A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/OJEMB.2020.2998051, IEEE Open Journal of Engineering in Medicine and Biology

IEEE OJEMB TECHNOLOGY LETTER

A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems

Thomas F. Quatieri, *Fellow,* Tanya Talkar, *Membe*r, and Jeffrey S. Palmer, *Senior Member IEEE*[1]

Abstract—

Goal: **We propose a speech modeling and signal-processing framework to detect and track COVID-19 through asymptomatic and symptomatic stages.**

Methods: **The approach is based on complexity of neuromotor coordination across speech subsystems involved in respiration, phonation and articulation, motivated by the distinct nature of COVID-19 involving lower (i.e., bronchial tubes, diaphragm, lower trachea) versus upper (i.e., laryngeal, pharyngeal, oral and nasal) respiratory tract inflammation [1], as well as by the growing evidence of the virus' neurological manifestations [2]–[5].**

Preliminary results: **An exploratory study with audio interviews of five subjects provides Cohen's d effect sizes between pre-COVID-19 (pre-exposure) from post-COVID-19 (after positive diagnosis but asymptomatic) using: coordination of respiration (as measured through acoustic waveform amplitude) and laryngeal motion (fundamental frequency and cepstral peak prominence), and coordination of laryngeal and articulatory (formant center frequencies) motion.**

Conclusions: **While there is a strong subject-dependence, the group-level morphology of effect sizes indicates a reduced complexity of subsystem coordination. Validation is needed with larger more controlled datasets and to address confounding influences such as different recording conditions, unbalanced data quantities, and changes in underlying vocal status from pre-to-post time recordings.**

Index Terms—**asymptomatic, COVID-19, respiration, vocal subsystems, motor coordination**

Impact Statement— ***The proposed sensing lends itself to nonintrusive widespread use through mobile devices. Thus, the approach provides a key capability for scalable, longitudinal studies that seek to capture human behavior dynamics in naturalistic environments for early warning and tracking of COVID-19.***

INTRODUCTION

COVID-19 is often characterized by specific dysfunction in respiratory physiology including the diaphragm and other parts of the lower respiratory tract, thereby affecting patterns of breathing during inhalation and exhalation of air from the lungs [1]. In speech production, during the exhalation stage, air from the lungs moves through the other essential vocal subsystems, i.e., through the trachea and larynx and into the vocal tract pharyngeal, oral and nasal cavities (Fig. 1). The manner in which we breathe in speaking, including the rate and length of an exhalation (coupled to the number of words in a phrase or sentence), and its intensity and variability, highly influences the quality of our voice. For example, the loudness, aspiration ("breathiness"), steadiness of fundamental frequency or "pitch" during phonation, and the mechanism by which we alter speaking rate all effect vocal quality. Furthermore, the respiratory system is highly coordinated with these primarily laryngeal-based subsystems [6][7] . Likewise, in turn, laryngeal activity is finely coupled to articulation in the oral and nasal cavities [8]. Although impact on speech subsystems and their coordination are often perceptually obvious with a condition involving inflammation, these changes can be subtle in the asymptomatic stages of an illness, either at onset or in recovery.

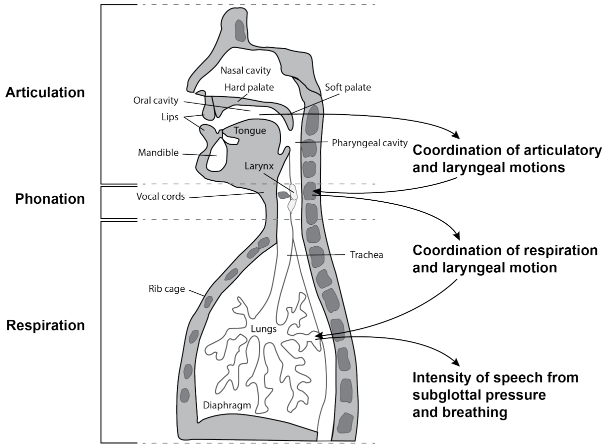

Fig. 1: Speech subsystems and their coordination hypothesized to be affected by COVID-19.

