. Using acoustic-based prosodic and spectral features along with a support vector machine classifier, Brown et al. [20] were able to attain an area under the receiver operating characteristic curve classification result of approximately 80% for automatically detecting COVID-19-positive and COVID-19-negative individuals. In another recent study [18], spoken sentences were automatically analyzed using acoustic-based spectral functionals to help detect three types of COVID-19-positive severities based on the number of days a patient was hospitalized. While the classification system using a support vector machine classifier in [18] produced a COVID-19-positive severity classification accuracy of 69%, these experiments omitted a healthy control or patients with other illnesses (i.e., contained only COVID-19-positive patients).

Still little is understood about the impact that COVID-19 has on healthy speech production and how changes in the voice may be used to help identify those infected [21, 22]. Although, recently in [23], it was shown that of sampled populations in the UK and USA that tested positive for COVID-19, approximately 25–32% reported having a hoarse glottal voice quality. Additionally, a recent voice pathology study conducted by [21] indicated that COVID-19-positive individuals with *mild* to *moderate* severity had abnormally high rates of vocal dysphonia (approximately 27%) likely due to glottic (e.g., vocal folds) edema and tissue inflammation. Further investigation into this new possible COVID-19 symptom is currently ongoing.

Among the aforementioned COVID-19 smartphone device acoustic recording studies [4, 16–20], most only examined recorded breathing and/or cough sounds, rather than natural speech. Therefore, these proposed systems may be limited in screening effectiveness in instances where individuals do not exhibit breathing difficulties or cough symptoms. Further, the majority of these previous studies did not evaluate individuals with *moderate* COVID-19-like symptoms that tested negative; they instead used a healthy control group with no or minimal COVID-19 symptoms to compare against a COVID-19-positive group. In a real-world application context, individuals exhibiting any COVID-19-like symptom will have careful cause for alarm; and there anon, they will likely seek further screening techniques to help decide if they require medical attention or self-quarantine.

There are several advantages to implementing remote smartphone digital health technology over traditional clinical visit screening for COVID-19. For instance, large-scale health screening implementation can be conducted via an app by simply using individuals' everyday smartphone device in non-clinical environments, such as the home, workplace, or vehicle. These remote respiratory health assessments can help to quickly log symptom location/time incidence in real-time. It has been shown that early detection of potential COVID-19 infections helps reduce the rate of transmission by alerting individuals to take more active precautions, such as limited contact with others during illness onset [24, 25].

In addition, remote evaluations can reduce overburden at healthcare emergency clinics, while serving to minimize cross infection at medical clinics, where other patients are already vulnerable. From a cost standpoint, smartphone digital health technologies can reduce healthcare-related expenses and also provide screening access to individuals living in remote places or where medical expertise might be limited [26]. Moreover, smartphone device screening has already proven useful in determining whether staff members should return to the workplace, providing a proactive approach to combating transmission of COVID-19 [27].

In this study, we investigate a new smartphone speech corpus, the Sonde Health COVID-192020 database, which contains participants with *mild* and *moderate* COVID-19-like symptoms that tested negative, along with participants with *mild* to *moderate* COVID-19 symptoms who tested positive. The experimental SHC speech corpus was collected via a digital health app using participants' personal smartphones in natural non-clinical environments. Using three distinct kinds of short-duration speech tasks, this article is the first to examine the open-source COVAREP acoustic speech feature set along with its different feature types (e.g., glottal, prosodic, spectral) for COVID-19 detection. Of particular interest are the glottal features because they have not yet been investigated in the previous automatic speech-based COVID-19 smartphone screening literature.

Automatic COVID-19 classification experiments presented herein focus on the reported accuracy of specific acoustic feature types (e.g., glottal, prosodic, spectral) and different speech tasks (e.g., held vowel, pataka phrase, nasal phrase), which have gone unexplored in previous literature [4, 16–20]. It is hypothesized that certain types of features extracted from specific speech tasks may yield more discriminative COVID-19-positive information than others. For instance, the uniformity of the held vowel task is anticipated to reveal more discriminative feature reliability on account that is an isolated single phoneme, requiring steady vocal fold activation and a degree of continued breath support. To help optimize COVID-19 identification performance, an automatic brute force *n*-best feature selection training technique is applied to help predetermine which features are most discriminative per speech task.

It is also hypothesized that by combining feature information from more than one speech task, further improvements in COVID-19 identification can be achieved. Therefore, a new task-based feature level fusion method is proposed, which utilizes decision tree classification probability outputs based on training data to evaluate combinations of different tasks and feature types and their effects on system accuracy. Preliminary experimental results herein indicate that relatively short recordings of speech (e.g., 6 s) collected from a smartphone carry sufficient discriminative information to aid in the identification of COVID-19-positive individuals, even when compared to participants that tested negative but exhibited *moderate* COVID-like symptoms. It is shown that by fusing multiple speech tasks systems that employ different feature types, further gains in automatic COVID-19 identification accuracy can be produced in comparison to using the entire baseline feature set.

2 Database

With the recent emergence of COVID-19, experimental speech recordings of individuals with positive test results are limited (i.e., relatively small number of patients, limited speech tasks). For instance, currently, there are no publicly available COVID-19 speech datasets. But there are a few notable efforts to collect larger-scale COVID-19-positive speech samples using smartphone apps [19, 28–31].

