COVID-19 Detection via Fusion of Modulation Spectrum and Linear Prediction Speech Features

Yi Zhu, *Student Member, IEEE*, Abhishek Tiwari, João Monteiro, Shruti Kshirsagar, and Tiago Henrique Falk

***Abstract*—The coronavirus disease 2019 (COVID-19) pandemic has drastically impacted life around the globe. As life returns to pre-pandemic routines, COVID-19 testing has become a key component, assuring that travellers and citizens are free from the disease. Conventional tests can be expensive, time-consuming (results can take up to 48 h), and require laboratory testing. Rapid antigen testing, in turn, can generate results within 15-30 minutes and can be done at home, but research shows they achieve very poor sensitivity rates. In this paper, we propose an alternative based on speech signals recorded at home with a portable device. It has been well-documented that the virus affects many of the speech production systems (e.g., lungs, larynx, and articulators). As such, we propose the use of new modulation spectral features and linear prediction analysis to characterize these changes via a two-stage classification system. Experiments on three COVID-19 speech datasets show that the proposed two-stage system outperforms several state-of-the-art benchmarks, relies on interpretable features, as well as generalizes well to unseen datasets. Overall, the proposed system shows promise as an accessible, low-cost, at-home method for COVID-19 detection.**

***Index Terms*—COVID-19 diagnostics, modulation spectrogram, linear predictive analysis, speech analysis.**

I. Introduction

AT THE time of writing, more than 230 million cases of the coronavirus 19 disease (COVID-19) had been reported worldwide, resulting in the death of 4.14 million people [1]. As vaccine rates are increasing around the world, travel is slowly returning to pre-pandemic rates and more establishments are opening their doors. An enabling factor for this has been wide-scale testing. For example, travellers arriving in countries are required to take tests pre and post arrival; students back to campus in-person that are not fully vaccinated are required to show negative tests on a weekly basis. Conventional testing, such as polymerase chain reaction (PCR) tests, however, are expensive and time-consuming, as well as arduous and extremely uncomfortable to the patient [2]. Rapid antigen tests, in turn, are emerging and can provide answers within 15-30 minutes and can be done at home. Research, however, has shown that such tests can have very poor sensitivity rates [3], [4], [5]. The false negative rate is estimated to be around 67% within the first 4-5 days following the onset of symptoms [6]. As such, alternate solutions are still needed.

It has been established that the COVID-19 virus can affect many of the systems involved with speech production [7]. For example, the virus affects the respiratory system, targeting the lungs and other airway passages, thus resulting in shortness of breath, coughs, and atypical breathing modulations, to name a few symptoms [8]. These, in turn, can irritate the vocal cords, resulting in inflammation, sore throats, hoarseness, and breathiness. Moreover, temporary neuromuscular deficits have been reported [9], thus also affecting speech articulators, causing atypical changes in the acoustic properties of the produced speech signal [7]. These factors suggest that automated analysis of speech signals for COVID-19 detection can be possible. This is the aim of the present study.

Existing speech-based COVID-19 detection systems have focused mainly on two aspects: i) the design of acoustic features and ii) data-driven machine learning algorithms that find non-linear (discriminatory) patterns in the acoustic signals. Regarding feature engineering, one of the most widely used feature sets explored for COVID-19 monitoring has been the so-called ComParE (Computational Paralinguistics ChallengE) acoustic feature set 2016 [10]. This ComParE set contains over 6,000 features that cover different speech signal representations, such as mel-frequency cepstral coefficients (MFCC), pitch contours, voicing-related information, as well as several other low-level descriptors (LLDs). In the recent INTERPSEECH 2021 ComParE COVID-19 speech sub-task, the ComParE set 2016 was used as one of the benchmark features [11].

With these features, Jing et al. showed that it could be possible to predict the severity of the disease, patient anxiety, sleep quality, and fatigue [12]. Alternatively, as mel-spectrograms can be seen as an image representation of the speech signals, several attempts have been made to explore their use with convolutional neural networks (CNNs) [13], [14], [15], [16]. In fact, the majority of the existing speech-based systems have

Manuscript received 30 September 2021; revised 23 September 2022; accepted 28 March 2023. Date of publication 7 April 2023; date of current version 21 April 2023. This work was supported by Institut national de la recherche scientifique (INRS) under Grant INRS-Covid19-2020-015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mathew Magimai Doss. *(Corresponding author: Yi Zhu.)*

Yi Zhu and Tiago Henrique Falk are with the Institut national de la recherche scientifique, INRS- EMT, University of Québec, Québec, QC G1K 9H7, Canada (e-mail: yi.zhu@inrs.ca; tiago.falk@inrs.ca).

Abhishek Tiwari is with the Myant Inc, Etobicoke, ON M9W 5Z9, Canada (e-mail: abhi.tiw1@gmail.com).

João Monteiro is with the ServiceNow Research, Montréal, QC H3C 2G9, Canada (e-mail: joaomonteirof@gmail.com).

Shruti Kshirsagar is with the EERS Global Technologies Inc., Montréal, QC H3C 2G9, Canada, and also with the Institut national de la recherche scientifique, INRS- EMT, University of Québec, Québec, QC G1K 9H7, Canada (e-mail: shruti.kshirsagar@inrs.ca).

Digital Object Identifier 10.1109/TASLP.2023.3265603

2329-9290 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

focused on the use of deep neural networks (DNNs) for the task at hand (e.g., [14], [15], [16]). Notwithstanding, the COVID-19 ComParE Challenge showed that support vector machines (SVM) could still outperform DNNs [11].

Ultimately, we are interested in a system that is i) accessible to all, i.e., does not require very complex models that cannot be run on legacy devices; ii) interpretable, hence the input features, as well as the predictions made by the system, can be explained and understood; and iii) generalizable to unseen conditions, such as speakers, languages, microphones, etc. To achieve these goals, we first designed a new feature set based on modulation spectral changes of the speech signal. Previous work has shown that the modulation spectral representation can accurately separate signal from noise [17], can characterize changes in human mental/emotional states in-the-wild [18], can characterize changes in speech production due to disease [19], as well as serve as improved input to DNNs, thus resulting in smaller, less complex models [20]. As such, modulation spectral features are expected to play a crucial role in robust, low-overhead diagnostics. Moreover, to further provide greater interpretability to the system, we propose to decompose the signal into excitation and vocal tract system components via linear prediction (LP) analysis [21]. A two-stage feature fusion system is then designed based on the observed complementarity between the two proposed feature sets, thus allowing for factors affecting both the articulators and the phonation systems to be characterized. Experimental results show the proposed system outperforming several DNN-based benchmarks across three different datasets, as well as achieving the highest cross-dataset accuracy, thus suggesting increased generalizability and interpretability of the system for COVID-19 detection.

The remainder of this paper is organized as follows. Section II describes the proposed features and their use in COVID-19 detection. Section III then describes the experimental setup, while Section IV describes and discusses the obtained results. Lastly, Section V presents the conclusions.

II. Proposed COVID-19 Detection System

In this section, we describe the proposed system and motivate the use of the modulation spectrum and LP features.

A. Modulation Spectral Features

The commonly-used spectrogram provides information on how frequency components change as a function of time. However, common noise sources overlap in both time and frequency, thus making the representation sub-optimal for classification tasks relying on noisy data. The modulation spectrum, in turn, characterizes the rate-of-change of these frequency components, thus better separating signal and noise components, making it a better candidate for in-the-wild applications [20], [22]. The modulation spectrum is also known to capture higher-order periodicities of the signal not otherwise obvious in time and time-frequency domains. This property has shown to be useful for disease characterization, such as autism spectrum disorder detection from toddler cries and non-verbal vocalizations [23], as well as automated intelligibility monitoring of dysarthric speech [19]. We explore it here as a useful, robust representation for COVID-19 detection.

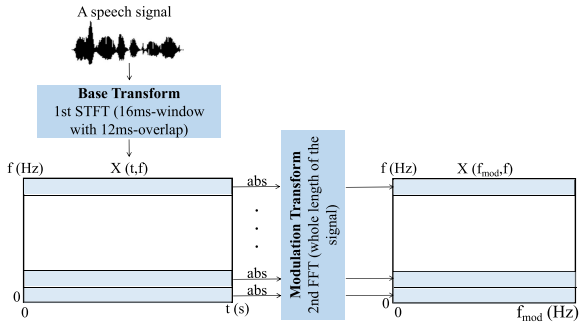


Fig. 1. Signal processing steps involved in the computation of the modulation spectrogram.

The signal processing steps involved in the computation of the modulation spectrogram are depicted in Fig. 1. First, the speech signal $x(t)$ is transformed to the time-frequency domain (spectrogram) via the short-time Fourier transform (STFT) (here, implemented using a 256-point FFT). A second transform is then applied across the time axis, for each frequency bin magnitude $|X(t, f)|$. This results in a frequency-frequency representation of the signal termed 'modulation spectrogram' ($X(f_m, f)$), which characterizes the rate-of-change of different spectral components. Here, f is used to characterize the conventional frequency (Hz) and f_m the modulation frequency. As the majority of the modulation spectral content of speech is known to lie below 20 Hz f_m [24], parameters used in the computation of the modulation spectrogram have been chosen here such that $f_m = 0 - 20$ Hz and $f = 0 - 8$ kHz.

