Spectral-Temporal Saliency Masks and Modulation Tensorgrams for Generalizable COVID-19 Detection

Yi Zhu and Tiago H. Falk
Institut national de la recherche scientifique, INRS-EMT, University of Québec, Montréal, Canada

***Abstract*—Speech COVID-19 detection systems have gained popularity as they represent an easy-to-use and low-cost solution that is well suited for at-home long-term monitoring of patients with persistent symptoms. Recently, however, the limited generalization capability of existing deep neural network based systems to unseen datasets has been raised as a serious concern, as has their limited interpretability. In this paper, we propose two innovations to help overcome these issues. First, we propose the use of a 3-dimensional modulation frequency tensor (called modulation tensorgram representation, MTR) as input to a convolutional recurrent neural network for COVID-19 detection. The representation is known to provide robustness against different environmental factors seen across datasets. Next, we propose the use of spectro-temporal saliency masking to aggregate regions of the MTR related to COVID-19, thus helping further improve the generalizability and interpretability of the model. Experiments are conducted on three public datasets and results show the proposed solution consistently outperforming two benchmark systems in within-, across-, and unseen-dataset tests. The proposed method relies on a similar number of parameters to the benchmark, thus a promising solution for at-home monitoring of COVID-19 infection.**

***Index Terms*—Generalizability, COVID-19 detection, modulation tensorgram, saliency map, spectral-temporal.**

I. Introduction

Since the outburst of coronavirus disease in 2019 (COVID-19), significant efforts have been made to investigate how to best control the pandemic through the development of reliable and accessible diagnostic tools [1]. It is known that the COVID-19 virus can induce infection in the respiratory tract, hence causing respiratory-related symptoms such as sore throat, cough, and shortness of breath [2]. Meanwhile, several studies have reported its effects on neuromuscular control, as well as proprioceptive functions [3], [4]. Together, these symptoms lead to degraded coordination of speech production, which further opens up the possibility of detecting COVID-19 via speech analysis. Such analyses may enable the development of remote, rapid, and low-cost diagnostic tools. Furthermore, as approximately 10% of COVID-19 survivors are reported to show prolonged respiratory symptoms [5], the development of long-term speech-based symptom monitoring tools could become an important ally for clinicians.

To accelerate the research in acoustics-based COVID-19 detection, several groups have collected COVID-19 speech samples from around the world and shared them publicly via challenges, including the 'ComParE' [6], 'DiCOVA', and 'DiCOVA2' [7], [8] Challenge datasets, as well as the complete Cambridge database [9]. The release of these datasets has facilitated the development of new acoustic features and machine learning based diagnostic models. As one of the earliest attempts, Schuller et al. proposed the use of openSMILE features [10] with a linear support vector machine (SVM) classifier and achieved an unweighted average recall (UAR) rate of 72.1% on the ComParE dataset [6]. This system was later employed as a baseline in the 2021 ComParE speech sub-challenge and was also the winning system of the challenge. The majority of the models proposed in subsequent studies, however, have relied on deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural network (RNN). Sharma et al., for example, achieved an area under the curve of the receiver operating characteristics (AUC-ROC) of 84.3% with a bi-directional long-short term memory (BiLSTM) model on the DiCOVA2 dataset [8]. A similar BiLSTM model was used in the winning system of the DiCOVA2 challenge [11]. Akman et al., in turn, proposed a 9-layer convolutional residual network and achieved an AUC-ROC of 78.7% on ComParE and 78.6% on DiCOVA [12].

Regarding input features, most deep learning based models have relied on the spectrogram representation [11]–[13]. CNN models usually take as input the mel-scaled spectrogram as a 2-dimensional (2D) image, while RNN models require spectrograms to be segmented into time frames, which are then flattened into 1-dimensional (1D) feature vectors. While the spectrogram provides details about linguistic and paralinguistic content [14], it is known that it is sensitive to environmental artifacts (e.g., background noise and/or room reverberation) and may be sub-optimal for speech diagnostics. In fact, recent studies have shown that the performance of CNN based COVID-19 detection systems could degrade to chance-level when tested on unseen datasets [12], [15], [16]. As COVID-19 speech recordings are often collected "in-the-wild", care must be taken to ensure that environmental noise does not hamper diagnostic accuracy.

One alternate representation that has been explored in speech applications due to its noise-robustness properties is the modulation spectrum representation (MSR) [17]–[19]. The MSR is a 2-dimensional (2D) frequency-frequency representation that decouples signal components from noise as their spectro-temporal dynamics differ. As such, it has become a prime candidate for analyzing in-the-wild speech recordings [20]–[22]. In fact, Zhu et al. recently proposed the use of the MSR to characterize the changes in the movement of articulators resulting from the disease. The system was shown to not only outperform several baseline COVID-19 detection systems but also generalize better across datasets [16], [23]. In the system described in [23], low-level descriptor features were

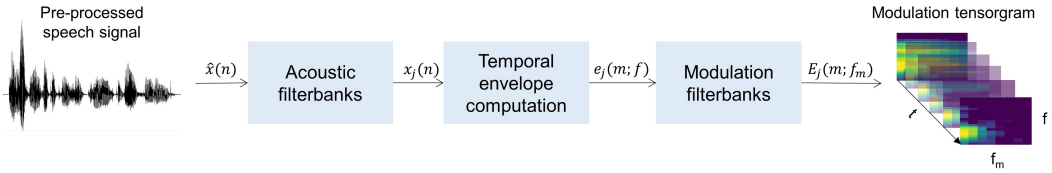


Fig. 1: Block diagram of the processing steps to compute the 3D modulation tensorgram.

extracted from the MSR (e.g., centroid at a specific modulation frequency band) and used as input to either SVM or deep machine learning models. Motivated by the gains seen in conventional spectrogram-based COVID-19 detection systems where the temporal dynamics of the spectrogram played a crucial role (e.g., [11]), it is hypothesized that similar benefits could be obtained from the MSR.