The Sonde Health COVID-19 2020 (SHC) database [31] is a proprietary speech corpus subset that consists of multiple recordings collected from participants' smartphones. All participants were recorded in a naturalistic, non-clinical environment (e.g., house, car, workplace) in the USA. The SHC database provides metadata that is unlike other previously published COVID-19 audio corpora [18–22]. For example, the SHC database contains a number of recently tested COVID-19 participants, reported COVID-19 self-questionnaire symptom scores (see Appendix, Table 5), demographic metadata (e.g., age, gender), subjective recording ratings (e.g., voice quality, noise quality), and three different speech task recordings per participant. Moreover, the SHC database supports the exploration of differences in speech characteristics between individuals who recently tested negative and those positive for COVID-19 infection.

The SHC database was reviewed and approved by Western IRB #20160262. Subjects were recruited via paid social media advertisement campaigns and then directed to download the free Sonde Health app from the App Store or Google Play. After creating a free app account, the IRB-approved consent form was displayed electronically within the app requiring subjects to indicate understanding and agreement with the research before proceeding to study activities. Subjects were not reimbursed or otherwise incentivized for study participation.

Presented in Table 1, the SHC database subset comprises (1) participants with a recent COVID-19-negative test result who exhibited *mild* COVID-19-like symptoms (Cneg group; $n = 22$); (2) participants with a recent COVID-19-negative test result who exhibited *moderate* COVID-19-like symptoms (CCneg group; $n = 22$); and (3) participants who recently tested positive for COVID-19 and exhibited *mild* to *moderate* COVID-19-like symptoms (Cpos group; $n = 22$).

Table 1 The SHC dataset contains a total of 66 participants. Each participant provided a held vowel, diadochokinetic pataka phrase, and nasal phrase recording

Participant groups	Metadata			
	Number of unique participants	*Number of female*	*Number of male*	*Age (years) mean ± std dev*
COVID-19-negative *Mild* symptoms (Cneg)	22	10	12	42.5 ± 13.5
COVID-19-negative *Moderate* symptoms (CCneg)	22	10	12	39.5 ± 12.4
COVID-19-positive *mild* to *moderate* symptoms (Cpos)	22	10	12	45.6 ± 12.7

In Fig. 1, the COVID-19 self-questionnaire individual symptom question score distributions per group were evaluated, creating the basis for symptom groupings (i.e., *mild* had total symptom scores ≤ 3, whereas *moderate* had total symptom scores ≥ 8). Specifically, the symptom score distributions for questions 1–11 and 14 were examined because these focused on key COVID-19 symptom indicators acknowledged by the WHO [2]. The COVID-19 self-questionnaire symptom score distributions reported by SHC database participants convey similar overlap in symptomology among the CCneg and Cpos groups. In Fig. 1, the CCneg group had equal or greater symptom severity scores than the Cpos group, whereas the Cneg group only demonstrated symptoms regarding nasal congestion. Per group, the COVID-19 self-questionnaire averages for these core symptoms (e.g., questions 3–11, 14) were Cneg (0.55), CCneg (19.00), and Cpos (8.23). Therefore, the CCneg group had more prevalent COVID-19-like symptoms than the actual Cpos group.

During a single-recorded session, participants were instructed via a free smartphone app (see https://www.sondehealth.com/sondeone-page) to verbalize three different unpracticed speech tasks. The recorded speech tasks were the following: (1) held vowel (e.g., take a breath and hold the vowel sound "*Ahh*," like in the word "*father*," for 6 s or until you run out of breath); (2) diadochokinetic phrase (e.g., quickly repeat the phrase "*pah-tah-kah*" as many times as you can until the timer ends); and (3) nasal phrase (e.g. , repeat "*Mama made some lemon jam*" as many times as you can until the timer ends). Once a speech task was completed and recorded, participants proceeded automatically to the next required task by the smartphone app interface.

All recordings were collected using the participants' personal smartphone device and in-built device digital recording software with a 16-bit, 44.1 kHz sampling rate (i.e., in uncompressed WAV file format). The participant audio recordings were all 6 s in length (i.e., maximum duration of task). The entire SHC dataset consisted of a total of 20 min of speech data recordings.

Illustrated in Fig. 2, acoustic spectrogram comparisons of held vowels indicated that the Cneg participants generally had consistently smoother, wider spectral energy contours over time, whereas the CCneg and Cpos participants often had more disrupted, weaker spectral energy contours in the mid to high frequencies (i.e., narrower spectral energy spread).