While one may choose to use the modulation spectrogram directly as input to machine learning algorithms (i.e., to be treated as an image, as most applications relying on spectrograms do [25]), it increases the model complexity and makes the system less interpretable. To improve the interpretability and simplicity of the proposed method, here we have decided to extract spectral power and descriptor features from the modulation spectrogram and use those instead to reduce the input dimensionality. Fig. 2 depicts this feature extraction process. We first quantize the modulation spectrogram into 400 bins by grouping the 0–20 Hz f_m axis into twenty 1-Hz bins. Similarly, the 0–8 kHz frequency axis is grouped into twenty 400-Hz bins. Next, the spectral power of each bin is then normalized by total power of the modulation spectrogram, then used as modulation spectral energy features. Next, we compute eight spectral shape descriptors for each of the 20 conventional frequency bins (i.e., descriptors computed across modulation frequency axis), as well as for the 20 modulation frequency bins (i.e., across conventional frequency). This results in an additional 320 features ($20 \times 8 \times 2$). The eight descriptors include: spectral centroid, entropy, spread, skewness, kurtosis, flatness, crest, and flux. A detailed description of these descriptors can be found in [18], [26]. Overall, a total of 720 modulation spectral features (MSF) are computed and tested.

B. Vocal Tract and Excitation Signal Decomposition

In an effort to build COVID-19 diagnostic tools that have improved interpretability, we part from the hypothesis that

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

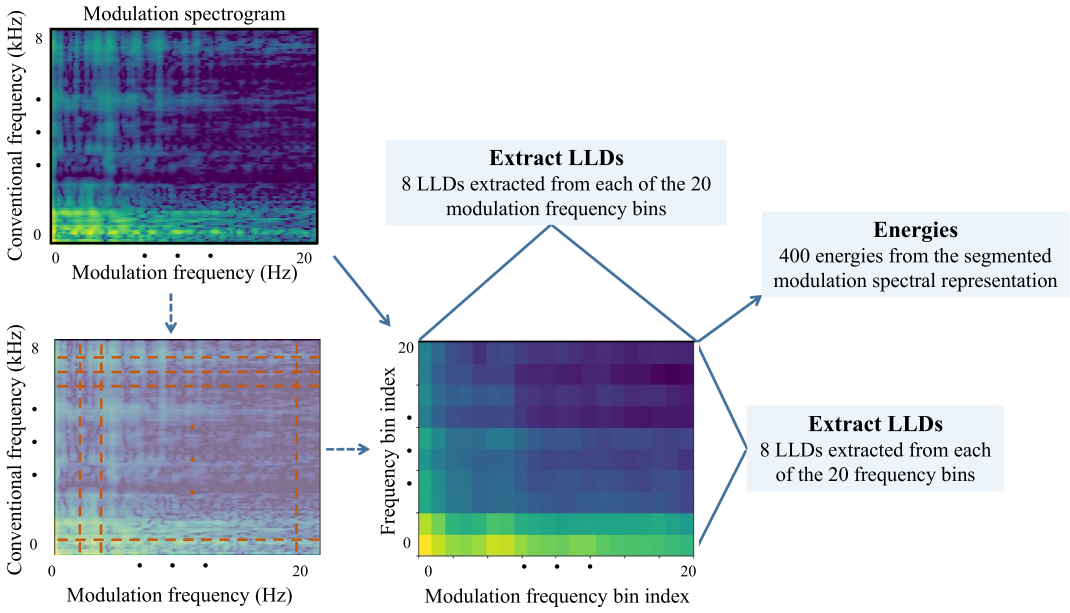


Fig. 2. Extraction of modulation spectrogram features.

COVID-19 can affect both the vocal tract properties (e.g., via increased muscle fatigue [9], [27]) and the excitation (e.g., impaired phonation [7]). While modulation spectral features may provide some insights from the former, we propose to also use linear prediction (LP) analysis to decompose the speech signal into vocal tract parameters (i.e., LP coefficients) and the excitation source (i.e., LP residual) [21]. Linear prediction analysis assumes that speech is generated by the excitation of a linear time-varying filter (vocal tract) by impulses for voiced speech segments or random noise for unvoiced speech segments [28]. The vocal tract can be modeled as an all-pole filter, of which the transfer function can be given by:

$$H(z) = \frac{1}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{1}$$

where G is the gain factor of the LP filter and is set to 1, p is the order of the LP filter and implies that the past p samples are used in the prediction of the current sample. More specifically, the speech signal $s(n)$ can be approximated by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n), \tag{2}$$

where the coefficients a_k are the linear predictive coefficients (LPCs) and $p = 18$ is used herein, as speech sampled at 16 KHz is used in our experiments.

The excitation signal, in turn, is represented by the prediction error signal $e(n)$, which is the difference between the estimated speech $\hat{s}(n)$ and the original speech $s(n)$, i.e.,

$$e(n) = s(n) - \hat{s}(n) = Gu(n), \tag{3}$$

where a_k in (2) can be estimated by minimizing the energy of $e(n)$ using Burg's method [29]. Henceforth, we utilize features extracted from the LP residual as features characterizing the excitation and the $a_k, k = 1, \ldots 18$ as features characterizing the vocal tract. The excitation signal, however, differs based on the periodicity of the speech segment. Voiced speech segments are produced when quasi-periodic pulses of air generated by the vibration of vocal folds resonate through the vocal tract, where the fundamental frequency of vibration of the vocal folds is usually interpreted as pitch [30]. As such, the residual signal has strong impulse-like peaks corresponding to the glottal pulses produced during voiced speech. Unvoiced segments, on the other hand, are produced without vibration of the vocal chords, and have residuals that are commonly modeled as noise [30]. In our analysis, we use the pYAAPT open source pitch tracker [31] in order to separate voiced and unvoiced segments prior to LP analysis. The following pitch tracking hyper-parameters were used in our analyses: 30 Hz minimum pitch searched, 400 Hz maximum pitch searched, 15 ms frame length, and 5 ms frame hops.

From the LP analysis, different features are then extracted from the vocal tract and excitation signals from the voiced and unvoiced segments. Fig. 3 depicts the steps taken for the calculation of the proposed LP features. First, LP analysis is performed on *voiced* speech frames and the following features are extracted: (i) 18 LPCs, to indicate the shape and resonance characteristics of the vocal tract, (ii) mean and standard deviation of the pitch values computed over all voiced speech frames in a recording, and the (iii) mean, standard deviation, kurtosis, and skewness of the voiced LP residual, computed per frame. These latter four per-frame features are finally aggregated using the mean and standard deviation statistics, thus resulting in eight voiced LP residual temporal dynamics features. Lastly, the same eight residual temporal dynamics features are extracted, but for the *unvoiced* frames. A total of 36 LP features are computed $(18 + 2 + 8 + 8)$.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

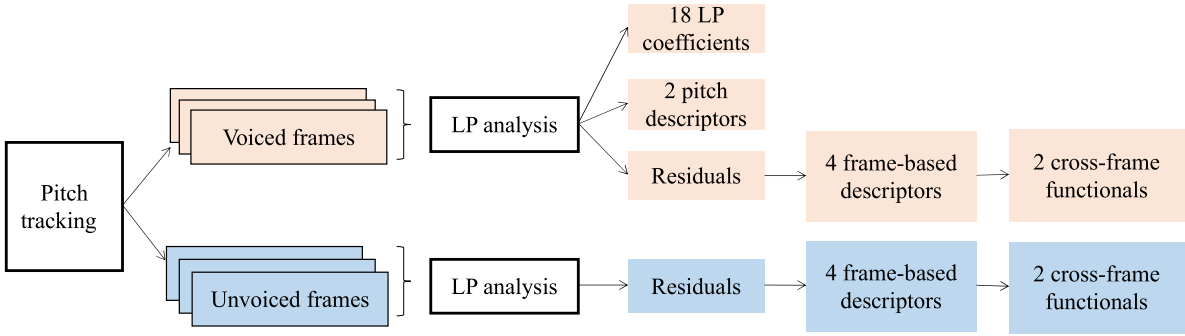


Fig. 3. Signal processing steps involved in the computation of the vocal tract and excitation signals from LP analysis.

C. Feature Selection

In order to reduce the number of features to be input to machine learning algorithms and better understand the importance of the different extracted features, the Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithm was used. The MRMR algorithm finds the optimal feature subset, such that the top-ranked features are mutually dissimilar, while being maximally related to the outcome variable (i.e., in our case, the COVID-19 diagnostic) [32]. It is a filter-based feature selection method in the sense that feature ranking is performed independent of the downstream machine learning algorithm, hence allowing for experiments with different models to be performed using the selected top features. Moreover, it allows for assessment of the potential of new engineered features and not necessarily on the machine learning algorithms. As we are interested in gauging the benefits and interpretability of the proposed features, a filter method is preferred. In this study, the number of features to be selected was sequentially experimented from 5 to 100, where the optimal number was determined by the performance achieved during cross-validation trials.