To validate this assumption, in this study we investigate the use of a 3-dimensional (3D) modulation tensorgram representation (MTR). Much like in spectrogram-RNN based solutions where spectrograms are computed over certain window sizes (which are then shifted and the process is repeated until the end of the speech utterance), the same is achieved here. The resulting time-frequency-frequency representation is thus comprised of temporally cascaded MSR snapshots, each computed over a certain window length, thus providing an overall depiction of the temporal changes of the MSR (more details to follow in the next section). As recurrent neural networks were shown in the past to take advantage of this temporal dynamics information, here we propose a customized convolutional recurrent neural network (CRNN) model trained on top of 3D MTRs. For simplicity, this system will be henceforth described as MTR-CRNN. Further, as interpretable models are desirable, especially for healthcare applications, we propose a spectral-temporal saliency map to identify the salient regions in the MTR being used by the diagnostic system. The saliency analysis allows not only for more interpretable findings but also enables the use of MTR masking, leading to a system that generalizes well across datasets.

The remainder of this paper is organized as follows. Section II describes the computational procedure to obtain the 3D MTR, the MTR-CRNN model architecture, and the details of the proposed spectral-temporal saliency maps. Section III describes the experimental setup, while Section IV describes and discusses the obtained results. Lastly, Section V presents the conclusions.

II. Proposed System Description

A. Modulation Tensorgram Representation

It is known that the MSR captures the long-term dynamics of the speech signal [24] and has been shown to carry meaningful information about vocal characteristics, such as vocal hoarseness and breathiness [23], [25]. To generate the 3D MTR, we here follow the same computational procedure described in [22]. The general processing pipeline is depicted in Fig. 1. First, as speech recordings are collected with different devices, the signal amplitude is normalized to remove unwanted amplitude variations caused by different loudness levels. Next, the pre-processed signal $\hat{x}(n)$ is filtered by acoustic filterbanks. While the gammatone filterbank is commonly used to mimic human perception of sound [26], it is not clear whether such a filterbank remains optimal for processing COVID-19 speech. For example, a recent study showed that a bio-inspired filterbank could outperform the conventional gammatone filterbank when analyzing COVID-19 coughs [27]. Hence, we experiment with two different 23-channel filterbanks: linear-scale and gammatone [28]. Furthermore, based on insights from [21], different lower and upper frequency ranges are explored. After applying the first filterbank, the temporal envelope $e_j(n)$ is computed from each filtered signal $\hat{x}_j(n)$ via the Hilbert transform:

$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + \mathcal{H}(\hat{x}_j(n))^2}, \tag{1}$$

where $\mathcal{H}(\cdot)$ denotes the Hilbert transform and the subscript j denotes jth filterbank. To obtain temporal dynamics information, each temporal envelope $e_j(n)$ is then windowed with a 256-millisecond (ms) Hamming window and an overlap of 216 ms. Such window length is relatively longer than that used in conventional spectrograms (e.g., 16 ms) and has been shown to provide appropriate resolution in low-frequency modulation frequencies [29]. To obtain the modulation spectrum $E_j(m;n)$ from each acoustic frequency component, the discrete Fourier transform $\mathcal{DFT}(\cdot)$ is applied to the temporal envelope $e_j(n)$:

$$E_j(m; f_m) = \left|\mathcal{DFT}(e_j(m;n))\right|, \tag{2}$$

where $|\cdot|$ denotes the absolute value operation, m denotes the frame number, and f_m denotes modulation frequency. An 8-channel modulation filterbank is then used to group neighboring modulation frequencies. Similar to the acoustic filterbanks, two different modulation filterbank types are tested, as are different lower/upper modulation frequency values. Table I summarizes the types of filterbanks tested and upper/lower frequency ranges. The optimal settings found through our experimentation are also reported in the table. Lastly, all MSRs computed per frame are aggregated into a final 3D representation called a "modulation tensorgram".

Figure 2 illustrates the importance of using a 3D tensorgram representation, as opposed to an averaged 2D representation, as in [23]. On the far left, the MSR averaged over all frames is shown. On the right, ten different MSR snapshots are shown. As can be seen, the modulation spectral patterns can differ greatly across time frames and such changes may carry important diagnostic cues.

B. Model architecture

The model architecture of MTR-CRNN is depicted in Fig.3. The CRNN model is comprised of two blocks, namely the 3D

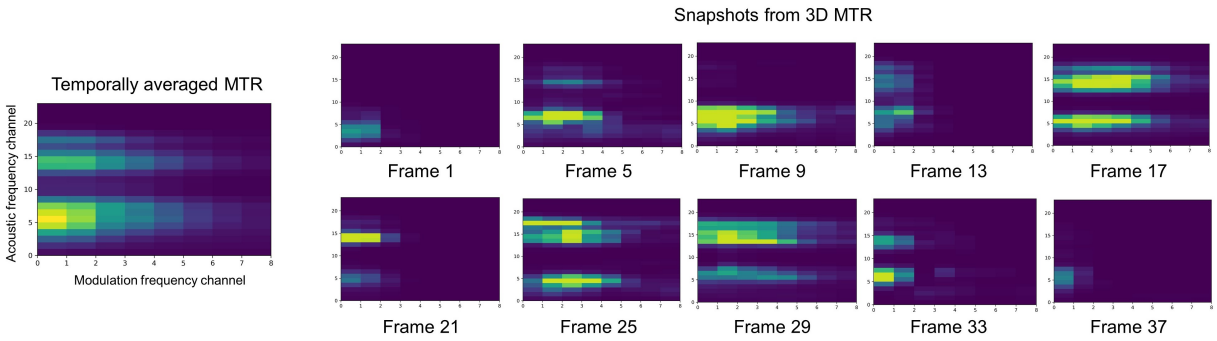


Fig. 2: Examples of the averaged 2D MTR and MTR snapshots at ten different frames. Examples are generated from the same speech sample.

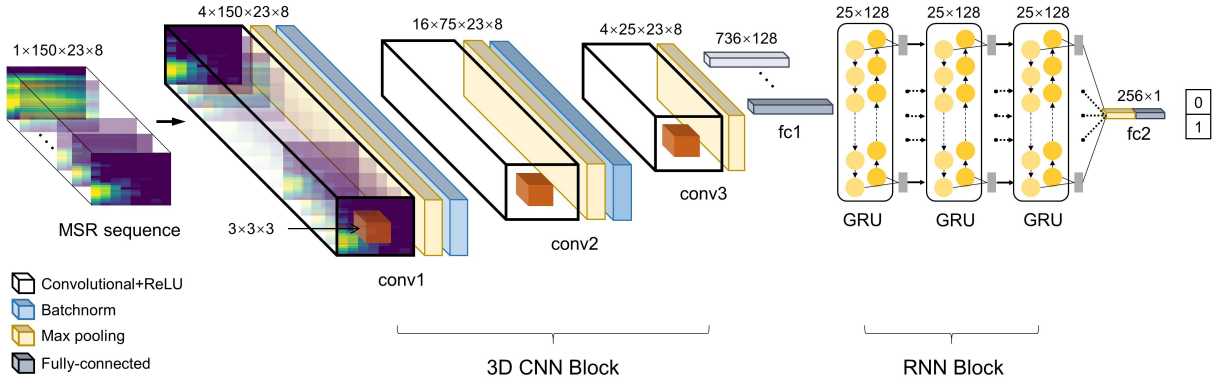


Fig. 3: Model architecture of the proposed MTR-CRNN system.