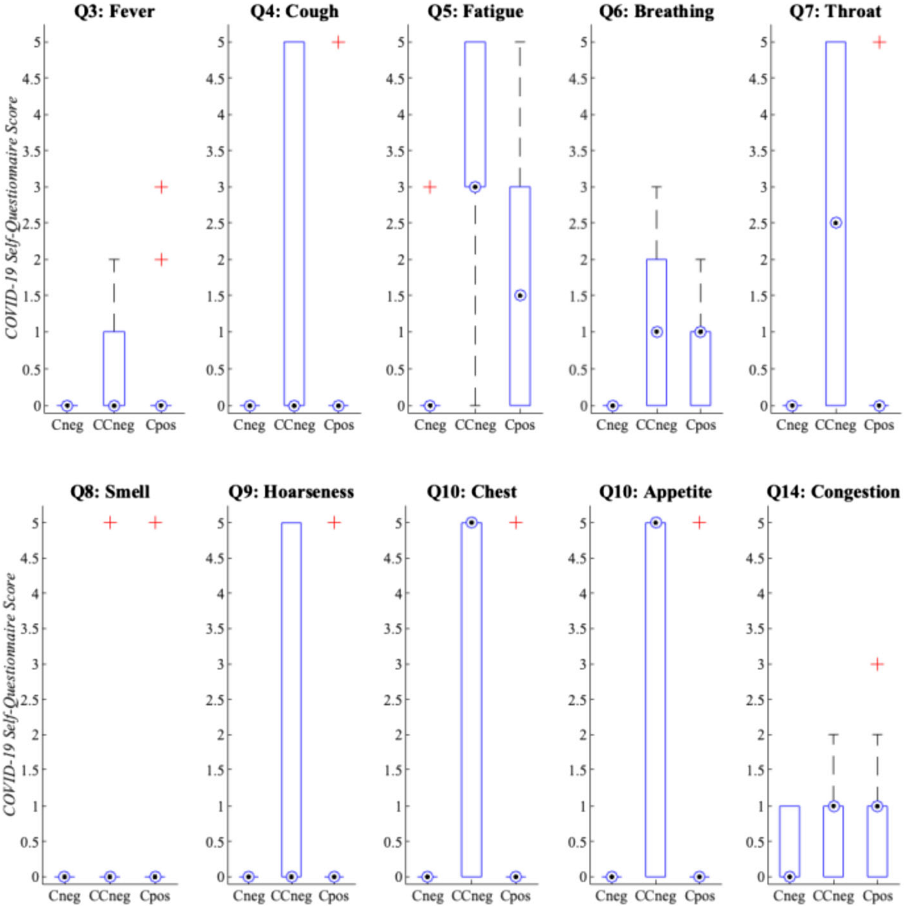

Fig. 1 SHC dataset median symptom severity COVID-19 self-assessment scores per individual question and participant group (see Appendix, Table 5). The dotted circle indicates the median, whereas the plus sign indicates extreme data outliers. In addition, the bar indicates the 25th to 75th percentiles, while the extended thin whisker line represents the remaining outer data values

3 Methods

3.1 System Configuration

The proposed system, shown in Fig. 3, begins with collecting voice samples from specific elicitation speech tasks. Individual systems and their fusion are investigated. From each speech recording, acoustic features are extracted at frame-level and aggregated over the whole speech recording. This process is followed by optional feature selection, which can help identify informative features and reduce overfitting. The features are then computed into a classifier for COVID-19 binary classification. Each individual system allows an investigation of a specific acoustic feature type (e.g., glottal, prosodic, spectral) derived from a particular speech task (e.g., held vowel, pataka phrase, nasal phrase). This is helpful for identifying reliable feature types and/or effective tasks for better discerning COVID-19 speech characteristics.

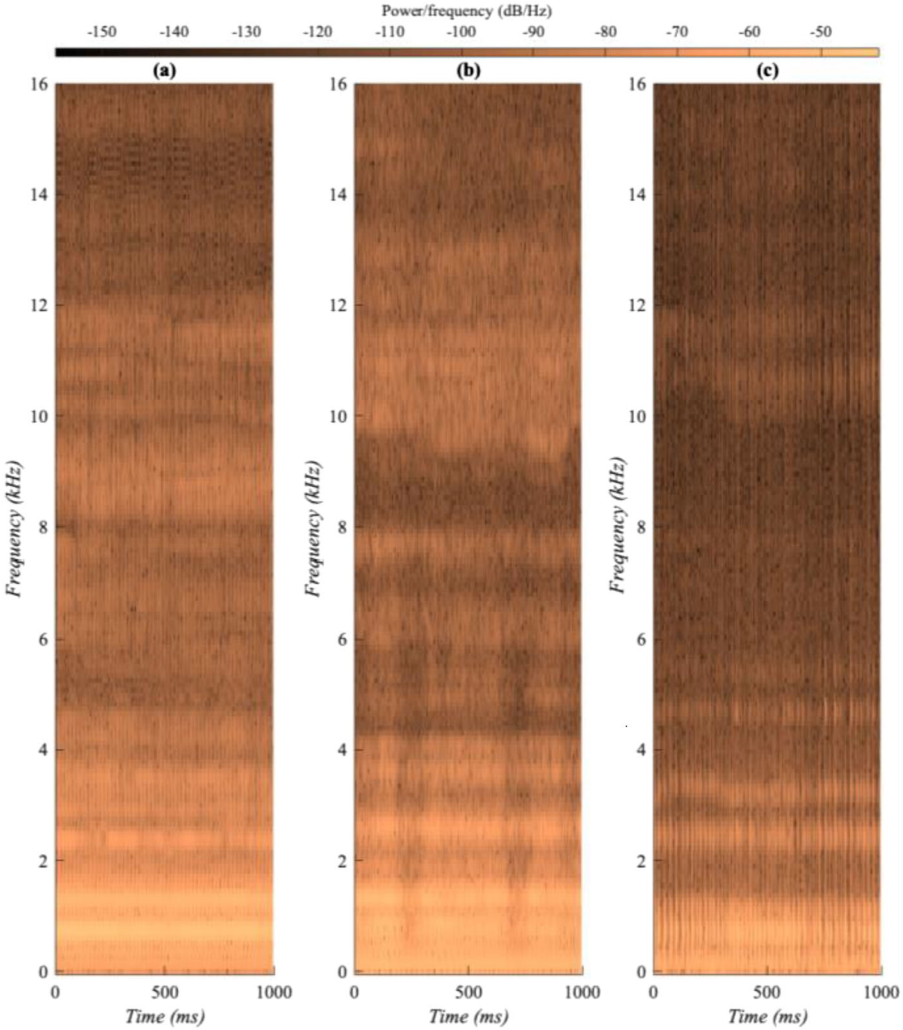

Fig. 2 SHC dataset spectrograms of three similarly aged female participants held vowel task recordings: (**a**) Cneg, (**b**) CCneg, and (**c**) Cpos. To observe the short-term temporal variation, only a 1-s segment of each recording is shown with a 0-16 kHz frequency range. For the particular spectrogram examples given above, there were relatively large differences in participants' fundamental frequency (F0): Cneg, 199 Hz; CCneg, 155 Hz; and Cpos, 63 Hz. A typical healthy adult female has a F0 of approximately 200 Hz