In addition to the physiological manifestations of COVID-19, recently there have been reported symptoms that relate to temporary neuromuscular impairments [2][3] and loss of taste and smell [4][5] which can have implications for muscular

[1] *DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.*

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering. © 2020 Massachusetts Institute of Technology. Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

All three authors are affiliated with MIT Lincoln Laboratory. Thomas F Quatieri and Tanya Talkar also have affiliation with the Harvard-MIT Speech and Hearing Bioscience and Technology Program, Harvard Medical Sciences. Tanya Talkar was supported in part by NIH-NIDCD T32 DC000038. There are no conflict of interests.

control and proprioceptive feedback, respectively, in speech production. Given the physiologically-based insult to breathing functions, as well as this growing evidence of neurological deficits present in COVID-19, we hypothesize that biomarkers derived from measures of vocal subsystem coordination may provide a sensitive indicator of COVID-19, most importantly in its asymptomatic stages.

I. MATERIALS AND METHODS

A. Dataset

Audio data for five subjects was obtained from YouTube, Instagram, and Twitter sources: pre-COVID-19 (before exposure) and post-COVID-19 (after positive but asymptomatic). Subject-only regions were segmented manually from the videos to exclude secondary speakers such as interviewers or interviewees and other interferences. The recordings are taken from press conferences and TV interviews all with celebrities or broadcast hosts, typically using high-quality recording facilities. Though consistent environment and high signal quality were sought across pre- and post-states, the data can have varying environmental and recording conditions. Post-recording times were in the range of days with pre-recording times in the range of days-to-years. Signal-to-noise ratios were fairly high and consistent across pre- and post-conditions, ranging per-subject from about 18 to 10 dB. The Supplementary Material Section provides subject-specific and group statistics on segment durations and counts, pre- and post-recording times and environmental noise conditions.

B. Subsystem model

Our speech feature selection is based on the physiologically-motivated speech production model in Fig. 2 where the airflow from the lungs during the exhalation phase of speech production passes through the bronchial tubes through the trachea and into the larynx. The 'intensity' of the airflow (velocity), that we refer to in this note as the *respiratory intensity,* governs time-varying loudness, and is coupled (coordinated) with phonation, i.e., the vibration of the vocal folds (fundamental frequency or 'pitch'), stability of phonation, and aspiration at the folds [7] all which are a function of laryngeal muscles and tissue, modulated by the respiratory intensity. Finally, in our model, the vocal fold source signal is modulated by, and coordinated with, the vocal tract movement during articulation.

C. Feature extraction

Standard *low-level* features associated with the various subsystems of Fig. 1 and Fig. 2 form the basis of our *high-level* features representing the coordination within and across these various subsystems and have been shown in previous research to be predictive of numerous neurocognitive conditions [9][10][11].

Low-level univariate features characterize basic properties of the three vocal subsystem components. The speech envelope is used as a proxy for respiratory intensity and is estimated using an iterative time-domain signal envelope estimation algorithm, providing a smooth contour derived from amplitude peaks [12][13]. At the laryngeal level, we estimate the fundamental frequency (pitch) using an autocorrelation approach [14][15] and cepstral peak prominence (CPP), which provides stability of vocal fold vibration [16]. CPP is based on the ratio of the pitch-related cepstral peak relative to aspiration noise level and is a widely used and robust measure for assessing pathological speech [17][18].

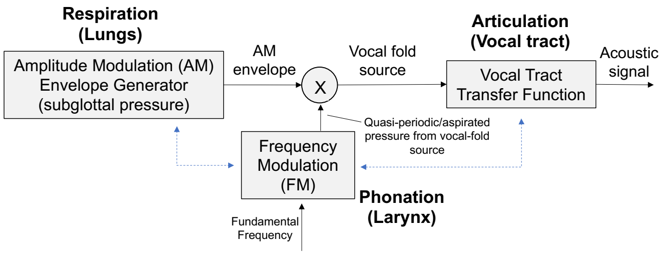
Respiration (Lungs)
Amplitude Modulation (AM) Envelope Generator (subglottal pressure)
AM envelope
X
Vocal fold source
Articulation (Vocal tract)
Vocal Tract Transfer Function
Acoustic signal
Quasi-periodic/aspirated pressure from vocal-fold source
Frequency Modulation (FM)
Phonation (Larynx)
Fundamental Frequency

Fig. 2: Fundamental speech-production subsystem model illustrating two of the potential points of coordination (dashed blue).