D. Classifiers and Fusion Methods

As mentioned above, results from the 2021 ComParE COVID-19 Challenge have shown that conventional models, such as support vector machines (SVM), can achieve similar results relative to DNNs [11]. Meanwhile, as is pointed out in [33], overly complex models could easily overfit to small datasets, hence leading to lower generalizability across unseen data conditions. Therefore, we explore the use of conventional classifiers in order to achieve better interpretability as well as lower computational complexity. In particular, a linear SVM classifier is used, as well as a decision tree (DT) classifier, as these have shown increased interpretability in the past [34]. Random forests (RF) are also explored, as they have shown high accuracy across different clinical applications and can be made interpretable [35]. In all cases, the scikit-learn toolkit was utilized [36]. For the SVM, a linear kernel was used. For the decision tree classifiers, a maximum tree depth of 10 was selected since we observed larger values to result in over-fitting; other parameters were set to default values.

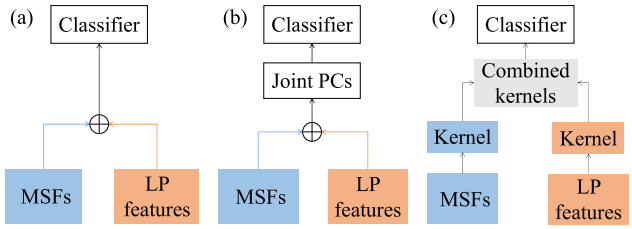


Fig. 4. Feature fusion schemes tested: (a) early-stage fusion, (b) intermediate-stage fusion via PCA, and (c) multiple kernel learning fusion.

Next, we take the advantage of the complementarity of the proposed features and explore different feature fusion schemes. First, three conventional feature fusion methods are explored, as shown in Fig. 4, namely: (a) early-stage fusion, where MSF and LP features are aggregated into one larger feature vector, (b) features from the two methods are aggregated via principal component analysis (PCA), and (c) in the case of the SVM classifier, a multiple kernel learning method is explored where different kernels are tested for each feature set, and the optimal kernel combination is found. Here, the open source MKLpy toolbox [37] was used for multi-kernel learning, where the regularization hyper-parameter λ was tuned empirically by experimenting values from 0 to 1.

In addition to feature fusion, decision level fusion combines decisions made by different systems. This architecture is favoured when errors made by different systems are mutually exclusive [38]. Decision level fusion, however, requires additional data to train the meta-classifier. Motivated by the previous findings where a reduced coordination of speech subsystems was found in COVID-19 patients [7], we here propose a two-stage system to capture complications in the articulators or the phonation system (or both) (depicted by Fig. 5). While the first stage targets deficits in the articulators, samples deemed negative pass through a second stage to avoid potential abnormalities in the phonation system being overlooked. As such, the two-stage approach should lead to higher sensitivity levels. Our experiments showed that combining the top-40 MSFs and top-10 LP features, as ranked by MRMR, led to improved results. A linear SVM classifier was used for the MSFs, while a DT-based classifier was used for the top LP features. With this two-stage approach,

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

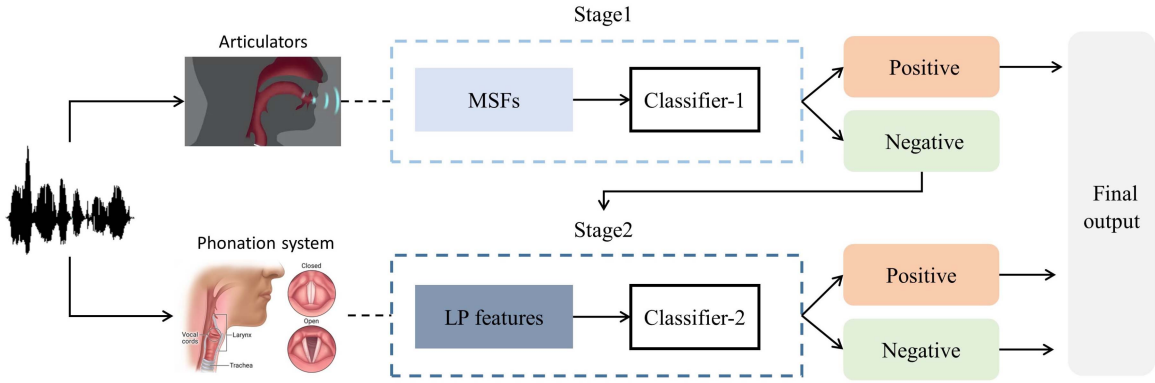


Fig. 5. Two-stage, decision-level fusion scheme proposed for improved COVID-19 detection.

TABLE I
DATASETS DEMOGRAPHICS (P-SYMP: COVID-POSITIVE WITH SYMPTOMS; P-ASYMP: COVID-POSITIVE WITHOUT SYMPTOMS; N-SYMP: NON-COVID WITH RESPIRATORY SYMPTOMS; N-HEALTHY: NON-COVID WITHOUT ANY RESPIRATORY SYMPTOMS)

Dataset	Samples	Gender			Language		COVID-positive		COVID-negative	
		Male	Female	N/A	English	Other	P-symp	P-asymp	N-symp	N-healthy
CSS	582	35%	45%	20%	45%	55%	72%	28%	41%	59%
Cambridge	1486	62%	36%	2%	100%	0%	87%	13%	54%	46%
DiCOVA2	965	70%	30%	0%	100%	0%	87%	13%	14%	86%

recordings are first processed by the MSF-based system. Positive predictions are kept as is, whereas those deemed to be negative, are further processed by the LP based system to be fine-tuned, as shown in the figure.

III. EXPERIMENTAL SETUP

In this section, we describe the experimental setup, including the database used, benchmark systems, and system evaluation approaches.

A. Database Description

In this study, three datasets are employed, including (1) the COVID-19 Speech Sub-challenge dataset of the INTERSPEECH 2021 Computational Paralinguistics Challenge (CSS) [11], (2) a subset of the English-language Cambridge COVID-19 sound database (Cambridge subset) [39], and (3) the Second Diagnosis of COVID-19 using Acoustics Challenge (DiCOVA2) dataset [40]. For both the CSS and Cambridge subsets, participants were asked to utter the same speech content ('I hope my data can help to manage the virus pandemic') in their mother tongue language. With DiCOVA2, in turn, participants did number counting from 1 to 10 in a normal pace. For all three datasets, participants were asked to self-declare whether they were COVID-19 negative (including healthy or having COVID-19 like symptoms or pre-existing medical conditions) or COVID-19 positive (including symptomatic and asymptomatic cases).

Statistics of the three COVID-19 speech datasets are provided in Table I. Regarding gender distribution, both the Cambridge and DiCOVA2 datasets are skewed towards a male population, while CSS is slightly skewed to the female population. Moreover, while all recordings from the Cambridge and DiCOVA2 datasets are in English, eight languages are included in the CSS set, with English being the largest language group corresponding to 45% of the entire set. Furthermore, it can be seen that all three datasets include asymptomatic COVID-19 subjects as well as non-COVID subjects with some respiratory symptoms. Among these three datasets, CSS has the highest percentage of asymptomatic COVID-19 samples (28%), while the Cambridge subset has the highest percentage of symptomatic non-COVID samples (54%).

TABLE II
DATA PARTITIONS OF THE DIFFERENT DATABASES

Dataset	Partition	#COVID-positive	#COVID-negative
CSS	Training set	56	243
	Test set	130	153
Cambridge	Training set	490	530
	Validation set	82	60
	Test set	162	162
DiCOVA2	Training set	170	702
	Test set	43	150

Table II lists the number of samples present in the disjoint training and testing subsets used herein, as well as the validation subset in the case of the Cambridge dataset. Since the "true" test set partitions used in the CSS and DiCOVA2 challenges were only accessible to challenge participants, we had to re-organize the training and validation partitions available into new train and test subsets. This was achieved as follows. For the CSS data, as speaker information was not provided, we utilized the original validation set as our test set to prevent leakage of speaker information into the test set, thus making it truly unseen. The original

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

training set, in turn, is used here for both training and validation. For the DiCOVA2 dataset, we divided the original development set into speaker-independent sets for training and testing with a ratio of 80:20%. Lastly, as the Cambridge dataset was not part of any challenge, the original training, validation and test partitions are used herein. It is important to emphasize that 28 files in the CSS dataset had originally been upsampled from 8 kHz to 16 kHz, thus resulted in artificial frequency content between 4-8 kHz [41]. To avoid any biases in the simulations, these 28 files were omitted from the CSS dataset, thus resulting in the 299 and 283 recordings, respectively, for the training and test sets shown in Table II.

B. Benchmark Systems

As two of the used datasets were part of international COVID-19 challenges, we utilize their benchmark systems as baselines in our experiments. As stated above, since data partitions are different from the original challenge ones due to limited data accessibility, we reproduced the baseline systems following the system architecture described in their corresponding papers [11], [39], [40]. A brief description of the baseline systems is presented below.

1) ComParE + SVM: 6,373 acoustic features are extracted using the openSMILE toolbox [42] and fed into a linear SVM for final predictions. This system has been used as a benchmark on all three datasets. Note that the work in [11] showed that using the bag-of-words (BoW) methodology on openSMILE features could lead to improved results. BoW on top of modulation features has also shown useful for speech emotion recognition [43]. As we are interested in gauging the benefits of the proposed features, we compare them with the original openSMILE set and leave BoW extended features as future work.