Fusion of the individual systems was also investigated (i.e., Fig. 3 lower section). For this process, each individual system trains a first-stage classifier to produce a probability of COVID-19-positive. Thereafter, probabilities of all individual systems are concatenated as features and computed into a second-stage classifer for COVID-19 classification. This fusion method can be more effective than direct concatenation of features and/or tasks within individual systems because the latter results in higher dimensionality and, hence, often overfits the training data.

As previously indicated in Table 1, the number of SHC database participants is relatively small ($n = 66$). Consequently, there is a high risk of overfitting if

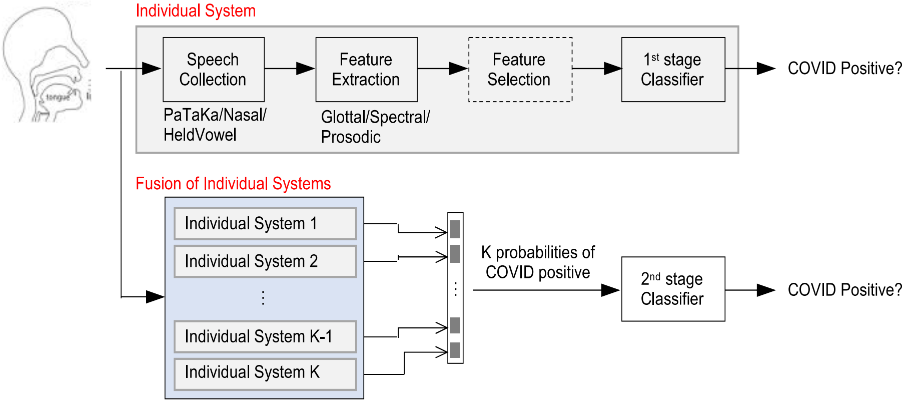

Fig. 3 Experimental design showing both individual systems and their fusion for COVID-19 binary classification (e.g., negative vs. positive). The proposed experimental feature selection method explored in this study herein is indicated by dashed lines

the feature dimensionality is too large. For example, if there are over one-hundred features, but only half as many participants, a machine learning system requires feature selection to alleviate potential overfitting. Therefore, our proposed system employed an automatic feature selection method that applied a brute-forced search of n-best features ($n = 5$) among up to 500 task-feature combinations. This method randomly sampled 500 combinations and evaluated these combinations in a leave-one-speaker-out cross validation manner. Thus, if the total number of possible features was less than 500, all combinations were explored.

3.2 COVAREP Feature Set

Since the COVID-19 illness is relatively new, its symptoms associated with abnormal vocal function are less known [21, 22]. Therefore, diverse acoustic feature–type groups were explored (e.g., glottal, prosodic, spectral) using the COVAREP speech toolkit [32]. Previously, COVAREP features have been utilized to investigate and automatically recognize voice quality [33, 34], respiratory [35], voice [36, 37], and psychogenic disorders [38, 39]. The COVAREP feature set includes 73 individual glottal, prosodic, and spectral features. All COVAREP features were calculated by extracting acoustic speech features from 20-ms frames with 50% frame shift overlap, wherein an aggregated mean and standard deviation functionals were computed over the entire recording per feature (i.e., 73 mean features + 73 standard deviation features = 146 features).

The standard COVAREP binary feature voice activity was used to indicate which frames contained speech; and therefore, only these speech frames were included for acoustic feature functional calculations. In addition, the number of indexed frames (total duration), silence frames (i.e., unvoiced frames, no pitch), speech frames (i.e., voiced frames, pitch), and percentage speech were also calculated as prosodic features. In total, there were 150 COVAREP features for experimentation.

3.3 Experimental Settings and Performance Metrics

Similarly to other speech-based respiratory disorder identification studies [40, 41], and due to the equal number of participants within each group (Cneg, CCneg, Cpos), COVID-19 classification accuracy is reported herein. Accuracy was computed per individual leave-one-speaker-out crossfold validation experiment, with all folds subsequently averaged. Therefore, during each fold experiment, no participant was duplicated in both training and test. Additionally, by implementing leave-one-speaker-out train and testing, the amount of training data available for class modeling was maximized (i.e., 43 speakers training and 1 speaker test per fold). A decision tree [42] backend by machine learning classifier with maximum of four splits was applied due to its simple interpretability, robustness to overfitting, and application in prior speech-based paralinguistic detection studies [43, 44].

Fusion of individual systems was also conducted using the same folds as in individual systems. Within each fold, the first-stage decision tree classifier was trained to produce probabilities on the training and test data, which is repeated for K different individual systems (e.g., with different feature types or elicitation tasks). The K probabilities were concatenated and applied as features to train a second-stage decision tree classifier; therefore, producing a final binary decision of the test participant within the fold. This process was repeated for all the folds, and led to binary predictions for all participants, one from each fold. Afterwards, accuracy between the binary predictions and ground truth was reported.

A randomized shuffle of features was further explored to help validate COVID-19 binary classification result accuracy gains presented herein (i.e., to indicate that the gain in accuracy was not due to random occurrence). Given an experimental dataset, $X \in \mathbb{R}^{N \times K}$, including N speakers and K features, the columns and rows were randomly shuffled to obtain $\widetilde{X} \in \mathbb{R}^{N \times K}$.