TABLE I: Overview of the modulation spectrogram parameter search detailing types of the filterbank, acoustic frequency (f) range, and modulation frequency (f_m) range.

Parameter	Range	Step	Optimal
Acoustic filterbank	gammatone, linear	-	gammatone
Modulation filterbank	log, linear	-	log
Lower bound f (kHz)	0-1	.125	.125
Higher bound f (kHz)	6-8	1	8
Lower bound f_m (Hz)	0-3	1	3
Higher bound f_m (Hz)	16-128	8	32

CNN block and the RNN block. As a uniform input shape is required for convolutional layers, speech samples are first unified to 10 s length by right zero-padding shorter recordings and segmenting longer recordings. This leads to a consistent 3D MTR shape $\{1 \times 150 \times 23 \times 8\}$ across all speech samples. Each input 3D MTR is then mean-variance normalized. Each part of the CNN block consists of a convolutional layer, a batch normalization layer, and a max pooling layer. A $\{3 \times 3 \times 3\}$ kernel is used for all three convolutional layers to extract meaningful modulation spectral patterns from neighboring MTR snapshots. The max pooling layer aims to remove the redundant MTR snapshots with relatively low energies, as these MTR snapshots usually correspond to silent frames. The output of the CNN block is then fed into a fully-connected layer to reduce feature dimensionality, leading to an output sequence of shape $\{25 \times 128\}$.

The subsequent RNN block has three cascaded bi-directional gated recurrent unit (GRU) layers to explore the temporal dependency of neighboring MTR snapshots. The output of the RNN block is then layer-normalized and fed into a pooling layer which finds the maximal value along the time dimension to generate a sequence-level embedding of shape $\{1 \times 256\}$. Lastly, a fully-connected layer is used to project the 256-dimension embedding to a 1-dimension output, which is then passed through a sigmoid layer to obtain the final COVID-19 probability score. To avoid over-fitting, a dropout factor of 0.7 is applied to the last fully-connected layer.

C. MTR Saliency Maps

Previous studies have suggested that different regions of the modulation spectrum correspond to different properties of the speech signal [23]–[25]. A better understanding of the MSR regions being used by the model would allow for better interpretation of the results and could lead to insights about the acoustic properties of COVID-19 speech. To this end, we propose a spectral-temporal saliency map based on the "vanilla gradient" saliency map algorithm originally invented for weakly supervised learning [30]. The method has been shown to be more robust than perturbation-based methods, thus is a good candidate for in-the-wild data [31].

The processing steps used to compute the spectral-temporal saliency maps is depicted in Fig. 4. First, the vanilla gradient method is used to compute raw saliency maps from a trained MTR-CRNN. The output map shape remains the same as the

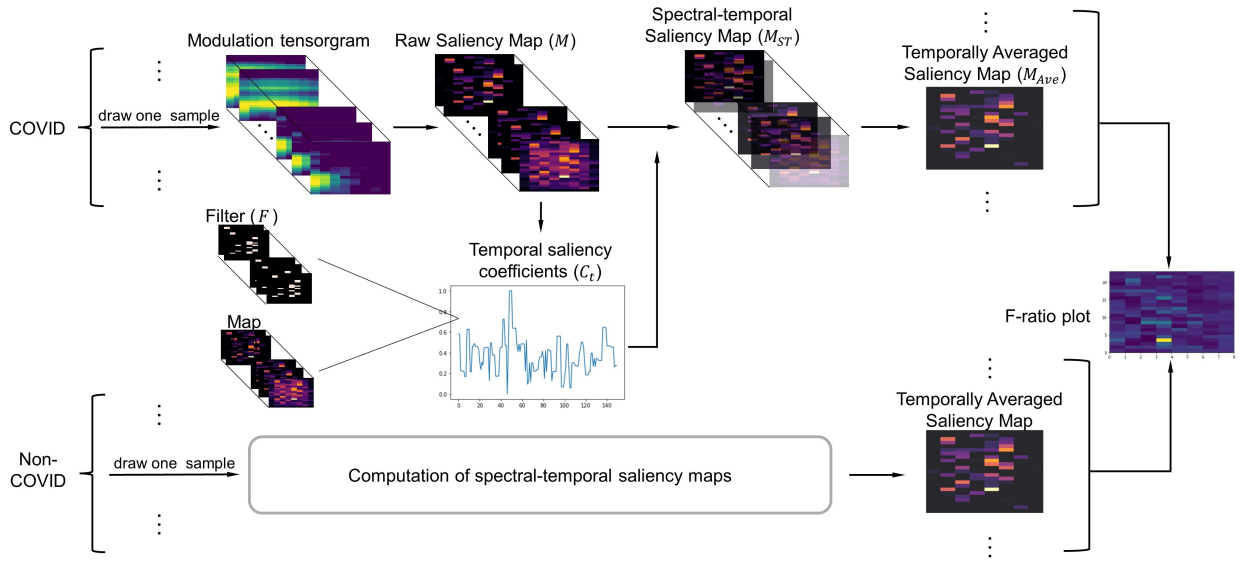


Fig. 4: Computation of the spectral-temporal saliency maps and the F-ratio plot. Only training data are used.

input shape, i.e., $\{150 \times 23 \times 8\}$. Although the raw saliency maps suggest attentive regions in each MTR snapshot, the temporal saliency is not well presented. Inspired by an audio-visual fusion saliency model [32], we solve this issue by first transforming the 3D raw saliency maps to a set of 1D temporal saliency coefficients. The transformation procedure is as follows. A 3D filter F is first used to discard the low-saliency regions in each 2D saliency map $M(t,i,j)$ with a pre-determined threshold:

$$F(t,i,j) = \begin{cases} 1 & \text{for} \quad |M_S(t,i,j)| \geq 0.2\max\{M\} \\ 0 & \text{for} \quad |M(t,i,j)| < 0.2\max\{M\} \end{cases} \tag{3}$$

where M denotes the raw 3D saliency map. Next, the filter is applied to each 2D saliency map and an averaging is used to obtain the temporal saliency coefficient C_t:

$$C_t(t) = \frac{\sum_{i,j}(F(t,i,j) \odot M(t,i,j))M(t,i,j)}{\sum_{i,j} F(t,i,j)M(t,i,j)}. \tag{4}$$

