FUSION OF MODULATION SPECTRAL AND SPECTRAL FEATURES WITH SYMPTOM METADATA FOR IMPROVED SPEECH-BASED COVID-19 DETECTION

Yi Zhu and Tiago H. Falk

Institut national de la recherche scientifique, INRS-EMT, University of Québec, Canada

ABSTRACT

Existing speech-based coronavirus disease 2019 (COVID-19) detection systems provide poor interpretability and limited robustness to unseen data conditions. In this paper, we propose a system to overcome these limitations. In particular, we propose to fuse two different feature modalities with patient metadata in order to capture different properties of the disease. The first feature set is based on modulation spectral properties of speech. The second comprises spectral shape/descriptor features recently used for COVID-19 detection. Lastly, we fuse patient metadata in order to improve robustness and interpretability. Experiments are performed on the 2021 INTERSPEECH COVID Speech Sub-Challenge dataset with several different data partitioning paradigms. Results show the importance of the modulation spectral features. Metadata, in turn, did not perform very well when used alone but provided invaluable insights when fused with the other features. Overall, a system relying on the fusion of all three modalities showed to be robust to unseen conditions and to rely on interpretable features. The simplicity of the model suggests that it can be deployed in portable devices, hence providing accessible COVID-19 diagnostics worldwide.

Index Terms— COVID-19 diagnostics, speech analysis, modulation spectrum, symptom metadata, ComParE

1. INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has disrupted life around the world. Recently, highly contagious variants of the disease are also causing spikes in cases worldwide [1]. Moreover, studies are showing that roughly 10% of individuals who have recovered from COVID-19 will experience prolonged symptoms [2]; these individuals have been termed "long haulers". As such, increasing attention has been given to the development of accurate and automatic diagnostic tools for patient pre-screening, as well as for long-term symptom monitoring [3].

With COVID-19 targeting directly the lungs, several respiratory symptoms have been reported, such as shortness of breath, coughing, and muscle fatigue [4]. Since the respiratory system also plays a key role in speech production [5], speech based diagnostics has shown great potential. Speech based systems have the advantage of being non-invasive, inexpensive, and can provide real-time results [3], something not possible with existing antigen and polymerase chain reaction testing. Moreover, since the majority of the worldwide population already possess a smartphone, speech-based diagnostics can become a worldwide reality with very little need for additional tools.

Existing remote speech-based diagnostic systems typically rely on two main components: (i) acoustic features extracted from the speech signal, and (ii) a downstream machine learning model to make predictions. To date, the most widely used feature set for COVID diagnostics has been the so-called ComParE acoustic feature set. In fact, the ComParE set has been proposed as a benchmark in the recent 2021 INTERSPEECH Computational Paralinguistics ChallengE (ComParE), in particular for the COVID-speech and COVID-cough sub-challenges [6]. The set contains 6,373 spectral features extracted from statistical properties (i.e., functionals) computed over hundreds of low-level descriptors, such as mel-frequency cepstral coefficients (MFCC) and spectral shape properties (e.g., centroid, entropy) [7]. Regarding machine learning strategies, deep neural networks (DNNs) have prevailed (e.g., [8]–[10]). Notwithstanding, 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 [6], [11]. A major limitation of DNN-based models lies in the poor interpretability (i.e., black-box behaviour) they provide, hence resulting in limited clinical use, as well as their poor robustness to unseen data conditions [12].

To tackle these limitations, we propose a remote COVID-19 diagnostic system that fuses three independent feature modalities: (i) an innovative set of features extracted from the modulation spectrum, (ii) the ComParE feature set, and (iii) metadata from COVID patients to provide additional interpretability. The robustness of the fusion-based system is evaluated using data from the INTERSPEECH ComParE COVID-Speech Challenge under different partitioning paradigms. Moreover, an in-depth analysis of the top-ranked features used by the machine learning models allowed us to better understand the underlying characteristics of COVID speech, hence building an interpretable tool that has greater potential for clinical use.