4 Results and Discussion

4.1 Speech Task and Feature Type

Experiments investigating the effect of different speech tasks and feature types were conducted. As shown in Table 2, using the entire COVAREP feature set, the nasal phrase task for the Cneg vs. Cpos experiments produced the highest COVID-19 classification baseline accuracy of 68%. The nasal phrase task contains more variety in phonemes (e.g. *m*, *ah*, *ay*, *d*, *s*, *uh*, *l*, *eh*, *n*, *j*, *aa*) than the other tasks, including many repetitive instances of nasal phonemes with nasal cavity resonance. Considering that the Cpos participants exhibited a greater level of congestion when compared with other groups (refer to Fig. 1), this symptom possibly narrowed their spectral energy spread and contributed to this higher accuracy.

In Table 2, an individual examination of COVAREP feature types for Cneg vs. Cpos classification demonstrated that prosodic features derived from the held vowel task produced the best COVID-19 classification accuracy (80%). The held vowel task elicits an isolated example of continuous vocal fold function, allowing any abnormal vocal characteristics to be easily captured. Thus, any abnormal shift in normal phonation ability caused by respiratory illness symptoms, such as respiratory muscle weakness or poorer voice quality, is more easily revealed during the held vowel task.

The high performance of the held vowel prosodic features is on account of the generally lower fundamental frequencies (F0) and instability found in the Cpos group when

Table 2 COVID-19 classification accuracy results using a decision tree classifier with leave-one-speaker-out cross validation

Cneg vs. Cpos	**Speech Tasks**		
Feature Types	**Vowel**	**Pataka**	**Nasal**
All COVAREP (150)	55%	32%	**68%**
Glottal (18)	30%	**68%**	32%
Prosodic (6)	**80%**	52%	52%
Spectral (126)	57%	52%	66%
CCneg vs. Cpos	**Speech Tasks**		
Feature Types	**Vowel**	**Pataka**	**Nasal**
All COVAREP (150)	52%	50%	27%
Glottal (18)	**71%**	52%	50%
Prosodic (6)	57%	39%	46%
Spectral (126)	50%	48%	39%

The total number of features per feature type is shown in parenthesis

compared with the Cneg group. The pataka task glottal features (68%) and nasal phrase spectral features (66%) produced relatively moderate COVID-19 classification accuracies. Further, in comparing the nasal phrase results of the entire COVAREP feature set accuracy (68%) to its individual feature type accuracies, is it evident that the spectral feature type (66%) contributed the most discriminative COVID-19 information.

Also shown in Table 2, with the exception of the held vowel glottal features (71%), the CCneg vs. Cpos experiments generally produced around chance-level accuracy. As somewhat expected, the CCneg vs. Cpos feature type classification experiments produced low accuracy because participants in both of these groups had more overlap in COVID-19-like symptoms and were, therefore, often more confusable. Nonetheless, the 71% classification accuracy result produced by the held vowel task glottal features indicates that there are glottal characteristic differences between Cneg and CCneg groups, both of which have COVID-19-like symptom overlap. Furthermore, investigation of results shown in Table 2 using the randomized shuffle of features showed that while Cneg vs. Cpos specific feature type accuracy results in general were only slightly higher than randomized feature results (~ 2%), the CCneg vs. Cpos specific feature type accuracy results were up to 12% (absolute) higher.

4.2 Automatic *n*-Best Feature Selection

To boost speech-based COVID-19 classification performance, an automatic brute-forced *n*-best feature selection approach was evaluated. During experimentation, a variety of *n* value parameter settings were explored. For both the Cneg vs. Cpos and CCneg vs. Cpos experiments and task-feature combinations, $n \leq 5$ value produced the best COVID-19 classification accuracy results. As shown in Table 3, the Cneg vs. Cpos experiments when compared against the entire COVAREP features set baseline, the *n*-best feature selection technique produced considerably higher COVID-19 classification accuracies, especially for the spectral features (82–86%). Also, for the Cneg vs. Cpos experiments using *n*-best feature selection, the held vowel task demonstrated the most consistent classification accuracy performance for all three feature types (80–82%).

Table 3 COVID-19 classification accuracy results using automatic brute-forced n-best feature selection and decision tree classifier with leave-one-speaker-out cross validation

Cneg vs. Cpos	**Speech tasks**		
Feature types	**Vowel**	**Pataka**	**Nasal**
Glottal n-Best (5)	80%	**82%**	75%
Prosodic n-Best (5)	**80%**	66%	61%
Spectral n-Best (5)	**82%**	**82%**	**86%**
CCneg vs. Cpos	**Speech tasks**		
Feature types	**Vowel**	**Pataka**	**Nasal**
Glottal n-best (5)	75%	73%	**80%**
Prosodic n-best (5)	61%	64%	66%
Spectral n-best (5)	**80%**	**82%**	77%

The total number of features per feature type is shown in parenthesis

As for the CCneg vs. Cpos classification experiments, shown previously in Table 3, by implementing the n-best feature selection, the spectral features generally produced the highest COVID-19 classification results over all three speech tasks (77–82%). In addition, the nasal phrase glottal n-best features produced a high COVID-19 classification accuracy (80%). Results in Table 3 indicate that, in addition to isolating different feature types per task, for this small dataset, further COVID-19 classification improvements can be made by reducing the number of features within feature types and optimizing key features that contain the most class discriminative power.