Each set of coefficients is of shape $\{150 \times 1\}$, corresponding to the temporal attention scores at each time step. To unify the coefficient range across samples, each set of coefficients is normalized between 0 and 1 and a 1D median filter is applied. Lastly, the temporal saliency coefficients C_t are multiplied with the raw saliency map to obtain the spectral-temporal saliency map M_{ST}:

$$M_{ST} = C_t(t)M(t). \tag{5}$$

As the final goal is to localize modulation spectral regions that are most closely related to COVID-19, the 3D spectral-temporal saliency map M_{ST} is then averaged over time which results in a single 2D saliency map M_{Ave} per sample. To further explore group differences, the Fisher ratio (F-ratio) is computed between two groups (COVID-19 positive and COVID-19 negative) of temporally averaged saliency maps:

$$\textit{F-ratio} = \frac{VAR_b}{VAR_w}, \tag{6}$$

where VAR_b represents the between-group variance, and VAR_w represents the within-group variance for each of the 23×8 saliency map values. Fisher ratio scores are then used to highlight the important discriminatory regions in the MTR.

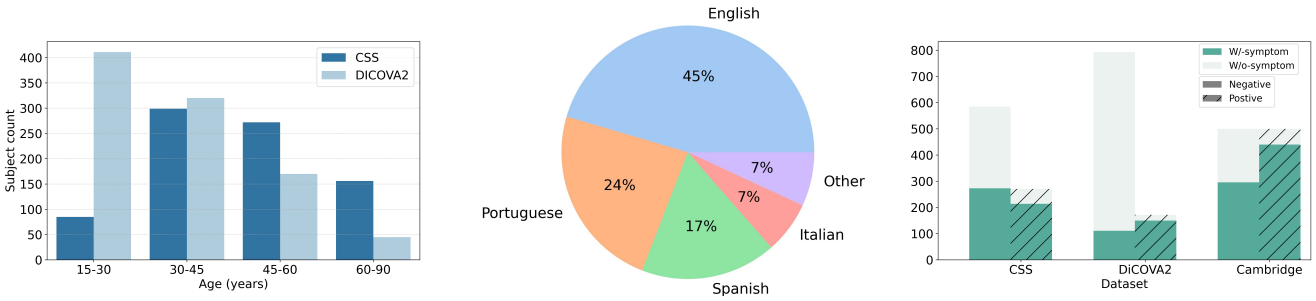
III. Experimental Setup

A. Dataset description

Three COVID-19 speech datasets are employed in our study: the ComParE COVID-19 Speech Sub-challenge dataset [6], the second DiCOVA Challenge dataset [11], and the English subset from the Cambridge COVID-19 sound database [9]. These datasets are referred to hereinafter as CSS, DiCOVA2, and Cambridge set, respectively. For all three datasets, volunteers across the world were encouraged to record and upload their voice data via Android and web apps[1]. With CSS, participants were asked to utter the sentence "I hope my data can help to manage the virus pandemic" at most three times in their mother tongue. The same speech content is used for the Cambridge set but uttered in English only. With DiCOVA2, participants did number counting from 1 to 10 in a normal pace in English. For all datasets, participants were asked to self-declare whether they were COVID-negative (including healthy or having COVID-like symptoms or pre-existing medical conditions) or COVID-positive (including symptomatic and asymptomatic cases).

An overview of the participant demographics from the three datasets is shown in Fig. 5. For CSS, a total of eight languages were included, with the majority of samples being uttered in English, Portuguese, Italian, and Spanish. The gender split was 45% female and 35% male (the remaining 20% chose *Prefer not to say* or *Other*). Close to 28% of the COVID-positive subjects were asymptomatic while only 59% of the COVID-negative subjects were without respiratory symptoms. For DiCOVA2, all recordings were in English only. The gender

[1]See https://www.covid-19-sounds.org and https://coswara.iisc.ac.in/



(a) (b) (c)

Fig. 5: Participant demographics for CSS, DiCOVA2, and Cambridge. (a) Subject age distribution for CSS and DiCOVA2, (b) Language distribution for CSS, (c) Respiratory symptoms distribution (W/: With; W/o: Without).

TABLE II: Data partition and class distribution.

Dataset	Partition	Positive	Negative	Total
CSS	Train	56	243	299
	Validation	130	153	283
	Test	87	189	266
DiCOVA2	Train	137	635	772
	Test	35	158	193
Cambridge	Test	500	500	1000

distribution was 70% male and 30% female. Among COVID-positive subjects, close to 87% were symptomatic. While for COVID-negative subjects, 86% were healthy without respiratory symptoms or other medical conditions. For the Cambridge set, no age information is provided. The symptom distribution is similar to that of CSS, where 88% of the COVID-positive subjects are symptomatic and 41% of the COVID-negative subjects are with COVID-like symptoms.

The CSS dataset was partitioned into three separate subsets by the challenge organizers, namely training, validation, and test. For comparisons, we employed the same challenge partition in this study. It should be emphasized that in the CSS dataset, several COVID-positive recordings were originally sampled at 8 kHz while the majority of the other files were sampled at 16 kHz. Keeping these up-sampled recordings has been shown to lead to over-optimistic results, thus we have removed them from our analysis, as suggested in [33]. The DiCOVA2 dataset, in turn, is comprised of development and evaluation subsets, with the evaluation data being accessible only to challenge participants. Hence, we performed a speaker-independent training-test split (80/20%) using the development subset only. Lastly, the Cambridge dataset originally contained 1,486 samples from 1,000 subjects. Since it was used only as a blind test set in our experiments, we removed the duplicated users to simulate a real-world setting, thus a total of 1,000 speech samples were used in our experiments. All recordings in the DiCOVA2 and Cambridge sets were sampled at 16 kHz. Table II summarizes the data split and class distributions for all three datasets.