2. FEATURES FOR COVID-19 DETECTION

In this Section, we describe the feature modalities used in the proposed system and the feature selection method used.

2.1. Modulation spectral features

Spectrograms have been widely used as a time-frequency representation of a speech signal. However, environment noise can overlap in both time and frequency, thus making spectrograms sub-optimal. The modulation spectrogram, in turn, captures the rate-of-change of frequency components thus becomes favored for in-the-wild speech analysis [13]. Previous work has shown the efficacy of the modulation spectrogram for speech quality assessment [14], speech emotion detection [15], and speech intelligibility assessment [16]. Here, we propose their use for COVID-19 detection.

The steps to compute the modulation spectrogram, as well as the proposed modulation spectral features (MSFs), 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 (via 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 (kHz) and f_m the modulation frequency (Hz). As the majority of the modulation spectral content of speech is known to lie below 20 Hz f_m [14], parameters used in the computation of the modulation spectrogram have been chosen as follows: $f_m = 0-20$ Hz, $f = 0-8$ kHz, 32 ms window length, and 16 ms hop length. The modulation spectrogram is then quantized 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, resulting in $\hat{X}(f_m, f)$. The normalized modulation energies from each of these 400 bins are used as modulation spectral energy features. Moreover, eight spectral shape descriptors were computed across the frequency axis as well as across the modulation frequency axis, including centroid, entropy, spread, skewness, kurtosis, flatness, crest, and flux. A detailed description of these descriptors can be found in [15]. Overall, a total of 720 MSFs are computed ($400 + 8 \times 20 + 8 \times 20$).

2.2. ComParE (benchmark) features

The ComParE set contains 6,373 features computed from various conventional speech representations, such as spectrograms, mel-spectrograms, as well as pitch and voicing related parameters and cepstral parameters using the openSMILE toolbox [17]. These features have been used widely by the speech community, including for COVID-19 detection [6]. They are included here to test their complementarity to the proposed MSFs. The interested reader is referred to [6] for more details about the ComParE features.

2.3. Metadata

While modulation/spectral features capture the changes to the respiratory system and articulators, metadata provides information about the patient that has been shown to also be useful for COVID-19 detection [18]. Here, metadata is divided into eight categories: age, gender, medical history, symptoms, smoking history, language, hospitalization, and the most recent COVID testing result. Only four categories directly associated with COVID-19 infection and symptoms are used, namely medical history, symptom, hospitalization, and smoking history. The other four were omitted to avoid introduction of any potential biases to the models. Metadata were converted into 25 numeric variables using a dummy encoding scheme.

2.4. Feature selection

To reduce the number of features to be input to a machine learning model, the Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithm was used. MRMR finds the optimal feature subset, such that the top-ranked features are mutually dissimilar, while being maximally related to the COVID diagnostic [19]. As a filter based method, MRMR selects features without relying on downstream machine learning models, thus, allows for assessment of the robustness of the top-ranked features. In this study, the optimal number of features to be selected was decided empirically based on cross-validation analysis.

3. EXPERIMENTAL SETUP

Here, we describe the datasets, benchmark, and figures-of-merit used, as well as the three different tasks explored.

3.1. Database description

In this study, the database provided by the COVID Speech Sub-challenge (CSS) of the INTERSPEECH 2021 ComParE challenge [6] was used. The CSS database is comprised of three disjoint datasets: (i) training set (315 recordings), (ii) validation set (295 recordings), and (iii) test set (283 recordings), resulting in a total of 893 recordings. Participants were asked to utter the sentence "I hope my data can help to manage the virus pandemic" 1-3 times in their mother tongue. From these sets, it was observed that some files had been upsampled from 8 kHz, thus had different spectral profiles between 4-8 kHz. To avoid any biased results, we omitted such files from our experiments. Table 1 summarizes the distribution of the recordings per class for each set. For convenience, henceforth these three sets are termed Sets 1-3.