2) Spectrogram + VGGish: Raw audio waveforms are first converted to 2-dimensional mel-spectrograms and then fed into a pre-trained VGGish network for classification. The VGGish backbone and the fully-connected layers are fine-tuned jointly on each dataset. This is a benchmark system used for the Cambridge subset.

3) Spectrogram Features + BiLSTM: 192 log mel-spectrogram features are extracted per speech frame, which are then fed into a bi-directional long short term memory (BiLSTM) recurrent neural network for classification. This is one of the benchmark systems for the DiCOVA2 dataset.

C. System Evaluation

Since the original validation set of the CSS and DiCOVA2 datasets are used as test sets in our study, we employ a 5-fold stratified cross-validation setup on the training set for hyper-parameter tuning. For consistency, we also aggregate the training and validation sets of the Cambridge dataset and follow the same cross-validation setup on the aggregated training set for hyper-parameter tuning. Models are then evaluated on the test set which remains unseen throughout training. As cross-validation randomly divides the training set into multiple folds, test scores might vary when using different cross-validation settings. To minimize the effect of such randomness, models are trained and tested repeatedly using 10 different cross-validation settings, hence resulting in 10 test scores per system for each given dataset. The Wilcoxon signed-rank test [44] is then conducted to statistically quantify the difference in performance between the proposed and benchmark systems.

In terms of evaluation metrics, the CSS challenge utilized the unweighted average recall (UAR) measure as the standard metric, whereas the DiCOVA2 challenge relied on the area under receiver operating characteristic curve (AUC-ROC). For fair comparisons with the benchmarks, here we report UAR for the CSS set and AUC-ROC scores for the DiCOVA2 and Cambridge subsets. The sensitivity (true positive rate) and specificity (true negative rate) are also included for system evaluation. To facilitate replication of the results reported herein, the scripts will be made available at https://github.com/zhu00121/two_stage_fusion.

D. Tasks

To develop an interpretable and generalizable system, our experiments are divided into two tasks:

Task-1: *Development of an interpretable system*. This task aims at finding and interpreting discriminatory features which can be linked to the underlying properties of COVID-19 speech. Different classifiers and feature fusion strategies are explored. Focus is placed on the CSS dataset to allow for the generalizability of the models to be explored in the next task.

Task-2: *Evaluation of system generalizability*. This task aims at evaluating the generalizability of the system developed in Task-1 via within-dataset and cross-dataset testing. For within-dataset evaluation, our system is first trained and tested within the DiCOVA2 and Cambridge datasets and performance is compared against the benchmark systems. This allows us to investigate whether the proposed interpretable features and system architecture remain effective on other unseen datasets. Next, we conduct a cross-dataset evaluation, where models are trained on data from one (or multiple) datasets and tested on the left-out unseen dataset. This cross-dataset testing setup allows us to test the generalizability of the proposed method and robustness of the system to unseen data conditions typically seen in a remote healthcare monitoring application.

IV. Experimental Results and Discussion

In this section, we describe the experimental results and discuss them in light of existing literature, including some study limitations.

A. Discriminatory Potential of the Proposed Features

As we are interested in developing interpretable models, we first explore the discriminatory potential of the two proposed feature sets using the CSS dataset. Modulation spectrograms, normalized and averaged over the training set files, are depicted

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

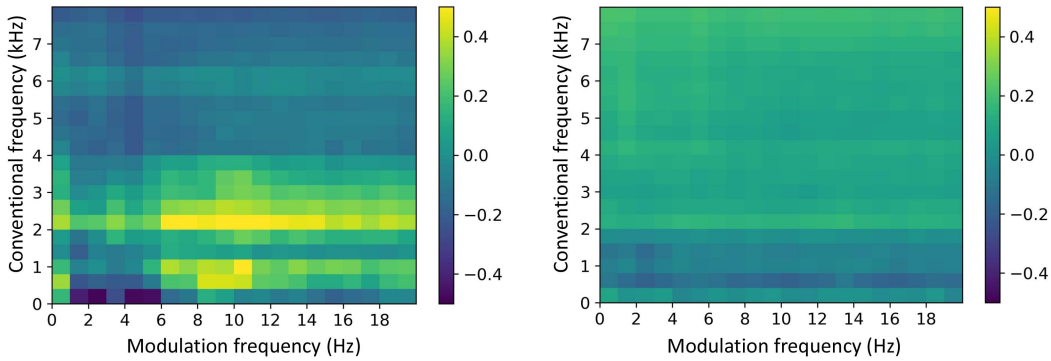


Fig. 6. Modulation spectrogram of COVID-19 speech (left) and non-COVID speech (right). Both are averaged across samples from the CSS training set.

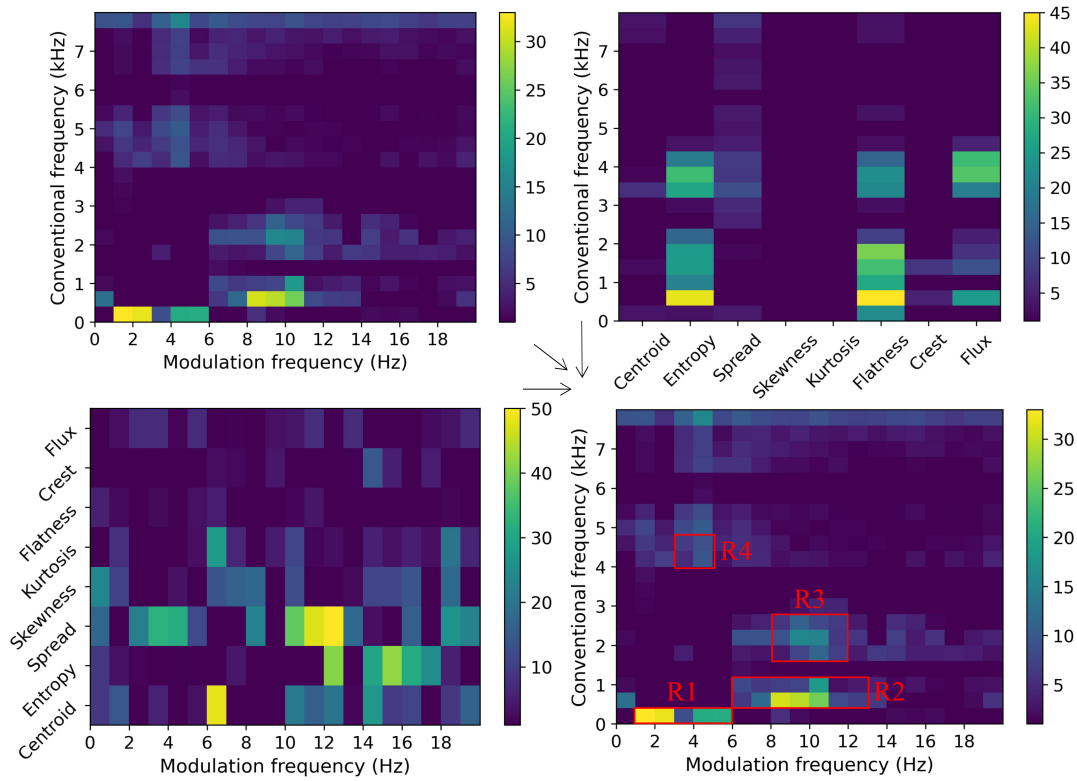


Fig. 7. F-ratio plots of modulation spectrogram energies (top-left), spectral shape descriptors computed across conventional frequency (top-right) and modulation frequency (bottom-left), and their four overlapping regions (bottom-right).

in Fig. 6 for the COVID-19 and non-COVID speech recordings. As can be seen, speech from COVID-19 patients shows some unique patterns, especially around $f_m = 6$–10 Hz and $f = 0.4$–1.2 kHz, and $f_m = 6$–14 Hz and $f = 2$–2.8 kHz.

To further explore the discriminatory potential of the modulation spectrum, the Fisher ratio (F-ratio) is used and computed between the modulation spectrograms of the two groups. To calculate the F-ratio, two estimates of the variance are made, including the between and within group variances, where the F-ratio is given by:

$$F - ratio = \frac{MS_{between}}{MS_{within}}, \tag{4}$$

where $MS_{between}$ represents the between group variance and MS_{within} represents the within group variance for each of the 400 entries of the modulation spectrogram (MS). Higher F-ratio values suggest increased group discrimination.