4.3 Automatic Task-Feature Fusion

It was hypothesized earlier that by combining acoustic feature information from more than one speech task, greater improvements in COVID-19 classification could be achieved. Results using the proposed task-feature fusion system (described previously in Section III. A and shown in Fig. 3) are shown in Table 4. Although an effort to boost COVID-19 classification performance was attempted by task-feature fusion, generally these accuracies could not surpass those just using the n-best feature selection method.

Among the best task-feature combination were the spectral features using all three tasks (80–84%). The held vowel using all three feature types in comparison to other speech tasks with all three feature types produced the highest overall accuracies (75–82%). Also, it is interesting to note that the combination of all speech tasks using glottal and spectral n-best features produced higher accuracy for CCneg vs. Cpos (80%) than Cneg vs. Cpos (73%). A similar trend also occurred during the held vowel task using all three feature types (82% and 75%, respectively).

In Table 4, an additional task-feature fusion classification experiment was investigated, Cneg vs. CCneg, which demonstrated that participants with *mild* and *moderate* COVID-19-like symptoms can be recognized from each other to a high degree (84%). These findings have implications in the wider field of smartphone speech-based respiratory illness identification, as they show that individuals with other possible respiratory illnesses can be identified via a simple speech task.

A further examination of results shown in Table 4 using the randomized shuffle of features indicated that the COVID-19 classification accuracies using all feature types

Table 4 COVID-19 classification accuracy results using task-based fusion with automatic brute-forced *n*-best feature selection ($n = 5$) with leave-one-speaker-out cross validation. Per row, results are shown for various task/feature combinations (indicated by dots). The selected tasks are indicated by dots for the held vowel (H), pataka phrase (P), and nasal phrase (N) abbreviations

Glottal			Prosodic			Spectral			*CCneg* vs. *Cpos*	*Cneg* vs. *CCneg*	*Cneg* vs. *Cpos*
H	*P*	*N*	*H*	*P*	*N*	*H*	*P*	*N*			
•	•	•							75%	77%	**84%**
			•	•	•				80%	57%	64%
						•	•	•	**84%***	80%	82%
•			•			•			75%	**82%**	82%
	•			•			•		77%	73%	82%
		•			•			•	80%	71%	75%
•	•	•	•	•	•				75%	66%	71%
•	•	•				•	•	•	73%	80%	77%
			•	•	•	•	•	•	77%	71%	**84%**
•	•	•	•	•	•	•	•	•	75%	77%	77%

*Produced measures of 0.91 *sensitivity* and 0.77 *specificity*

(e.g., glottal, prosodic, spectral) for pataka or nasal phrases produced an average of 5% absolute higher accuracy than randomized shuffle feature results.

5 Conclusion

With the global rapid spread of COVID-19, new techniques for quickly screening for this highly contagious disease are needed. In this study, we proposed using speech data collected via smartphone recordings to help identify COVID-19-positive individuals. When analyzing results based on subset feature types (e.g., glottal, prosodic, spectral) rather than the entire set, our COVID-19 classification accuracy results showed that the held vowel task outperformed the pataka and nasal phrase task. For instance, up to 14% absolute accuracy gain for COVID-19-negative participants with *mild* symptoms (Cneg) and up to 21% absolute accuracy gain for COVID-19-negative participants with *moderate* symptoms (CCneg) were recorded.

According to aforementioned automatic speech-based COVID-19 identification literature, glottal features have not yet been explored. Results herein demonstrated that glottal features perform well in instances where there is a high degree of COVID-19-like symptomology overlap in both negative and positive individuals. For example, the COVAREP glottal feature type for CCneg vs. Cpos generated an accuracy of 71% higher than the prosodic and spectral feature types (see Table 2).

The proposed brute-forced *n*-best feature selection method generated the best COVID-19 accuracies due to its ability to choose the most optimal features per task and feature type. While an examination of task-feature fusion did not produce further gains in COVID-19 accuracy over the non-fused brute-forced *n*-best approach, it did however provide further insights, such as which feature types were least affected by speech tasks and what task combinations produced consistently high performance.

These preliminary COVID-19 classification results were competitive with recent COVID-19 cough/breathing identification studies [4, 20]. However, unlike [4, 20], SHC audio recordings were collected less invasively (i.e., collected over short chosen period, chosen location/time, under naturalistic conditions) and speech data was automatically analyzed using only 6 s per recording (i.e., did not require searching hours of speech data for particular speech sound segment). Additionally, one of the experimental COVID-19 negative groups evaluated herein had moderate COVID-19-like symptoms that subjectively seemed to have a stronger effect on the speech produced than those of the actual COVID-19-positive group.

6 Future Work

Future studies using speech-based smartphone data collected from embedded apps can become a convenient tool in helping combat the spread of COVID-19, along with other common respiratory illnesses. More research concerning the effects of COVID-19 on speech behaviors, both during its acute phase and post-recovery phase, is desirable because it will lead to a better understanding of this new disease and more effective automated speech-based smartphone screening/monitoring applications.

Acknowledgments The authors would like to thank Sonde Health for collecting and providing the SHC dataset. The work of Brian Stasak, Zhaocheng Huang, and Julien Epps is partly supported by ARC Linkage Project (LP160101360) and by Data61, CSIRO.