B. Baseline systems and evaluation metrics

To gauge the performance gains obtained with the proposed MTR-CRNN, comparisons with two top-performing systems presented in different COVID-19 detection challenges is performed. For the CCS dataset, the benchmark system is comprised of 6,373 openSMILE features (INTERSPEECH ComParE 2016 format), which are extracted per speech sample, and served as input to a linear kernel SVM classifier. For the DiCOVA2 challenge, in turn, the system relies on mel-spectrograms input to a BiLSTM classification model. Details of these two methods can be found in [6] and [8], respectively. Since most of the employed datasets have imbalanced class distributions, the area under the receiver operating characteristic curve (AUC-ROC) is used as the evaluation metric.

C. Training and inference strategies

Training: Since CSS has a pre-defined training and validation set, hyper-parameters were optimized using the pre-defined validation set. With DiCOVA2, we implemented a 3-fold cross-validation on the training set. This is to best optimize the hyper-parameters with a small data set while maintaining the same test set for direct comparison with other models. The binary cross entropy (BCE) loss and Adam optimizer were used to train our CRNN model. The optimal hyper-parameters used for the final CRNN model are summarized in Table III. In contrast to several existing systems (e.g., [8], [12]), no oversampling or data augmentation techniques were employed with the MTR-CRNN system. This was chosen as one of our goals is to better interpret the features being used by the model and data augmentation methods could bias our findings. As data augmentation has been shown to improve the performance of DNN-based models, the results reported herein could be regarded as a lower bound on what the proposed system could achieve.

Inference: To make full use of the available data, training and validation sets were aggregated and used to train the model from scratch with the optimal hyper-parameters. To avoid over-fitting, the model was trained with a fixed number of epochs which was used to achieve the best validation accuracy during the training phase. As this would lead to different test scores when using different initialized parameters, we report the average test score and standard deviation obtained across 10 random initializations.

TABLE III: Hyper-parameter search ranges and optimal values used. Same parameters were used across tasks.

Hyper-parameter	Search range	Optimal value
Learning rate	($1e^{-5}$, $1e^{-4}$)	$6e^{-5}$
Batch size	(8, 128)	32
Weight decay	($1e^{-5}$, $1e^{-4}$)	$1e^{-4}$

D. Classification task types

Commonly, COVID-19 detection systems report only the within-dataset performance, which has been shown to be often over-optimistic and curtails clinical use [15], [34]. Here, to better objectively evaluate system generalizability across datasets, three different classification tasks are implemented:

Task-1: *Within- and cross-dataset evaluation*. For the within-dataset evaluation, systems are trained and tested using the same dataset. Alternately, for cross-dataset evaluations, systems are trained with the training set of one dataset and blindly tested on the test set of another dataset.

Task-2: *MTR masking*. The goal of this task is to find MTR regions that have a higher potential to generalize across different datasets. The spectral-temporal saliency maps computed from trained CRNN models in Task-1 are used to search for these regions. MTR masks are then applied and the masked representation is then input to the CRNN model for within-dataset and cross-dataset evaluations.

Task-3: *Unseen dataset evaluation*. The primary goal of this task is to explore if the salient MTR regions found in Task-2 remain generalizable when tested on a completely unseen dataset. To this end, we use the Cambridge set as a blind test set. For a fair comparison, all systems (proposed and benchmarks) are trained with the same datasets (i.e., either CSS or DiCOVA2 individually, or the combination of both) and tested on the Cambridge set. This task is designed to simulate a strict setting where no prior knowledge about the test set is known.

IV. Results and discussion

A. Task-1 system performance

The within- and cross-dataset performance of all three tested systems are reported in Table IV. As can be seen, the within-dataset results on CSS show the proposed MTR-CRNN system outperforming even the CSS benchmark, resulting in a final average AUC-ROC of 0.770. On the DiCOVA2 dataset, the obtained results were in line with those obtained from the DiCOVA2-optimized benchmark. Overall, systems achieved a somewhat lower accuracy on the CSS dataset, suggesting that COVID-19 detection with CSS could be a more challenging task. This is likely due to the varied language distribution of the dataset, as well as the higher percentage of asymptomatic COVID-19 samples present in CSS. When tested in the cross-database setting, the proposed MTR-CRNN system showed a substantial improvement relative to the benchmarks. As can be seen, the benchmarks dropped to chance levels when tested on unseen sets. As CSS and DiCOVA varied greatly in demographics (e.g., language, gender, age) and speech content, these results suggest that the proposed method achieved greater generalizability and robustness to unseen data, whilst requiring only 0.1 million more parameters than the DiCOVA2 benchmark. Notwithstanding, the drops seen in accuracy suggest that further improvements may be possible.

TABLE IV: Task-1 performance comparison. Average and standard deviation of AUC-ROC scores are calculated from 10 different initializations. Bold values indicate the highest AUC-ROC. 'DiC' corresponds to DiCOVA2 and 'Param' to number of parameters in the deep learning models.

System	Param	Within-dataset		Cross-dataset	
		CSS	DiC	CSS→DiC	DiC→CSS
CRNN	0.9M	**.770±.019**	.781±.011	**.600±.023**	**.509±.004**
CSS	-	.758±.008	.756±.010	.511±.007	.486±.009
DiC	0.8M	.714±.015	**.789±.016**	.483±.020	.462±.019

B. Task-1 ablation study

To further investigate the role of each module on overall proposed system accuracy, an ablation study is done where the individual 3D CNN and RNN blocks are tested, as well as different temporal pooling schemes and the inclusion of an attention mechanism. For simplicity, only within-database experiments are conducted; results are reported in Table V. As can be seen, for individual blocks, the shallow 2D CNN block shows similar performance as a 9-layer ResNet, suggesting that deeper CNNs may overfit on small datasets. When extending the convolution kernel from 2D to 3D, an improvement is achieved on both datasets. Such improvement is in line with previous findings in video analysis, where the 3D convolution is shown to outperform 2D convolution in terms of capturing the temporal relationships between cascaded images [35]–[37]. Using only the RNN block, in turn, slight improvements relative to the 2D CNN block on DiCOVA2 could be seen, but overall lower accuracy relative to the 3D CNN block was achieved on both datasets.

Next, we investigate the effect of the temporal pooling size of the 3D CNN block. As can be seen from the table, temporal pooling size has an effect on the MTR-CRNN performance, with CSS showing greater variability. A recent study showed the effects of temporal pooling on language identification [38], suggesting that the greater variability seen with CSS could be due to the greater number of languages available in the

TABLE V: Performance of individual blocks and pooling sizes. Same hyper-parameters from Table III are used.