The top-left plot in Fig. 7 shows the F-ratio map indicating which parts of the modulation spectrogram provide the most discriminatory information. F-ratio analysis suggests increased discrimination around $f_m = 8$–11 Hz and $f = 0.4$–1.2 kHz, as well as between $f_m = 1$–6 Hz and $f = 0$–400 Hz, and to a lesser extent between $f = 6$–8 kHz. Similar F-ratio computations are also made with the eight spectral shape descriptors computed across the 20 conventional frequency bins (top-right plot in Fig. 7) and the descriptors computed across the 20 modulation frequencies (bottom-left plot in Fig. 7). Lastly, four regions

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

TABLE III
WELCH'S T-TEST RESULTS OF LP FEATURES FOR VOICED SEGMENTS. AN $*$ INDICATES FEATURES THAT ACHIEVED $p \leq 0.01$

Feature	t	p
LPC-1*	-3.1714	0.0016
LPC-2	1.8291	0.0682
LPC-3	-0.9879	0.3239
LPC-4	1.3276	0.1852
LPC-5	-2.3710	0.0185
LPC-6*	2.8390	0.0050
LPC-7	-2.2956	0.0226
LPC-8	2.1144	0.0356
LPC-9	-2.1762	0.0307
LPC-10	2.4304	0.0160
LPC-11*	-2.9794	0.0032
LPC-12*	3.3049	0.0011
LPC-13*	-2.9700	0.0033
LPC-14	2.0232	0.0441
LPC-15	-1.7128	0.0878
LPC-16*	2.6769	0.0078
LPC-17*	-3.0076	0.0028
LPC-18	1.5792	0.1152
Pitch (Mean)	-1.1187	0.2640
Pitch (Std)	-1.1843	0.2370

TABLE IV
WELCH'S T-TEST RESULTS OF LP RESIDUAL FEATURES FOR VOICED AND UNVOICED SEGMENTS. AN $*$ INDICATES FEATURES THAT ACHIEVED $p \leq 0.01$, WHILE $**$ INDICATES $p \leq 0.001$

Segments	Feature	t	p
Voiced	Mean (mean)	-0.1935	0.8466
	Std (mean)*	-2.8676	0.0043
	Kurtosis (mean)**	6.3310	0.0000
	Skewness (mean)**	6.1813	0.0000
	Mean (std)**	-4.2438	0.0000
	Std (std)*	-3.0236	0.0026
	Kurtosis (std)**	5.0586	0.0000
	Skewness (std)**	5.5291	0.0000
Unvoiced	Mean (mean)	-0.1169	0.9070
	Std (mean)	-2.5331	0.0116
	Kurtosis (mean)**	4.3677	0.0000
	Skewness (mean)*	2.6003	0.0097
	Mean (std)**	-4.0154	0.0001
	Std (std)**	-4.0621	0.0001
	Kurtosis (std)	1.6488	0.1000
	Skewness (std)	2.0208	0.0440

are found that overlap between the three MSF features classes, hence suggesting that these are the modulation frequency ranges providing the most discriminatory potential for COVID-19 detection from speech (bottom-right plot in Fig. 7). These ranges include: (1) $f_m = 1$–6 Hz and $f = 0$–0.4 kHz, (2) $f_m = 6$–13 Hz, $f = 0.4$–1.2 kHz, (3) $f_m = 8$–12 Hz, $f = 2$–2.4 kHz, and (4) $f_m = 3$–5 Hz, $f = 4$–4.8 kHz.

For the proposed LP features, a Welch t-test is used to test for significance between the two groups. The Welch's t-test is a parametric test that compares the means between two independent groups without assuming equal population variances [45], thus, is favoured when the sample sizes of two groups are unequal. The t-statistic and p-value are used to gauge the discriminatory potential of the LP features. Table III reports the statistical test results for the 20 features obtained from the voiced segments. A significance level of 99% ($p \leq 0.01$) is used and seen in 7 of the 18 LPCs. In particular, it is observed that the 6th, 12th, and 16th LPC of COVID-19 speech are significantly higher than those from non-COVID speech, whereas the 1st, 11th, 13th, and 17th LPC are higher for non-COVID speech. No significant differences are observed for the two pitch descriptors.

Table IV shows the test results for the residual features of the voiced and unvoiced segments. All residual features, except the mean, show significant differences between the two groups. In particular, COVID-19 speech demonstrates significantly higher values for average kurtosis and skewness, as well as the variation of kurtosis and skewness of voiced residual segments. Meanwhile, the non-COVID group shows higher variations of voiced residual signal values. Similar to the voiced residual features, COVID-19 speech shows significantly higher value of average kurtosis and skewness, while non-COVID speech demonstrates higher cross-frame variations of unvoiced residual signal values.

Based on these three tests, four residual features are selected for further investigation, including the average and standard deviation of kurtosis and skewness for the voiced and unvoiced residuals. Fig. 8 depicts the histograms of these four features (from the training set) for both COVID-19 and non-COVID groups. As can be seen, for the average kurtosis of the voiced residual signal, although the two groups share a similar center, the distribution for the COVID-19 group is positively skewed with a few samples distributed at higher values. For the average skewness of voiced residual signal, in turn, distributions of the two groups are centered differently, with the COVID-19 group showing a peak at a higher value. A similar pattern is observed with the two unvoiced residual features, with the COVID-19 group demonstrating higher values at the peak of the kurtosis and skewness distributions.

B. Feature Ranking Based on MRMR

In order to validate the aforementioned findings with the top-ranked features found by MRMR, Fig. 9 shows a heatmap indicating the region from the modulation spectrogram in which the top-40 features came from. Brighter areas indicate those used by a larger number of top-ranked features. As can be seen, the $f_m = 7 - 10$ Hz range shows greater importance, as do the $f = 0.8 - 1.2$ kHz, $f = 2 - 2.8$ kHz, and $f \geq 4$ kHz ranges. Comparisons between the MRMR feature heatmap in Fig. 9 and the F-ratio plots in Fig. 7 show several overlapping regions, especially regions R1, R3, and R4 in the latter figure, thus corroborating the efficacy of MRMR in selecting useful features and the MSFs in providing discriminatory information for the task at hand.

For the LP features, the top-10 selected by MRMR include: LPC-6, LPC-9, LPC-11, LPC-12, mean skewness of the residual voiced signals, mean average, standard deviation, kurtosis, and skewness of the unvoiced residuals, and the standard deviation of the mean residual for unvoiced segments. Comparisons with the Welch test show an overlap between seven of the top-10 MRMR selected features, hence again corroborating the effectiveness of the feature selection method and of the proposed features.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

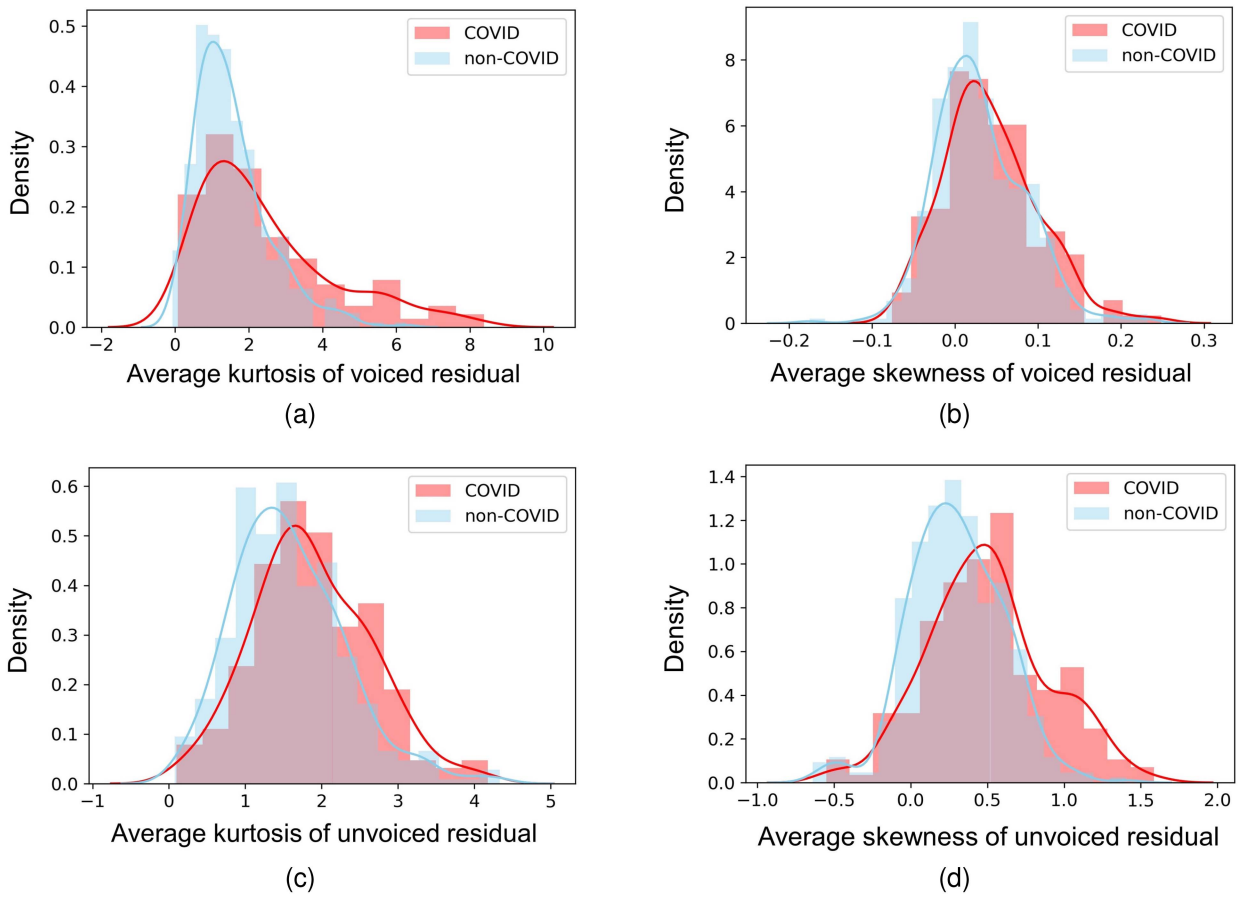


Fig. 8. Histograms of residual features for COVID-19 and non-COVID speech.