Appendix

Table 5 Example of the Center for Disease Control informed self-questionnaire completed by each participant in the SHC dataset

Q1	*Have you ever been tested for COVID-19?*
	(0) No, I have never had a COVID test
	(1) Yes, my test result was positive
	(2) Yes, my result was negative
	(3) Yes, I am waiting for my test result
	(4) Yes, my test was inconclusive
Q2	*If you have been tested for COVID-19, when was your test performed?*
	(0) I have never had a COVID test
	(1) I was tested today
	(2) I was tested 1–2 days ago
	(3) I was tested 3–6 days ago
	(4) I was tested 1–2 weeks ago
	(5) I was tested more than 2 weeks ago
Q3	*Do you have a fever or feel too hot?*

Table 5 (continued)

	(0) No
	(1) Yes, I have not measured my temperature
	(2) Yes, less than 100F
	(3) Yes, between 100 and 102 F
Q4	*Do you have a persistent cough (coughing a lot for more than an hour, or three or more coughing episodes in 24 h)?*
	(0) No
	(5) Yes
Q5	*Are you experiencing unusual fatigue?*
	(0) No
	(1) Mild fatigue
	(2) Severe fatigue—I struggle to get out of bed
Q6	*Are you experiencing unusual shortness of breath or have trouble breathing?*
	(0) No
	(1) Yes, mild symptoms—slight shortness of breath during ordinary activities
	(2) Yes, significant symptoms—breathing is comfortable only at rest
Q7	*Do you have a sore or painful throat?*
	(0) No
	(5) Yes
Q8	*Do you have loss of smell or taste?*
	(0) No
	(5) Yes
Q9	*Do you have an unusually hoarse voice?*
	(0) No
	(5) Yes
Q10	*Are you feeling an unusual chest pain or tightness in your chest?*
	(0) No
	(5) Yes
Q11	*Have you been skipping meals?*
	(0) No
	(5) Yes
Q12	*Please select if a doctor has ever diagnosed you with any of the following conditions?*
	(0) Asthma
	(1) Chronic obstructive pulmonary disease (COPD)
	(2) Congestive heart failure (CHF)
	(3) Other condition that may make it difficult to breathe
	(4) None of the above
Q13	*Do you smoke tabacco?*
	(0) Yes, everyday
	(1) Yes, some days
	(2) Not currently, but I did in the past
	(3) Never smoked
Q14	*How congested is your nose right now?*
	(0) Not congested at all

Table 5 (continued)

(1) Mild congestion
(2) Moderate congestion
(3) Severe congestion

These questions are based on several key symptoms reported by many individuals infected with COVID-19 illness

References

1. John Hopkins University and Medicine (2020) COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE). https://coronavirus.jhu.edu/map.html. Accessed September 2020
2. World Health Organization (WHO) (2020) Coronavirus Symptoms. https://www.who.int/health-topics/coronavirus#tab=tab_3. Accessed September 2020
3. Goldsobel AB, Chipps BE (2010) Cough in the pediatric population. J Pediatr 156(3):352–358
4. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz S, Ali K, et al. (2020) AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app, Informatic in medicine unlocked
5. Batra S, Baker RA, Wang T, Forma F, DiBiasi F, Peters-Strickland T (2017) Digital health technology for use in patients with serious mental illness: a systematic review of the literature. Med Devices 10:237–251
6. Bauer M, Glenn T, Geddes J, Gitlin M, Grof P, Kessing LV et al (2020) Smartphones in mental health: a critical review of background issues, current status and future concerns. Int J Bipolar Disord 8(2):1–19
7. Majumder S, Deen MJ (2019) Smartphone sensors for health monitoring and diagnosis. Sensors 19(9):1–45
8. Mosa ASM, Yoo I, Sheets L (2012) A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Making 12(67):1–31
9. Tabatabaei SAH, Fischer P, Schneider H, Koehler U, Gross V, Sohrabi K (2020) Methods for adventitious respiratory sound analyzing applications based on smartphones: a survey. IEEE reviews in biomedical engineering.
10. Yang X, Kovarik CL (2019) A systematic review of mobile health interventions in China: identifying gaps in care. J Telemed Telecare 0(0):1–20
11. Petrizzo D, Popolo PS (2020) Smartphone use in clinical voice recording and acoustic analysis: a literature review. J Voice 28;S0892–1997(19)30284-X. https://doi.org/10.1016/j.jvoice.2019.10.006
12. Merkus J, Hubers F, Cucchiarini C, Strik H (2019) Digital eavesdropper – acoustic speech characteristics as markers of exacerbations in COPD patients. In: Proc. Lang. Res. and Eval. Conf. (LREC), pp. 78–86
13. Nathan V, Rahman M, Vatanparvar K, Nemati E, Blackstock E, Kuang J (2019) Extraction of voice parameters from continuous running speech for pulmonary disease monitoring. In: Proc. IEEE Intern. Conf. on Bioinformatics and Biomedicine (BIBM), pp. 859–864
14. Song I (2015) Diagnosis of pneumonia from sounds collecte using low cost cell phones. In: Proc. Intern. Joint Conf. on Neural Networks (IJCNN), Killarney - Ireland, pp 1–8
15. Rudraraju G, Palreddy S, Mamidgi B, Sripada NR, Sai YP, Vodnala NK, Haranath SP (2020) Cough sound analysis and objective correlation with spirometry and clinical diagnosis. Informatics Med Unlocked 19:1–11
16. Deshpande G, Schuller BW (2020) An overview on audio, signal, speech, & language processing for COVID-19. arXiv:2005.08579v1, pp. 1–5
17. Faezipour M, Abuzneid A (2020) Smartphone-based self-testing of COVID-19 using breathing sounds. Telemed J E Health 26(10):1202–1205. https://doi.org/10.1089/tmj.2020.0114
18. Han J, Qian K, Song M, Yang Z, Ren Z, Liu S, et al (2020) An early study on intelligent analysis of speech under COVID-19: severity, sleep quality, fatigue, and anxiety. arXiv:2005.00096v2, pp. 1–5
19. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, Nirmala R, et al. (2020) Coswara – a database of breathing, cough, and voice sounds for COVID-19 diagnosis. arXiv:2005.10548v1, pp. 1–5

20. Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, Xia T, Cicuta P, Mascolo C (2020) Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In: Proc. KDD '20 (Health Day), San Diego, CA – USA
21. Lechien JR, Chiesa-Estomba CM, Cabaraux P, Mat Q, Huet K, Harmegnies B et al (2020) Features of mild-to-moderate COVID-19 patients with dysphonia. J Voice
22. Quatieri TF, Talkar T, Palmer JS (2020) A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. IEEE Open J Eng Med Biol 1:203–206
23. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, Ganesh S, Varsavsky T, Cardoso MJ, el-Sayed Moustafa JS, Visconti A, Hysi P, Bowyer RCE, Mangino M, Falchi M, Wolf J, Ourselin S, Chan AT, Steves CJ, Spector TD (2020) Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 26:1037–1040
24. Salathé M, Althaus C, Neher R, Stringhini S, Hodcroft E, Fellay J et al (2020) COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. Swiss Med Wkly 150:1–3
25. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC (2020) Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19). JAMA 324(8):782–793
26. Goodridge D, Marciniuk D (2016) Rural and remote care: overcoming the challenges of distance. Chron Respir Dis 13(2):192–203
27. Vo A, Brooks GB, Farr R, Raimer B (2011) Benefits of telemedicine in remote communities & use of mobile and wireless platforms in healthcare. UTMB Telemedicine and Center for Telehealth Research and Policy, pp. 1–9
28. Subirana B, Hueto F, Rajasekara P, Laguarta J, Puig S, Malvehy J, et al (2020) Hi Sigma, do I have the coronavirus? MIT Auto-ID Lab Report, pp. 1–13
29. Carnagie Mellow University (CMU), COVID voice detector, Pittsburgh, PA – USA, https://cvd.lti.cmu.edu. Accessed September 2020
30. University of Cambridge (UC), COVID-19 sounds app, Cambridge – UK, https://www.covid-19-sounds.org/en/. Accessed September 2020
31. Sonde Health (SH), Sonde One, Boston, MA – USA, https://www.sondehealth.com/sondeone-page. Accessed September 2020
32. Degottex G, Kane J, Drugman T, Raitio T, Scherer S (2014) COVAREP – a cooperative voice analysis repository for speech technologies. In: Proc. of Acoustics, Speech and Signal Processing (ICASSP), pp 960–964
33. Borsky M, Mehta DD, Van Stan J, Gu∂nason J (2017) Modal and nonmodal voice quality classification using acoustic and electroglottographic features. IEEE/ACM Trans Audio Speech Lang Process 25(12): 2281–2291
34. Villegas J, Markov K, Perkins J, Lee SJ (2020) Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness. IEEE J Select Top Signal Proc 14(2):355–366
35. Szklanny K, Gubrynowicz R, Tylki-Szymansak A (2018) Voice alterations in patients with Morquio A syndrome. J Appl Genet 59:73–80
36. Amara F, Fezari M (2015) Laryngeal pathologies analysis using glottal source features. In: Proc. Intern. Conf. on Automatic Control, Telecomm. and Signals (ICATS15), Annaba – Algeria, pp. 1–6
37. Szklanny K, Gubrynowicz R, Iwanicka-Pronicka K, Tylki-Szymańska A (2016) Analysis of voice quality in patients with late-onset Pompe disease. Orphanet J Rare Dis 11(1):1–9
38. Stasak B, Epps J, Goecke R (2017) Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect", In: Proc. INTERSPEECH, Stockholm – Sweden, pp. 834–838
39. Vâzquez-Romero A, Gallardo-Antolin A (2020) Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy 22(688):1–17
40. Cummins N, Schmitt M, Amiriparian S, Krajewski J, Schuller B (2017) You sound ill, take the day off": automatic recognition of speech affected by upper respiratory tract infection", 2017 38th Ann. Intern. Conf. of the IEEE Eng. in Med. And Biology Society (EMBC), pp. 3806–3809
41. Pozo F, Murillo RB, Hernández-Gómez JL, López-Gonzalo E, Alcázar-Ramírez J, Toledano DT (2009) Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. EURASIP J Adv Signal Proc 2009:1–11. https://doi.org/10.1155/2009/982531
42. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall, Boca Raton
43. Yang L, Jiang D, He L, Pei E, Oveneke MC, Sahli H (2016) Decision tree based depression classification from audio video and language information. In; Proc. 6th Int. Work. Audio/Visual Emot. Chall. (AVEC '16), pp. 89–96

44. Yüncü E, Hacihabiboglu H, Bozsahin C (2014) Automatic speech emotion recognition using auditory models with binary decision tree and svm. In: Proceedings of the 2014 22nd Intern. Conf. on Pattern Recognition, Stockholm - Sweden, 2014, pp. 773–778

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Brian Stasak[1] · **Zhaocheng Huang**[1] · **Sabah Razavi**[2] · **Dale Joachim**[2] · **Julien Epps**[1]

Zhaocheng Huang
zhaocheng.huang@unsw.edu.au

Sabah Razavi
srazavi@sondehealth.com

Dale Joachim
djoachim@sondehealth.com

Julien Epps
j.epps@unsw.edu.au

[1] School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, NSW, Australia

[2] Sonde Health, Boston, MA, USA