3.2. Benchmark system and figures-of-merit

The CSS Challenge benchmark system utilized ComParE acoustic features as inputs to a linear SVM classifer. For the purpose of fair comparison, the proposed system is based on

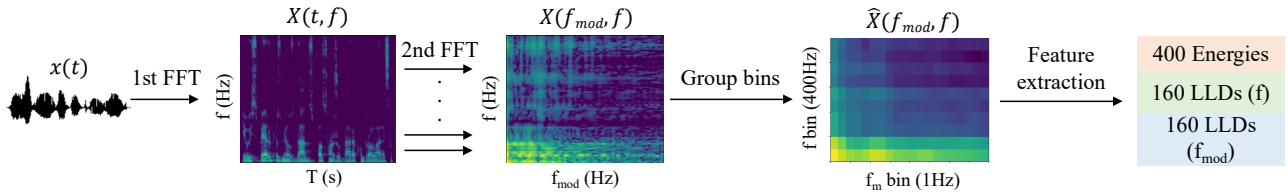


Fig. 1: Steps involved in computing modulation spectral features

Table 1: Partitioning of the "unbiased" database where upsampled files were omitted.

Partition	Positive	Negative
Set-1 (training set)	56	243
Set-2 (validation set)	130	153
Set-3 (test set)	87	189

the same linear SVM following the same hyper-parameter tuning strategy described in [6]. Since the CSS database is imbalanced, three metrics were used to evaluate model performance, including the unweighted average recall (UAR), true positive rate (TPR) and true negative rate (TNR). It is important to emphasize that [6] reports the benchmark system achieving an UAR of 0.58 on Set-2 (original validation set) and 0.72 on Set-3 (original test set). These results take into consideration the potentially-biased upsampled files mentioned above. After removing such files, it is observed that the UAR drops to 0.54 on Set-2 and 0.60 on Set-3. Henceforth, we will use these UAR scores in our comparisons.

3.3. Tasks

Several tasks are performed to test system performance:
Task-1: Aims at evaluating the *individual* performance of the three feature sets. Each model is trained on Set-1 and tested on Set-2 and Set-3 separately.
Task-2: Aims at evaluating the performance of the models after *feature fusion* and feature selection with MRMR. Here, the fused set of dimensionality 7,118 $(6,373 + 720 + 25)$ is reduced to the 35 top features, found empirically to strike a balance between accuracy and complexity. As before, models are trained on Set-1 and tested on Set-2 and Set-3 separately.
Task-3: Aims at quantifying the complementarity of the three different feature sets, as well as the robustness of the models to unseen data. To this end, three schemes are used: (1) training with Set-1 and Set-2 testing with Set-3, (2) training with Set-1 and Set-3, testing with Set-2, and (3) training with Set-2 and Set-3, testing with Set-1. In all feature fusion schemes, the top-35 features from each category are fused together.

4. RESULTS AND DISCUSSION

In this Section, we describe and discuss the results obtained.

4.1. Classification results

To evaluate the performance of proposed features, Table 2 reports the UAR, TPR, and TNR achieved with each feature modality. Results with the different number of features are reported to allow for fair comparisons among feature sets. As can be seen, the accuracy of the tested features varied based on the test set used. While MSFs were shown to be the best for test Set-2, the ComParE benchmark features showed improved accuracy on test Set-3. When comparing results achieved with the same number of features (i.e., 720 or 35), in turn, the MSFs outperformed the benchmark on test Set-2 and achieved comparable results on test Set-3. On the other hand, feature fusion showed stable accuracy for both test sets. These results are important, as they not only show the robustness of the proposed system to unseen data, but also show that reliable accuracy can be achieved with as little as 35 features. This can be important for interpretability, as well as for running diagnostic systems directly from (potentially low-cost) portable devices, hence accessible to all.

Next, we explore the robustness of different feature fusion schemes and unseen dataset combinations to further validate these findings. To this end, the feature categories are combined two-by-two and then all together, where the top 35 features from each category are fused. Table 3 shows the obtained results across the three schemes described in Section 3.3. As can be seen, the fusion-based system showed the most stable accuracy across different test sets, hence suggesting its robustness to unseen data. Moreover, including metadata showed to improve accuracy over using modulation spectral/spectral features alone, hence showing the importance of symptomatic metadata for improved COVID-19 diagnosis.