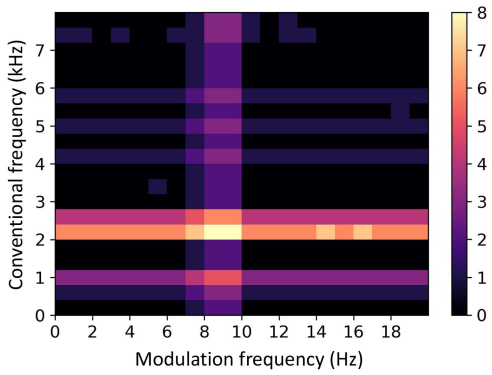


Fig. 9. Heatmap from top-40 MRMR-selected features showing the number of features selected from the different modulation spectrogram regions.

C. Feature Interpretation

One common symptom of COVID-19 is hoarseness due to inflammation of the vocal folds. In severe cases, the voice can become almost a whisper. Previous work on the modulation spectral analysis of whispered speech has shown that whispers can be manifested below $f = 1$ kHz and between 3 kHz$\leq f \leq 4.5$ kHz and $f_m \geq 10$ Hz [19]. These findings corroborate some of the regions found with both the F-ratio and MRMR, hence suggesting that some of the modulation spectral changes seen could be due to increased hoarseness of COVID-19 speech. Moreover, temporary neuromuscular impairments have been reported with COVID-19 [9], hence potentially affecting muscular control, and ultimately speech production. The seven LPC coefficients that show significant differences could be linked to such neuromuscular deficiencies. Moreover, as mentioned in [7], speech production can be modeled as a combination of airflow from the lungs, passing into the larynx, where coordinated coupling with phonation is achieved, followed by vocal tract shaping during articulation. As such, the proposed MSFs can also be seen as a correlate of the breathing amplitude modulations, especially those in the lower f_m range. Plots of COVID-19 modulation spectrograms in Fig. 6 demonstrate pronounced effects in this range, potentially due to breathing issues caused by impaired lung function [46]. Lastly, the significant changes seen in the LP residuals for both voiced and unvoiced segments also point towards phonation impairments, potentially also caused by inflammation of the larynx. For example, the increase in the LP residual kurtosis values in COVID-19 speech could be indicative of higher levels of vocal harshness, whereas the lower variability of LP residuals' values could suggest more breathy signals.

D. Classification Accuracy

To evaluate the efficacy of proposed features, Table V reports the UAR, sensitivity, and specificity levels achieved on the CSS dataset along with the top-performing classifier and the number

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

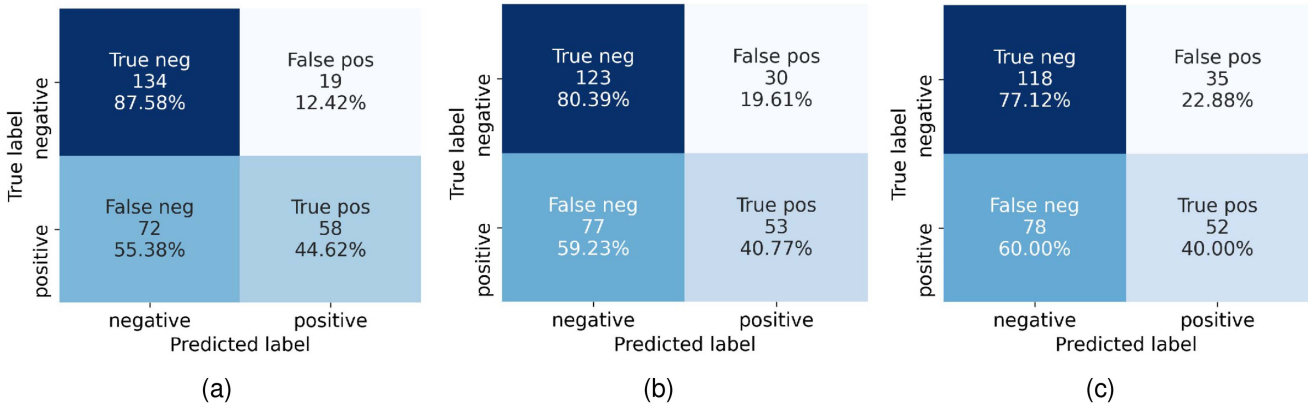


Fig. 10. Confusion matrices of systems tested on CSS (with best results chosen) for (a) MSFs based system, (b) LP features based system, and (c) Benchmark system.

TABLE V
PERFORMANCE OF PROPOSED AND BENCHMARK FEATURES USED INDIVIDUALLY WITH THE CSS DATASET. BOLD VALUES INDICATE BEST SYSTEM BASED ON A GIVEN FIGURE-OF-MERIT. STATISTICALLY SIGNIFICANT IMPROVEMENT RELATIVE TO THE HIGHEST BENCHMARK UAR IS HIGHLIGHTED WITH AN ASTERISK

Feature	#No.	Classifier	UAR	Sensitivity	Specificity
MSF	40	SVM	**0.661***	0.447	0.876
		DT	0.578	**0.515**	0.641
		RF	0.602	0.269	0.935
LP features	10	SVM	0.551	0.220	0.882
		DT	0.571	0.369	0.771
		RF	0.606	0.408	0.804
Benchmark	6373	SVM	0.537	0.139	0.935
		DT	0.566	0.426	0.706
		RF	0.586	0. 400	0.771
Benchmark	50	SVM	0.540	0.153	0.927
		DT	0.507	0.223	0.791
		RF	0.534	0.108	**0.961**

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT FUSION METHODS EVALUATED WITH CSS. BOLD VALUES INDICATE BEST SYSTEM BASED ON A GIVEN FIGURE-OF-MERIT. STATISTICALLY SIGNIFICANT IMPROVEMENT RELATIVE TO THE HIGHEST UAR OBTAINED BY SINGLE FEATURE MODALITY IS HIGHLIGHTED WITH AN ASTERISK

Fusion type	Configuration	UAR	Sensitivity	Specificity
Early-stage	SVM	0.664	0.380	**0.948**
	DT	0.607	0.347	0.867
	RF	0.630	0.418	0.842
Intermediate-stage	PCA+SVM	0.641	0.493	0.789
	PCA+DT	0.593	0.311	0.895
	PCA+RF	0.610	0.346	0.874
MKL	Polynomial	0.652	**0.746**	0.558
Two-stage	SVM+DT	0.674	0.508	0.840
	SVM+RF	**0.682***	0.531	0.833

of top-selected features used. As can be seen, the proposed MSF feature set achieves the highest UAR (0.661) and sensitivity (0.515). This is followed by the combined LP feature set, which achieves a UAR of 0.606 and sensitivity of 0.408. Both of the proposed feature sets outperform the benchmark system (ComParE+random forest) in terms of these two performance metrics, whilst requiring a significantly lower number of features (40). When comparing results achieved with the similar number of features (i.e., 40-50), the benchmark features, on the other hand, achieve the highest specificity value of 0.961. The marked increase in specificity achieved relative to sensitivity suggests that all models are better at classifying non-COVID samples rather than COVID-19 samples. Confusion matrices achieved for the three systems are shown in Fig. 10. As can be seen, only 24.7% of positive predictions made by MSFs were false alarmed. The LP features, however, render a false alarming rate that is almost twice as high (40.2%). On the other hand, the correct rejection rate of MSFs and LP features are relatively close (34.9% and 39.7%). Based on the high precision achieved by MSFs, the development of a hierarchical system is motivated, where MSFs pre-screen COVID-19 samples at the first stage, with the negative predictions being then sent as inputs to a second stage classification using LP features (see Fig. 5).