Model	Detail	Within-dataset	
		CSS	DiC
ResNet	9-layer	.691±.023	.683±.015
2D CNN block	3×3 k	.689±.026	.674±.019
3D CNN block	$3 \times 3 \times 3$ k	.725±.017	.721±0.15
RNN block	6×temporal pooling	.671±.018	.696±.022
CRNN	0×temporal pooling	.680±.009	.757±.016
	2×temporal pooling	.702±.015	.753±.007
	6×temporal pooling	**.770±.019**	**.781 ±.011**
	30×temporal pooling	.625±.020	.762±.013
CRNN+Attention	CBAM	.715±.023	.747±.016

database. Temporal pooling layers are known to aggregate information from neighboring time frames, thus changes in pooling strategy alter the size of the temporal receptive field of resultant feature embeddings. In this ablation study, a temporal pooling factor of 6 showed to achieve the best accuracy across both datasets. This optimal configuration downsizes the number of time steps from the initial 150 to 25 after the CNN block, leading to a 6× larger temporal receptive field for each time step. Given that a window size of 256 ms with 216ms-overlap was used to compute the MTR, the resultant temporal receptive field of each time step is 456 ms. Conventional window length used for speech analysis usually ranges from 8 to 32 ms [39], which helps to capture transient changes in speech content. Our findings here, in turn, suggest that speech-based diagnostics could benefit from larger window size and longer-term changes, thus further motivating the use of the MSR/MTR.

Next, we explore the benefits of including an attention mechanism into the proposed system. Rather than adopting the commonly used transformer architecture, a convolutional block attention module (CBAM) is attached to the 3D CNN block [40]. This approach was chosen as it does not require major changes to the model architecture nor the input, while a transformer would require patched features. Results show that adding a CBAM does not lead to performance improvements, likely due to the limited size of training data. Moreover, as CBAM was originally designed for 2D image analysis, it might require careful modification for 3D tensor processing. This investigation is left for a future study to investigate the optimal approach to combine attention mechanism with a 3D MTR.

Lastly, to better understand the decisions made by each block, output embeddings from both the 3D CNN block and the final fully-connected layer of the CRNN model are projected to a 3D space with the maximum variance by performing the principal component analysis (PCA). Figure 6 shows the PCA plots using the training data for both CSS and DiCOVA2 datasets. Using only the embeddings from the 3D CNN block, a small group of COVID-19 samples from CSS can already be distinguished from the non-COVID cluster. For both datasets, it can be visually observed that COVID-19 clusters are better separated using the embeddings from the last layer of CRNN model, which is also reflected by the higher variance achieved with the first three principal components. Taken together, these results suggest that the intermediate representations extracted by the CNN block can provide certain separability between positive and negative COVID-19 speech. This separability is further improved by introducing the temporal dependencies between neighboring MSR snapshots via the RNN block. This temporal dependency further motivates the need for a spectral-temporal saliency map, as explored in Task-2.

C. Task-2 and Task-3 system performance

To better localize the COVID-related regions present in the MTR, Fisher ratio (F-ratio) plots generated from training data from both datasets are compared to each other. For example, the far right and far left plots in Figure 7 show the raw F-ratio plots for CSS and DiCOVA2, respectively, computed directly from the average (raw) MSRs of the two classes. Brighter colors in the plots show modulation spectral regions which better discriminate between the positive and negative COVID-19 classes. When temporal information is considered via Eq. 5, the two middle plots in the figure are obtained. Comparisons between the middle plots suggest that some MSR regions are consistent across datasets, thus suggesting these would be ideal regions for classification.

To avoid measuring energy in individual acoustic-modulation frequency bins, here we propose to group neighboring frequency-frequency bins into patches. We empirically propose patches of shape $\{6 \times 3\}$ where modulation spectral energy values within the patches are summed and min-max normalized. As shown in the figure, two patches, denoted by R1 and R2, are chosen to represent the two most discriminant regions consistently present across the two datasets. R1 corresponds to $f = 650 - 1600$ Hz and $f_m = 5 - 13$ Hz while R2 to $f = 125 - 500$ Hz and $f_m = 3.5 - 10$ Hz. Previous studies have shown that whispered speech is usually manifested at $f < 1$ kHz and $f_m = 5 - 13$ Hz [25], which partly overlaps with the location of R1 and R2. This finding could be linked to an increased level of vocal hoarseness that has been commonly reported with COVID-19 speech, possibly caused by inflammation of the vocal tract area. In turn, the highest F-ratio values for both datasets was found around

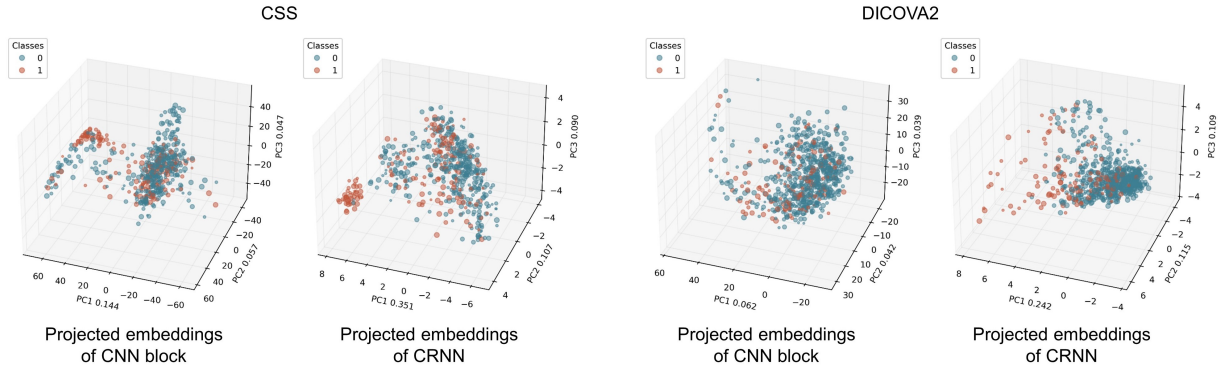


Fig. 6: 3D PCA projection of embeddings extracted from the last layer of CNN block and the CRNN model using training data from CSS and DiCOVA2. Class 0: COVID-negative; Class 1: COVID-positive.