As a measure of vocal fold stability, CPP has the potential to reflect change in coupling of subglottal respiratory and laryngeal subsystems with COVID-19. Finally, formant center frequency tracks (vocal tract resonance frequencies), used as a proxy for articulation, relies on a robust Kalman-filter-based tracking algorithm [19]. Features are computed only during speaking using a speech activity detector [19].

High-level features involve multivariate auto- and cross-correlations of low-level features to produce measures of coordination within and across the underlying mechanisms of speech subsystems. Correlation functions for each (subject-only) segment are sampled at a time-delay scale of 10 ms. Within a segment, masking is applied to exclude speech pauses in computing correlations. The eigenspectra of each *correlation matrix*, formed from various sets of samples from correlation functions, quantifies and summarizes the frequency properties of the set of feature trajectories (see Supplemental Material).

Higher *complexity* across multiple channels is reflected in a more uniform distribution of eigenvalues, and more independent "modes" of the underlying system components, while lower complexity is reflected in a larger proportion of the overall signal variability being concentrated in a small number of eigenvalues. In the latter case, the eigenspectral concentration typically manifests with high-rank eigenvalues being lower in amplitude and thus reflecting more dependent or 'coupled' system components. For the five subjects, independently and combined, the Cohen's d effect sizes pre-versus post-COVID-19 were computed (for all segments in each category) based on the eigenspectra for low-level respiration intensity, fundamental frequency, cepstral peak prominence, and formant center frequencies. More details about computing low- and high-level coordination features, effect sizes, and their interpretation are provided in the Supplementary Materials section.

II. Results

The example given in Fig. 3, with pre-COVID-19 and post-COVID-19 (asymptomatic) conditions, shows Cohen's d effect sizes with three measures of coordination: respiration and pitch (fundamental frequency), respiration and stability of pitch periodicity (CPP), and pitch (fundamental frequency) and articulation (formant center frequencies). Effect size patterns for the two cases involving respiration show similar high-to-low trends across many of the subjects, with high-rank eigenvalues tending toward relatively lower energy for the post-COVID-19 cases.

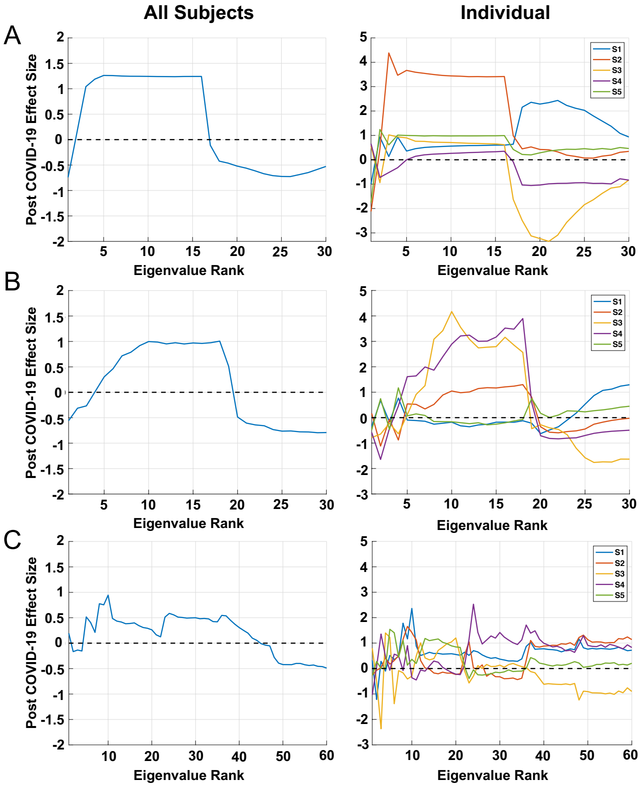

Fig. 3: COVID-19 influence on coordination of respiration (as measured through the speech envelope) and laryngeal characteristics as well as coordination of respiration with articulation (as measured through vocal tract formant center frequencies). Group and subject-dependent effect sizes are shown in left and right columns, respectively. (A): respiration intensity and pitch (fundamental frequency) (30 eigenvalues obtained from 2 features x 15 correlation samples); (B): respiration intensity and stability of periodicity (CPP) (30 eigenvalues obtained from 2 features x 15 correlation samples); (C): articulation (3 formant center frequencies) and pitch (fundamental frequency) (60 eigenvalues obtained from 4 features x 15 correlation samples). Effect sizes greater in magnitude than 0.37 in the comparison across all subjects have corresponding p < 0.05.