Next, we explore the effects of fusion on system performance. Table VI presents the results obtained with the four fusion methods described in Section II-D. As can be seen, the proposed two-stage system achieves the highest UAR (0.682), hence a 3.2% increase relative to the best single feature reported in Table V, and a 16.4% increase relative to the benchmark performance. The multi-kernel learning fusion method, in turn, results in the highest sensitivity (0.746), thus a 44.9% improvement relative to the sensitivity of the best single feature. Early-stage fusion achieves the highest specificity. Finally, the two-stage fusion method based on SVM and a random forest classifier achieves the best trade-off between sensitivity and specificity, hence could be a better candidate for use in the clinic. Overall, it is clear that fusion of both MSF and LP features results in the best accuracy, likely due to the fact that the different features are capturing different aspects of the disease in a complementary manner (see Section IV-C). Overall, the two-stage system is shown to statistically outperform the benchmark system in terms of UAR, whilst requiring a significantly lower number of features (i.e., 50 compared to over 6,000), thus is used throughout the remainder of this paper.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
PERFORMANCE COMPARISON ON CAMBRIDGE AND DICOVA2 DATASETS. SCORES REPORTED ARE AVERAGED OVER 10 DIFFERENT CROSS-VALIDATION RUNS. BOLD VALUES INDICATE THE BEST SYSTEM FOR A GIVEN METRIC. THE STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN THE HIGHEST AUC-ROC OBTAINED BY BENCHMARK SYSTEMS AND THE TWO-STAGE SYSTEM IS HIGHLIGHTED WITH AN ASTERISK

Dataset	System	Evaluation metrics		
		AUC-ROC	Sensitivity	Specificity
Cambridge	MSF+SVM	0.543	**0.760**	0.366
	LP+RF	0.595	0.575	**0.625**
	Two-stage	**0.612**	0.722	0.413
	ComParE+SVM	0.521	0.431	0.619
	Spec+VGGish	0.608	0.582	0.615
DiCOVA2	MSF+SVM	0.693	0.729	0.567
	LP+RF	0.633	0.401	**0.806**
	Two-stage	0.711	0.708	0.689
	ComParE+SVM	0.751	**0.731**	0.671
	Spec+LSTM	**0.770***	0.653	0.791

E. System Generalizability

While the two-stage system performs significantly better than the benchmark system on the CSS dataset, it is crucial to examine whether such system architecture and the proposed features can also be useful on other datasets. Table VII compares the performance of the proposed 2-stage system, as well as systems trained on single feature modalities and two benchmark systems, on the DiCOVA2 and Cambridge datasets. With an AUC-ROC of 0.612 achieved with the Cambridge set, the two-stage system outperforms both ComParE and VGGish-based benchmark systems. However, the improvement over the VGGish system is not statistically significant. Meanwhile, it can be noticed that using a small number of LP features alone achieves comparable performance relative to the more complex VGGish benchmark. Notwithstanding, the spectrogram with BiLSTM classifier remains the top-performer system on the DiCOVA2 dataset with an AUC-ROC of 0.770.

Moreover, it can be observed that the performance of all tested systems varies greatly across datasets. For example, the ComParE benchmark achieves an AUC-ROC of 0.751 on DiCOVA2 but only 0.521 on Cambridge and an UAR of 0.537 on the CSS dataset. A similar, but relatively smaller, variance can be observed with our proposed two-stage system, which achieves an AUC-ROC of 0.711 on DiCOVA2 and 0.612 on the Cambridge set. These findings suggest that the prediction tasks with CSS and Cambridge datasets may be more challenging relative to DiCOVA2, which may be due to the high percentage of asymptomatic COVID-19 samples and symptomatic non-COVID samples, as highlighted in Table I.

Notwithstanding, across all three datasets the proposed two-stage system performs consistently better than using any single feature modality. To better understand the trade-off between TPR and FPR, and the contribution of each feature modality in the fusion system, ROC curves for the Cambridge and DiCOVA2 datasets are shown in Fig. 11. As can be seen, the proposed two-stage system not only results (on average) in a higher AUC-ROC, but also in a more balanced trade-off between sensitivity and specificity.

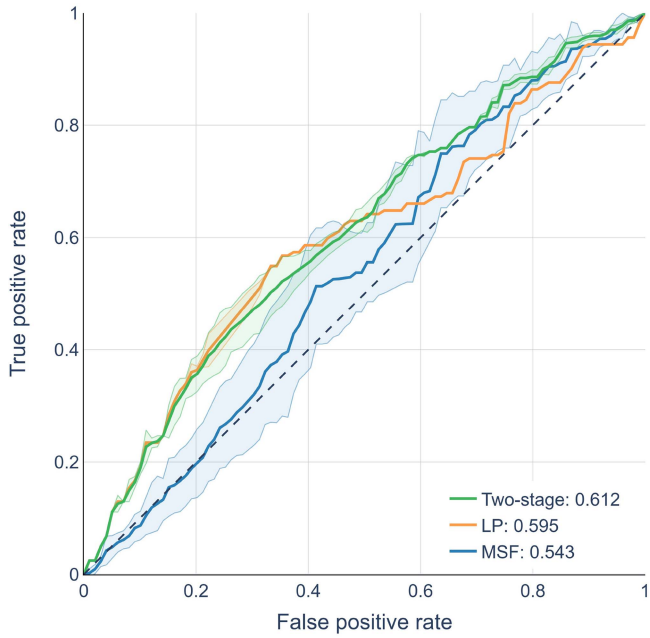


Fig. 11. ROC curves achieved by the two-stage system and single feature modality systems with (a) Cambridge subset and (b) DiCOVA2. Solid lines correspond to the average test AUC-ROC achieved across 10 different cross-validation settings. Shadowy region represents the range of 95% confidence intervals.

Finally, we test the generalizability of the investigated systems. Table VIII summarizes the AUC-ROC achieved with the different systems within cross-dataset testing conditions. In this experiment, training is performed on two combined datasets and tested on the third unseen set without any model fine-tuning. As can be seen, a marked drop is observed for all systems, which is

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
PERFORMANCE COMPARISON IN CROSS-DATASET TESTING CONDITIONS. SCORES REPORTED ARE AVERAGED OVER 10 DIFFERENT CROSS-VALIDATION RUN. BOLD VALUES INDICATE THE BEST SYSTEM FOR A GIVEN CONDITION. THE SYSTEM WITH A SIGNIFICANTLY IMPROVED AVERAGE AUC-ROC RELATIVE TO THE OTHER SYSTEMS IS HIGHLIGHTED WITH AN ASTERISK. CA:CAMBRIDGE; DI:DICOVA2

System	CSS+CA→DI	CSS+DI→CA	CA+DI→CSS	Ave
Two-stage	**0.606**	**0.554**	**0.578**	**0.579***
ComParE+SVM	0.570	0.515	0.506	0.530
Spec+LSTM	0.489	0.484	0.477	0.483
Spec+VGGish	0.491	0.502	0.485	0.493

in line with findings reported with other modalities (e.g., cough and image) [47], [48]. Interestingly, for the deep learning based systems, all performances degrade to below chance levels, suggesting that the systems may be overfitting to specific database nuances and not necessarily COVID-19 information. In comparison, our two-stage system maintains an average AUC-ROC of 0.579 across the three cross-dataset conditions, significantly outperforming all other benchmarks. These findings suggest that the proposed two-stage system may indeed be capturing relevant phonation and articulators information important for COVID-19 detection.

F. Study Limitations

Database imbalance plays a critical role in studies like the one described here. Among the three datasets included in this study, two of them (CSS and DiCOVA2) have fewer than 35% COVID-positive samples. Such imbalance can cause issues with classifier training and can lead to poor sensitivity values. One common approach to tackle this issue is data augmentation, which has been proven effective in many other studies (e.g., [49], [50]). Until a clear understanding of the disease and its effect on speech is obtained, data augmentation may lead to algorithmic biases, thus hampering interpretability of the results. Furthermore, different language distributions across portions of CSS dataset could lead to potentially biased results. In fact, our recent study has shown that fusing modulation spectrogram features with openSMILE features and patient metadata can help tackle this issue [51]. Finally, since only COVID-19 and healthy samples are included in the existing databases, it is not clear whether the proposed speech "biomarkers" are truly indicative of COVID-19 or of general respiratory abnormalities. Future studies are needed to validate the findings on other respiratory disease voice datasets.

V. CONCLUSION

In this paper, we explored the use of different feature sets to characterize changes across all stages of the speech production system (i.e., breathing modulation from the lungs, phonation, and vocal tract shaping) that can be affected by a COVID-19 infection. In particular, changes in the modulation spectrum, linear prediction coefficients, and linear prediction residuals are investigated and the top-ranked features were interpreted in light of existing literature. With the obtained insights, a two-stage COVID-19 prediction system is then proposed and tested on three different COVID-19 speech datasets. Experimental results show the proposed two-stage feature fusion system outperforming several benchmarks, especially those based on complex deep neural networks. Lastly, the two-stage system is tested in a cross-dataset evaluation scheme and shown to achieve greater generalizability across unseen conditions. Together, these findings suggest that our proposed speech-based system could be a promising alternative to rapid, low-cost, at-home testing for COVID-19.

DISCLAIMER AND ACKNOWLEDGEMENT

The authors would like to acknowledge the University of Cambridge for sharing their COVID-19 speech dataset. We would also like to acknowledge the organizers of both DICOVA2 and COMPARE challenges for the data collection and event organization. The aforementioned organizations do not bear any responsibility for the analysis and results presented in this paper. All results and interpretation only represent the view of the authors. Authors also acknowledge funding from INRS, NSERC, and CIHR.