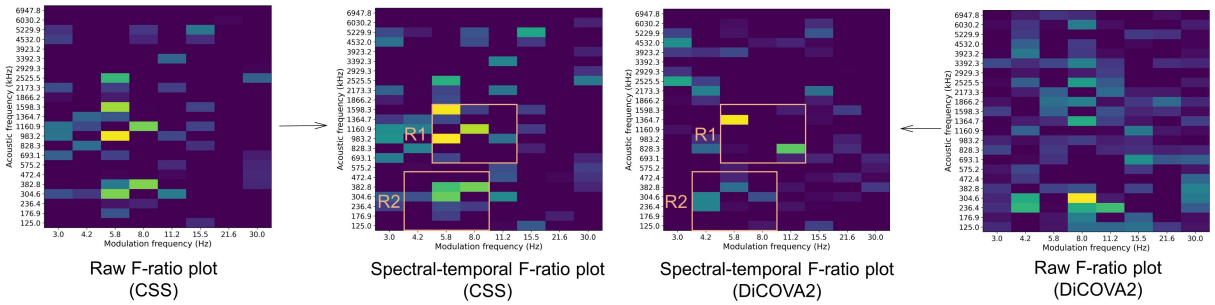


Fig. 7: Discriminative patches found consistently in the spectral-temporal F-ratio plots. Two middle spectral-temporal F-ratio plots are generated with the spectral-temporal saliency maps, while the raw F-ratio plots are generated directly from the raw saliency maps. Brighter areas represent higher discrimination between positive and negative COVID-19 speech samples.

$f = 1.6$ kHz and $f_m = 5 - 6$ Hz. This corroborates previous findings where COVID-19 speech showed more centralized spectral energy at around $f = 2$ kHz [23].

Moreover, a direct comparison of the raw and spectral-temporal F-ratio plots shows that both are almost identical for CSS, but a marked difference can be seen for DiCOVA2. In the raw F-ratio plot of DiCOVA2, the highlighted R1 area in the spectral-temporal plot is barely noticeable. Brighter regions appear more often in the higher acoustic and modulation frequency ranges, which have been linked in the past to represent room acoustic effects, such as reverberation [29]. It is suspected that when the DiCOVA2 data was collected, isolation was still required when testing positive, thus COVID-19 positive speech samples could be affected by e.g., room reverberation, which is more pronounced in enclosed environments. This further shows the importance of aggregating the temporal aspect within the saliency map, thus allowing the model to focus on true COVID-19 discrimination properties and not potential database biases due to e.g., room properties resulting from quarantine isolation.

With these insights and obtained saliency maps, masking is applied to the MTR to extract the two regions only. Task-2 then tests the within and cross-dataset accuracy using the proposed CRNN model with the masked MTRs. Table VI (columns 3-6) reports the accuracy achieved when only R1, only R2, and both R1+R2 regions are used in the mask for Task-2. For comparisons, the original results with the full MTR (as per Table˜IV) are also listed. As can be seen, for the within-dataset test, the original configuration outperformed the masked ones for both datasets. Notwithstanding, the masked version resulted in the highest accuracy in the cross-dataset task, with a substantial margin of improvement relative to the original version. Interestingly, using only R1 resulted in the highest cross-database accuracy, with no benefits seen by adding information from R2.

Lastly, Task-3 provides the most stringent test where a completely unseen test dataset is used. Columns 7-9 show the accuracy achieved when a COVID-19 detection model is trained on only the CSS dataset (and tested on the unseen Cambridge set), only DiCOVA2, and on the combined CSS+DiCOVA2 sets, respectively. As can be seen, accuracy drops for all tested algorithms. All proposed solutions outperform the two benchmarks. In this setting, aggregating information from both R1 and R2 regions showed the greatest accuracy across most conditions. Moreover, increasing the training set size by aggregating two datasets showed some improvement in accuracy, but not substantial. For comparisons, a within-dataset accuracy on the Cambridge set for the CSS and DiCOVA2 benchmarks of .521 and .543 were achieved. As such, our proposed system with patched input outperforms the two benchmarks even in a more stringent testing condition, which shows the robustness of patches found in Task-2 and the generalizability across datasets.

D. Limitations and future work

While the obtained results have been promising, it is important to emphasize that the model can only discriminate between COVID-19 positive and negative, as these were the only labels available in the public datasets. As such, it is not clear if the model is discriminating between healthy and unhealthy participants or COVID-19 itself. As the vocal characteristics of COVID-19 speech found (e.g., vocal hoarseness,

TABLE VI: Performance comparison of different MTR masks. The last column reports AUC-ROC averaged across all tasks. The same hyper-parameters from Table III are used. Bold values indicate the best system for a given task.

System	Input patches	Within-dataset: Task-2		Cross-dataset: Task-2		Unseen dataset: Task-3			Average
		CSS	DiC	CSS→DiC	DiC→CSS	CSS→Cam	DiC→Cam	CSS+DiC→Cam	
MTR-CRNN	R1	.732±.019	.741±.017	**.705±.015**	**.651±.013**	.512±.006	.531±.007	.554±.010	**.632±.091**
	R2	.591±.018	.756±.010	.479±.017	.524±.019	.514±.008	.542±.009	.552±.011	.565±.084
	R1+R2	.656±.015	.775±.017	.602±.010	.558±.016	.538±.009	**.556±.010**	**.560±.007**	.606±.078
	Original	**.770±.019**	.781±.011	.600±.023	.509±.004	**.540±.011**	.541±.007	.543±.008	.612±.106
CSS	-	.758±.008	.756±.010	.511±.007	.486±.009	.504±.006	.489±.009	.506±.006	.572±.117
DiC	-	.714±.015	**.789±.016**	.483±.020	.462±.019	.471±.007	.483±.010	.486±.011	.556±.126

as seen here) could exist in other respiratory diseases, such as asthma or chronic obstructive pulmonary disease, the obtained findings could indeed be more general and represent groups of diseases and not only COVID-19. Initial attempts at addressing this issue have been taken by also including cough sounds in the analysis [41], [42]. Future work should explore the use of speech and cough sounds, as well as integrate other pulmonary diseases within the speech databases. Moreover, the obtained findings may be subject to some inherent data collection biases. As shown in Fig. 5, the positive/negative COVID-19 sample distributions differed across e.g., language, gender, and age groups. As such, the developed models (and implemented benchmarks) may indeed be biased by certain participant demographics rather than the disease itself. As such, a systematic investigation should be performed to test if there are any inherent biases present in the datasets. The authors in [33] already signaled a bias from sample rates in CSS, but there may be other factors.