Effect sizes for the combined subjects indicate a similar but more distinct group-level morphology in these cases. On the other hand, effect sizes for coordination of pitch (fundamental frequency) and articulation (formant center frequencies) is more variable across subjects, but the combined counterpart shows a high-to-low trend, albeit weaker than those involving coordination involving respiration. Although a strict interpretation is not possible due to the small cohort, at a group level, the morphology of effect sizes in Fig. 3 indicates a reduction in the complexity of coordinated subsystem movement, in the sense of less independence of coordinated respiratory and laryngeal motion and likewise, but to a lesser extent, for laryngeal and articulatory motion.

III. Discussion

Although the group-level eigenspectra-based effect size trends indicate a reduced complexity in coordination, clearly a larger cohort is warranted as well as addressing a number of confounders, including subject- and recording-dependences, in any validation procedure. For example, across all variables, inter-subject analysis shows for some subjects a distinctly different trend of larger high-ranking eigenvalues that indicates a more complex but more erratic (or variable) coordination. Regarding signal quality, due to the nature of the online video sources, there is a variety of inter- and intra- subject recording variability, the most perceptually notable effect being reverberation, possibly modifying the true effect sizes, over- or underestimating their importance. An example given in the Supplementary Material, isolating two of the subjects with more consistent, least reveberant environments, enhances the combined effect sizes relative to the N=5 case.

IV. Conclusion

We have established a framework for discovery of vocal biomarkers of COVID-19 based on the coordination of subsystems of speech production involving respiration, phonation, and articulation. Our preliminary results, using a very limited data set, hint at support of the hypothesis that biomarkers derived from measures of vocal subsystem coordination provide an indicator of COVID-19 impact on respiratory function, particularly in its asymptomatic stage. Given a sample size of five subjects, however, validation of our hypothesis will clearly require additional data and analysis to address potential confounders such as different recording environments and channels, unbalanced data quantities, and changes in underlying vocal status from pre-to-post time recordings. It will also be important to expand the suite of vocal features, introducing neurophysiological modeling of subsystem interactions, to address the increasing evidence of neurological insult arising from COVID-19 and feature specificity relative to typical flu and flu-like symptoms.

V. Supplemental Material

The Supplementary Material section provides the following expansions of the main body topics: (1) more detailed description of the physiological motivation for the coordination model of Fig. 1 and Fig. 2; (2) more details of our standard low-level feature extraction, as well as introducing other vocal source features such as harmonic-to-noise ratio, vocal creak, and glottal open quotient; (3) effect sizes of summary statistics of the low-level features across the pre- and post-COVID-19 conditions as a comparative reference to the high-level feature effect sizes; (4) further description of the subject- and sessiondependent environmental conditions; (5) more detailed description of the correlation methodology; and (6) expanded algorithm descriptions and software references to expedite use by others in the field.

ACKNOWLEDGMENT

The authors thank Drs. Brian Telfer, Timothy Dasey, William Streilein, and Christopher Smalt for discussions and references on the nature of COVID-19 and various data sources, and Drs. Daryush Mehta and James Williamson for use of their software packages in computing certain low-and high-level features, respectively. We also acknowledge Eric Palmer for the initial YouTube sourcing.

REFERENCES

[1] World Health Organization, “Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020,” 2020. Accessed: May 02, 2020. [Online]. Available: https://www.who.int/.

[2] L. Mao *et al.*, “Neurological Manifestations of Hospitalized Patients with COVID-19 in Wuhan, China: A Retrospective Case Series Study,” 2020. doi: 10.1101/2020.02.22.20026500.

[3] J. Helms *et al.*, “Neurologic Features in Severe SARS-CoV-2 Infection,” *N. Engl. J. Med.*, Apr. 2020, doi: 10.1056/nejmc2008597.