4.2. Feature interpretation

To improve the interpretability of the top-ranked features, we compare the importance of the top-35 features across all three testing schemes. It is noted that three features consistently appeared in the top-10 list for each scheme. These included 'spectral kurtosis across modulation frequencies at $f = 2 - 2.4$kHz' (MSF), 'pcm_fftmag_spectral entropy' (ComParE), and 'smell and taste loss' (metadata). To further interpret the selected modulation spectral descriptor, the modulation spectrograms, normalized and averaged over all speech files, are depicted in Fig. 2 for COVID and non-COVID speech.

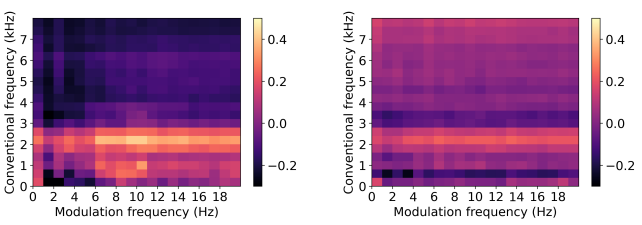
As can be seen, for COVID speech a unique pattern

Table 2: Performance comparison per feature set and feature fusion. Bold values indicate best system for a given metric.

Set #	Feature	# Feats	UAR	TPR	TNR
2	ComParE	6,373	.547	.149	**.945**
		720	.545	.175	.915
		35	.546	.177	.915
	MSF	720	**.647**	**.439**	.856
		35	.647	.385	.906
	Metadata	25	.549	.215	.882
	Fused	35	.610	.331	.889
3	ComParE	6,373	.604	.287	.921
		720	.580	.331	.826
		35	.610	.318	.896
	MSF	720	.570	.241	.900
		35	.550	.161	**.942**
	Metadata	25	.557	.310	.804
	Fused	35	**.614**	**.333**	.894

Table 3: Performance comparison for different feature combinations and testing schemes. Bold values indicate best system for a given metric.

Scheme	Features	UAR	TPR	TNR
1	ComParE+MSF	.602	.310	.894
	ComParE+Metadata	.640	.310	**.974**
	MSF+Metadata	.622	.345	.900
	Fused (all)	**.643**	**.356**	.931
2	ComParE+MSF	.592	.269	**.915**
	ComParE+Metadata	.769	.631	.909
	MSF+Metadata	.783	.685	.882
	Fused (all)	**.799**	**.723**	.876
3	ComParE+MSF	**.815**	**.750**	.881
	ComParE+Metadata	.761	.571	.951
	MSF+Metadata	.680	.393	**.967**
	Fused (all)	.782	.625	.938



(a) (b)

Fig. 2: Average modulation spectrograms for (a) COVID speech, and (b) non-COVID speech

is found at $f < 1.2$ kHz and $f_m = 6 - 14$ Hz, and $f = 2 - 2.4$ kHz. Part of these discriminative regions found are in line with the selected modulation spectral descriptor, hence validating the importance of the selected MSF. Additionally, previous work on whispered speech showed that whispers can be manifested below $f = 1$kHz and $f_m > 10$Hz [16]. These findings suggest that the changes to MSFs could be due to increasing vocal hoarseness of COVID patients, possibly caused by inflammation of the respiratory system. Additionally, existing literature has shown that an increase in spectral entropy is seen in muffled speech compared with clean speech [20], as well as in reduced speech quality due to poor manner of articulation [21]. As the same pattern is observed here for COVID speech, it could be associated with the difficulty in articulation and oral-facial muscle fatigue in COVID-19 [22], as well as nasal congestion causing speech to sound e.g., more muffled. Lastly, clinical studies have shown that smell dysfunction is one of the first symptoms of a COVID-19 infection [23] and a useful predictor [18], thus corroborating our findings. Interestingly, in [24], loss of taste/smell, nasal congestion, and facial pain were reported in the majority of COVID-19 patients. The proposed features seem to be measuring these parameters quantitatively and in a complementary manner.