V. Conclusions

In this paper, we proposed a novel speech-based COVID-19 detection system called MTR-CRNN. The system is based on a 3D modulation tensorgram representation (MTR) combined with a spectro-temporal saliency map mask. Masking enables the system to focus on discriminant COVID-19 regions of the modulation spectrum, across both spectrum and temporal dimensions and bypasses potential database nuances, such as room acoustics. It also allows for greater interpretability of the data serving as input to deep learning models. Experiments on three datasets show the proposed system consistently outperforming two COVID-19 detection challenge top-performing benchmarks on both within-dataset and cross-dataset tasks, as well as on a completely unseen dataset. Together, these findings show that the proposed system is able to generalize well to unseen data and to provide users with a more interpretable and reliable COVID-19 detection solution.

Acknowledgments and disclaimer

The authors would like to thank the developers of the CSS, DiCOVA2, and Cambridge datasets for making them available to the community for research purposes. The developers of the datasets do not bear any responsibility for the analysis and results presented in this paper. All results and interpretations only represent the view of the authors. The authors would also like to thank Prof. Alex Mariakakis and Prof. Eyal de Lara for their insightful discussions and INRS, NSERC, and CIHR for funding this research.

References

[1] A. Scholtz, A. Ramoji, A. Silge, *et al.*, "Covid-19 diagnostics: Past, present, and future," *ACS Photonics*, vol. 8, no. 10, pp. 2827–2838, 2021.

[2] C.-C. Lai, W.-C. Ko, P.-I. Lee, S.-S. Jean, and P.-R. Hsueh, "Extra-respiratory manifestations of covid-19," *International journal of antimicrobial agents*, vol. 56, no. 2, p. 106 024, 2020.

[3] V. K. Paliwal, R. K. Garg, A. Gupta, and N. Tejan, "Neuromuscular presentations in patients with covid-19," *Neurological Sciences*, vol. 41, no. 11, pp. 3039–3056, 2020.

[4] J. Mullol, I. Alobid, F. Mariño-Sánchez, *et al.*, "The loss of smell and taste in the covid-19 outbreak: A tale of many countries," *Current allergy and asthma reports*, vol. 20, no. 10, pp. 1–5, 2020.

[5] T. Greenhalgh, M. Knight, M. Buxton, L. Husain, *et al.*, "Management of post-acute COVID-19 in primary care," *British Medical Journal*, vol. 370, 2020.

[6] B. W. Schuller, A. Batliner, C. Bergler, *et al.*, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.

[7] A. Muguli, L. Pinto, N. Sharma, *et al.*, "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[8] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The second dicova challenge: Dataset and performance analysis for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2110.01177*, 2021.

[9] T. Xia, D. Spathis, J. Ch, *et al.*, "Covid-19 sounds: A large-scale audio dataset for digital respiratory screening," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

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

[11] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 556–560.

[12] 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," *arXiv preprint arXiv:2107.14549*, 2021.

[13] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for covid-19," *arXiv preprint arXiv:2005.08579*, 2020.

[14] M. Benzeghiba, R. De Mori, O. Deroo, *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[15] H. Coppock, L. Jones, I. Kiskin, and B. Schuller, "Covid-19 detection from audio: Seven grains of salt," *The Lancet Digital Health*, vol. 3, no. 9, e537–e538, 2021.

[16] Y. Zhu, A. Mariakakis, E. De Lara, and T. H. Falk, "How generalizable and interpretable are speech-based covid-19 detection systems?: A comparative analysis

and new system proposal," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2022, pp. 1–5.

[17] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech commun.*, vol. 25, no. 1-3, pp. 117–132, 1998.

[18] A. Haridas, R. Marimuthu, and B. Chakraborty, "A novel approach to improve the speech intelligibility using fractional delta-amplitude modulation spectrogram," *Cybern. Systems*, vol. 49, no. 7-8, pp. 421–451, 2018.

[19] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[20] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[21] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, "Objective speech intelligibility measurement for cochlear implant users in complex listening environments," *Speech communication*, vol. 55, no. 7-8, pp. 815–824, 2013.

[22] A. Tiwari, R. Cassani, S. Kshirsagar, D. P. Tobon, Y. Zhu, and T. H. Falk, "Modulation spectral signal representation for quality measurement and enhancement of wearable device data: A technical note," *Sensors*, vol. 22, no. 12, p. 4579, 2022.

[23] Y. Zhu and T. H. Falk, "Fusion of modulation spectral and spectral features with symptom metadata for improved speech-based COVID-19 detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8997–9001.

[24] S. Greenberg and B. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 3, 1997, pp. 1647–1650.

[25] M. Sarria-Paja and T. H. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 783–786, 2013.

[26] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, 1987.

[27] T. K. Dash, S. Mishra, G. Panda, and S. C. Satapathy, "Detection of covid-19 from speech signal using bio-inspired based cepstral features," *Pattern Recognition*, vol. 117, p. 107 999, 2021.

[28] M. Slaney *et al.*, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[29] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, 2009.

[30] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[31] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[32] P. Koutras, G. Panagiotaropoulou, A. Tsiami, and P. Maragos, "Audio-visual temporal saliency modeling validated by fmri data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2000–2010.

[33] H. Coppock, A. Akman, C. Bergler, *et al.*, "A summary of the compare COVID-19 challenges," *arXiv preprint arXiv:2202.08981*, 2022.

[34] M. Roberts, D. Driggs, M. Thorpe, *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.

[35] X. Du, Y. Li, Y. Cui, R. Qian, J. Li, and I. Bello, "Revisiting 3d resnets for video recognition," *arXiv preprint arXiv:2109.01696*, 2021.

[36] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[37] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[38] J. Monteiro, M. J. Alam, and T. Falk, "On the performance of time-pooling strategies for end-to-end spoken language identification," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 3566–3572.

[39] K. K. Paliwal, J. G. Lyons, and K. K. Wójcicki, "Preference for 20-40 ms window duration in speech analysis," in *2010 4th International Conference on Signal Processing and Communication Systems*, IEEE, 2010, pp. 1–4.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[41] A. Imran, I. Posokhova, H. N. Qureshi, *et al.*, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100 378, 2020.

[42] C. Brown, J. Chauhan, A. Grammenos, *et al.*, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," *arXiv preprint arXiv:2006.05919*, 2020.