REFERENCES

[1] G. Heckman, M. Saari, C. McArthur, N. Wellens, and J. Hirdes, "COVID-19 outbreak measures may indirectly lead to greater burden on hospitals," *CMAJ*, vol. 192, no. 14, pp. E384–E384, 2020.
[2] K. Green et al., "What tests could potentially be used for the screening, diagnosis and monitoring of COVID-19 and what are their advantages and disadvantages," *Centre Evidence-Based Med.*, vol. 13, 2020. Accessed: Apr. 16, 2023. [Online]. Available: https://www.cebm.net/wp-content/uploads/2020/04/CurrentCOVIDTests_descriptions-FINAL.pdf
[3] F. Olearo et al., "Handling and accuracy of four rapid antigen tests for the diagnosis of SARS-CoV-2 compared to RT-qPCR," *J. Clin. Virol.*, vol. 137, 2021, Art. no. 104782.
[4] E. Surkova, V. Nikolayevskyy, and F. Drobniewski, "False-positive COVID-19 results: Hidden problems and costs," *Lancet Respir. Med.*, vol. 8, no. 12, pp. 1167–1168, 2020.
[5] S. S. Khandker, N. H. H. Nik Hashim, Z. Z. Deris, R. H. Shueb, and M. A. Islam, "Diagnostic accuracy of rapid antigen test kits for detecting SARS-COV-2: A systematic review and meta-analysis of 17,171 suspected COVID-19 patients," *J. Clin. Med.*, vol. 10, no. 16, 2021, Art. no. 3493.
[6] L. M. Kucirka, S. A. Lauer, O. Laeyendecker, D. Boon, and J. Lessler, "Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-COV-2 tests by time since exposure," *Ann. Intern. Med.*, vol. 173, no. 4, pp. 262–267, 2020.
[7] T. F. Quatieri, T. Talkar, and J. S. Palmer, "A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 203–206, 2020.
[8] P. Vetter, D. Vu, A. L'Huillier, M. Schibler, L. Kaiser, and F. Jacquerioz, "Clinical features of COVID-19," *BMJ*, Brit. Med. J. Publishing Group, vol. 369, 2020.
[9] J. Helms et al., "Neurologic features in severe SARS-CoV-2 infection," *New England J. Med.*, vol. 382, no. 23, pp. 2268–2270, 2020.
[10] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Front. Psychol.*, vol. 4, 2013, Art. no. 292.
[11] B. Schuller et al., "The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates," *Interspeech*, p. 431, 2021.
[12] J. Han et al., "An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety," in *Proc. Annu. Conf. Int. Speech. Commun. Assoc.*, pp. 4946–4950, 2020.
[13] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for COVID-19," 2020.
[14] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, and V. Aharonson, "SARS-CoV-2 detection from voice," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 268–274, 2020.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

[15] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Comput. Biol. Med.*, vol. 135, 2021, Art. no. 104572.
[16] M. A. Nessiem, M. M. Mohamed, H. Coppock, A. Gaskell, and B. Schuller, "Detecting COVID-19 from breathing and coughing sounds using deep neural networks," in *Proc. IEEE 34th Int Symp. Comput.-Based Med. Syst.*, 2021, pp. 183–188.
[17] T. Falk and W.-Y. Chan, "Spectro-temporal features for robust far-field speaker identification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 634–637.
[18] S. Wu, T. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.
[19] T. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Commun.*, vol. 54, no. 5, pp. 622–631, 2012.
[20] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 177–188, Jan.–Mar. 2021.
[21] J Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
[22] T. H. Falk, W.-Y. Chan, E. Sejdic, and T. Chau, "Spectro-temporal analysis of auscultatory sounds," *New Developments in Biomedical Engineering*. London, U.K.: IntechOpen, 2010, pp. 93–104.
[23] S. Bedoya, N. Katz, J. Brian, D. O'Shaughnessy, and T. Falk, "Acoustic and prosodic analysis of vocalizations of 18-month-old toddlers with autism spectrum disorder," in *Acoustic Analysis of Pathologies*. Berlin, Germany:Walter De Gruyter, 2020, pp. 93–126.
[24] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.
[25] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020.
[26] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO Ist Project Rep.*, vol. 54, pp. 1–25, 2004.
[27] N. P. Solomon, "What is orofacial fatigue and how does it affect function for swallowing and speech?," *Seminars Speech Lang.*, vol. 27, no. 4, pp. 268–282, 2006.
[28] J. Markel and A. Gray, *Linear Prediction of Speech*, vol. 12. Berlin, Germany:Springer, 2013.
[29] J Makhoul, "Stable and efficient lattice methods for linear prediction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 5, pp. 423–428, Oct. 1977.
[30] Z. Zhang, "Mechanics of human voice production and control," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. 2614–2635, 2016.
[31] S. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *J. Acoust. Soc. Amer.*, vol. 123, no. 6, pp. 4559–4571, 2008.
[32] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinf.*, vol. 18, no. 1, pp. 1–14, 2017.
[33] J. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, 1996.
[34] D. Slack, S. Friedler, C. Scheidegger, and C. Roy, "Assessing the local interpretability of machine learning models," *NeurIPS Workshop Human-Centric Mach. Learn.*, 2019.
[35] S. Hara and K. Hayashi, "Making tree ensembles interpretable: A Bayesian model selection approach," in *Proc. Int. Conf. Artif. Intell. Statist*, 2018, pp. 77–85.
[36] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[37] I. Lauriola and F. Aiolli, "MKLpy: A python-based framework for multiple kernel learning," 2020, *arXiv:2007.09982*.
[38] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.
[39] T. Xia et al., "COVID-19 sounds: A large-scale audio dataset for digital respiratory screening," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/e2c0be24560d78c5e599c2a9c9d0bbd2-Abstract-round2.html
[40] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The second dicova challenge: Dataset and performance analysis for diagnosis of COVID-19 using acoustics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 556–560.
[41] H. Coppock et al., "A summary of the ComParE COVID-19 challenges," *Frontiers Digit. Health*, vol. 5, p. 7, 2022.
[42] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[43] S. R. Kshirsagar and T. H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1892–1905, Oct.–Dec. 2022.
[44] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley Encyclopedia Clin. Trials*, vol. 8, 2007, pp. 1–3.
[45] B. L. Welch, "The generalization of 'STUDENT'S' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.
[46] M. Chowdhury, N. Hossain, M. Kashem, A. Shahid, and A. Alam, "Immune response in COVID-19: A review," *J. Infection Public Health*, vol. 13, no. 11, pp. 1619–1629, 2020.
[47] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Mach. Intell.*, vol. 3, no. 3, pp. 199–217, 2021.
[48] A. Akman, H. Coppock, A. Gaskell, P. Tzirakis, L. Jones, and B. W. Schuller, "Evaluating the COVID-19 identification resnet (CIdeR) on the INTERSPEECH COVID-19 from audio challenges," *Frontiers Digit. Health*, vol. 4, 2022.
[49] G. Deshpande and B. Schuller, "Audio, speech, language, & signal processing for COVID-19: A comprehensive overview," 2020, *arXiv:2011.14445*.
[50] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020.
[51] Y. Zhu and T. H. Falk, "Fusion of modulation spectral and spectral features with symptom metadata for improved speech-based COVID-19 detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8997–9001.

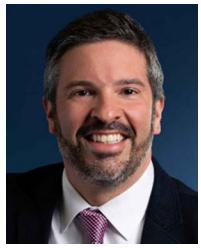
Yi Zhu (Student Member, IEEE) received the B.S. degree in 2018 from Southeast University, Nanjing, China, in 2018, and the Master's degree in biomedical engineering from University of Minnesota – Twin Cities, Twin cities, MN, USA, in 2020. He is currently working toward the Ph.D. degree with INRS-EMT, University of Quebec, Quebec City, QC, Canada. His research interests include human-centered signal analysis, machine learning, cybersecurity, and audio-based healthcare applications.

Abhishek Tiwari received the Ph.D. degree in biomedical signal processing for wearable devices from INRS-EMT, University of Quebec, Quebec city, QC, Canada, in 2021. He is currently with Myant Inc. on developing models for wellness and health detection from data collected using smart textile. His interests include signal quality assessment, feature engineering, and quality-aware physiological signal modelling.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.

João Monteiro is currently a Research Scientist at ServiceNow Research with a keen interest in creating learning algorithms and models that are robust against variations in data properties during training and testing. He focuses on developing model classes that can be verified at testing time, resulting in predictors with rejecting capabilities and improved adversarial robustness. His recent research work includes projects such as defining efficient and general out-of-distribution detection approaches, adjusting classifiers outputs to improve robustness against spurious features, and developing multi-lingual code search without parallel training data. He has also been leading initiatives for training large language models of code, involving models with parameters ranging from hundreds of millions to a few billions, in distributed settings.

Shruti Kshirsagar received the Ph.D. degree from INRS-EMT, Montreal, QC, Canada, in 2022. She was short-term visiting Researcher with IIT Bombay (DAPLAB), Mumbai, India, NTU, Singapore (Temasek lab) and has interned at Lisnen and Bosch on acoustic scene analysis. She is currently with EERS global Inc. as an Audio, R&D Scientist.



Tiago Henrique Falk received the B.Sc. degree in electronics engineering from UFPE, Recife, Brazil, in 2002, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from Queen's University, Kingston, ON, Canada, in 2005 and 2008, respectively. From 2008 to 2010, he was a Postdoctoral Fellow with the University of Toronto, Toronto, ON, Canada. Since 2010, he has been with INRS-EMT, University of Quebec, Quebec city, QC, Canada, where he is currently a tenured Full Professor, the Director of the Multimodal/Multisensory Signal Analysis and Enhancement Lab, and a Founding/Regular Member of the INRS-UQO Joint Research Unit on Cybersecurity. His research interests include the signal processing and machine learning for next-generation human-machine interfaces.

Authorized licensed use limited to: WICHITA STATE UNIVERSITY LIBRARIES. Downloaded on March 06,2025 at 19:17:39 UTC from IEEE Xplore. Restrictions apply.