5. CONCLUSION

In this paper, we have proposed a speech-based COVID-19 detection system based on the fusion of three different feature sets: a novel set based on modulation spectral features, a benchmark set based on spectral descriptors recently proposed for COVID-19 detection, and symptom metadata. A number of experiments are performed to test the robustness of the different features sets and their combinations on unseen data. The novel modulation feature sets are shown to provide complementary information to existing features and, most importantly, when combined with benchmark spectral features and symptom metadata, robust accuracy was shown across varying dataset partitions. The overall simplicity of the proposed system, both in terms of the number of features used and their interpretability, suggests the proposed system could be a good candidate for remote COVID-19 detection that is accessible to all around the world.

6. ACKNOWLEDGEMENT AND DISCLAIMER

The authors would like to acknowledge the University of Cambridge for sharing the COVID-19 speech database and for the INTERSPEECH 2021 ComParE Challenge organizers. The University of Cambridge does 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 also acknowledge INRS for funding.

7. REFERENCES

[1] D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M. M. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, *et al.*, "Reduced sensitivity of SARS-CoV-2 variant delta to antibody neutralization," *Nature*, vol. 596, no. 7871, pp. 276–280, 2021.

[2] 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.

[3] G. Deshpande and B. W. Schuller, "Audio, speech, language, & signal processing for COVID-19: A comprehensive overview," *arXiv:2011.14445*, 2020.

[4] P. Vetter, D. L. Vu, A. G. L'Huillier, M. Schibler, L. Kaiser, and F. Jacquerioz, "Clinical features of COVID-19," *British Medical Journal*, vol. 369, 2020.

[5] P. F. Macneilage, "Speech production," *Language and Speech*, vol. 23, no. 1, pp. 3–23, 1980.

[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:2102.13468*, 2021.

[7] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

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

[9] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, and V. Aharonson, "SARS-CoV-2 detection from voice," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 268–274, 2020.

[10] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, p. 104 572, 2021.

[11] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 8328–8332.

[12] 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.

[13] T. H. Falk, W.-Y. Chan, E. Sejdic, and T. Chau, "Spectro-temporal analysis of auscultatory sounds," in. InTech, 2010, pp. 93–104.

[14] 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.

[15] 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.

[16] T. H. 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 Communication*, vol. 54, no. 5, pp. 622–631, 2012.

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

[18] C. Menni, A. M. Valdes, M. B. Freidin, *et al.*, "Real-time tracking of self-reported symptoms to predict potential COVID-19," *Nature medicine*, vol. 26, no. 7, pp. 1037–1040, 2020.

[19] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–14, 2017.

[20] S. A. Memon, "Acoustic Correlates of the Voice Qualifiers: A survey," *arXiv:2010.15869*, 2020.

[21] F. Llanos, J. M. Alexander, C. E. Stilp, and K. R. Kluender, "Power spectral entropy as an information-theoretic correlate of manner of articulation in american english," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, EL127–EL133, 2017.

[22] J. Helms, S. Kremer, H. Merdji, *et al.*, "Neurologic features in severe SARS-CoV-2 infection," *New England Journal of Medicine*, vol. 382, no. 23, pp. 2268–2270, 2020.

[23] S. T. Moein, S. M. Hashemian, B. Mansourafshar, A. Khorram-Tousi, P. Tabarsi, and R. L. Doty, "Smell dysfunction: A biomarker for COVID-19," in *International forum of allergy & rhinology*, Wiley Online Library, vol. 10, 2020, pp. 944–950.

[24] M. A. Callejon-Leblic, R. Moreno-Luna, A. Del Cuvillo, *et al.*, "Loss of smell and taste can accurately predict COVID-19 infection: A machine-learning approach," *Journal of Clinical Medicine*, vol. 10, no. 4, p. 570, 2021.