[4] C. H. Yan, F. Faraji, D. P. Prajapati, C. E. Boone, and A. S. DeConde, “Association of chemosensory dysfunction and Covid-19 in patients presenting with influenza-like symptoms,” *Int. Forum Allergy Rhinol.*, Apr. 2020, doi: 10.1002/alr.22579.

[5] J. F. Gautier and Y. Ravussin, “A New Symptom of COVID-19: Loss of Taste and Smell,” *Obesity*, vol. 28, no. 5. Blackwell Publishing Inc., p. 848, May 01, 2020, doi: 10.1002/oby.22809.

[6] Z. Zhang, “Respiratory Laryngeal Coordination in Airflow Conservation and Reduction of Respiratory Effort of Phonation,” *J. Voice*, vol. 30, no. 6, pp. 760.e7-760.e13, Nov. 2016, doi: 10.1016/j.jvoice.2015.09.015.

[7] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins, “Relationship between changes in voice pitch and loudness,” *J. Voice*, vol. 2, no. 2, pp. 118–126, Jan. 1988, doi: 10.1016/S0892-1997(88)80067-5.

[8] V. L. Gracco and A. Löfqvist, “Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements,” *J. Neurosci.*, vol. 14, no. 11 I, pp. 6585–6597, Nov. 1994, doi: 10.1523/jneurosci.14-11-06585.1994.

[9] T. F. Quatieri, J. Williamson, C. Smalt, J. Perricone, T. Patel, L. Brattain, B. Helfer, D. Mehta, J. Palmer, K. Heaton, and M. E. Moran., “Multimodal Biomarkers to Discriminate Cognitive State,” in *The Role of Technology in Clinical Neuropsychology*, R. L. Kane and T. D. Parsons, Eds. Oxford University Press, 2017, p. 409.

[10] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, “Tracking depression severity from audio and video based on speech articulatory coordination,” *Comput. Speech Lang.*, vol. 55, pp. 40–56, 2019, doi: 10.1016/j.csl.2018.08.004.

[11] T. Talkar, J. Williamson, D. Hannon, H. Rao, S.Yuditskaya, D. Sturim, K. Claypool, L. Nowinski, H. Saro, C. Stamm, M. Mody, C. McDougle, and T.F. Quatieri, “Assessment of Speech Motor Coordination in Children with Autism Spectrum Disorder,” in *Oral Presentation at Motor Speech Conference, Santa Barbara, CA*, 2020.

[12] A. Röbel and X. Rodet, “Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation,” in *8th International Conference on Digital Audio Effects, DAFx 2005 - Proceedings*, 2005, pp. 30–35, Accessed: May 02, 2020. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01161334.

[13] R. L. Horwitz-Martin, T. F. Quatieri, E. Godoy, and J. R. Williamson, “A vocal modulation model with application to predicting depression severity,” in *BSN 2016 - 13th Annual Body Sensor Networks Conference*, 2016, pp. 247–253, doi: 10.1109/BSN.2016.7516268.

[14] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer [Computer program],” 2018. http://www.praat.org.

[15] P. Boersma, “Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound,” *Proc. Inst. Phonetic Sci.*, vol. 17, pp. 97–110, 1993.

[16] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, “Acoustic Correlates of Breathy Vocal Quality,” *J. Speech, Lang. Hear. Res.*, vol. 37, no. 4, pp. 769–778, Aug. 1994, doi: 10.1044/jshr.3704.769.

[17] Y. D. Heman-Ackah, D. D. Michael, and G. S. Goding, “The relationship between cepstral peak prominence and selected parameters of dysphonia,” *J. Voice*, vol. 16, no. 1, pp. 20–27, Mar. 2002, doi: 10.1016/S0892-1997(02)00067-X.

[18] R. Fraile and J. I. Godino-Llorente, “Cepstral peak prominence: A comprehensive analysis,” *Biomed. Signal Process. Control*, vol. 14, pp. 42–54, 2014, doi: https://doi.org/10.1016/j.bspc.2014.07.001.

[19] D. D. Mehta, D. Rudoy, and P. J. Wolfe, “Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking,” *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, Sep. 2012, doi: 10.1121/1.4739462.