2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

COVID-19 Biomarkers in Speech: On Source and Filter Components

Gauri Deshpande[1] and Björn W. Schuller[2]

Abstract— **This paper analyses the source of excitation and vocal tract influenced filter components to identify the biomarkers of COVID-19 in the human speech signal. The source-filter separated components of cough and breathing sounds collected from healthy and COVID-19 positive subjects are also analyzed. The source-filter separation techniques using cepstral, and phase domain approaches are compared and validated by using them in a neural network for the detection of COVID-19 positive subjects. A comparative analysis of the performance exhibited by vowels, cough, and breathing sounds is also presented. We use the public Coswara database for the reproducibility of our findings.**

I. INTRODUCTION

As the source of excitation and vocal tract filter parameters govern the speech production mechanism, it is critical to understand their effects on the produced sound. For the production of vowels and voiced consonants, the quasi-periodic glottal pulses are the source of excitation, whereas, for cough, it is high-velocity expiration from the lungs [1]. The unvoiced consonants and breathing sounds also originate from the lungs.

The speech 'source component' analysis is used in numerous applications such as in [2] for enhancing the quality of speech from multiple microphones, in [3] for speaker localisation, in [4] for detecting the number of distinct speakers, in [5] for detecting the perceived loudness of speech, in [6] for audio clip classification, and in [7], [8] for emotion recognition, using the excitation source information from Linear Prediction (LP) residuals. The LP residual is the minimum error signal calculated as the difference between the speech sample and its predicted value obtained from linear prediction analysis. Similarly, the vocal tract parameters are used for classifying the speech into a low, medium, and high cognitive load in [9] and for improving the speech recognition performance in [10]. The authors of [11] present several efforts for collecting COVID-19 (C19) patients audio data; however, a tiny percentage of it is on analysing the data. In [12], the authors have collected speech from C19 patients and have performed classification of the data from four aspects: severity of illness, sleep quality, fatigue, and anxiety. In another study presented in [13], the authors have analyzed the group correlation of Mel-Frequency Cepstral Coefficients (MFCCs) from the speech of C19 and healthy individuals. The dynamics of the glottal flow waveform are examined in [14] to identify relevant parameters for detecting C19 from voice.

To understand the effect of C19 infection on the human respiratory system, in turn causing changes in the production of speech, cough, and breathing sound, we compare the performance of classification systems built using the source and filter components of human produced audio signals. As shown in Figure 1, the public Coswara database [15], collected at IISc Bangalore India, is used for this comparative analysis. We present two different approaches for the separation of source and filter components. This is explained in detail in Section II. A neural network consumes either the source or the filter component of the human audio signal as features. This network intends to classify the samples of healthy from that of C19 positive individuals. Section IV presents the architecture details. We also compare the performance exhibited by speech, cough, and breathing sounds. The comparative results are as presented in Section V.

II. SOURCE-FILTER SEPARATION

The source of excitation leads to high frequency (HF) oscillations of the vocal cords; hence, the HF components (HFCs) of the speech signal represent the influence of the source of excitation on the produced speech signal. As these HFCs pass through the vocal tract, they are modulated by its filtering effects, thereby inducing low frequency components (LFCs) in the produced speech. To separate the two components, HFCs and LFCs, we bring the speech signal in a domain where they add up.

A. Cepstral Domain Separation

In the cepstral domain (CD), 40 mel filters convert the signal onto the Mel scale, where a filter bank is calculated as per Equation 1. As shown in Equation 2, a discrete cosine transform is performed to de-correlate the components obtained through the Mel filters. The initial 12–13 out of 40 coefficients thus obtained contain the LFCs reflecting the influence of the vocal tract filter properties. The later coefficients are usually discarded, as they contain the HFCs reflecting the influence of the source of excitation. In our analysis, we compare the performance of LFCs & HFCs in classifying C19 subjects from healthy ones.

$$Mel(f) = 2595 * \log(1 + (f/700)) \tag{1}$$

$$C(i) = \sqrt{2/N} \sum_{j=1}^{N} M^{j} \cos\left((\pi * i)/N * (j - 0.5)\right) \tag{2}$$

[1]Gauri Deshpande is working as Senior Scientist at TCS Research Pune, India and, pursuing PhD from University of Augsburg, Germany under the guidance of Prof. Björn W. Schuller. gauri1.d@tcs.com

[2]Björn W. Schuller is full professor and head of the chair of Embedded Intelligence for Health Care and Wellbeing, at the University of Augsburg, Germany. He is full professor of Artificial Intelligence and Head of GLAM - Group on Language, Audio & Music, at Imperial College London. schuller@ieee.org

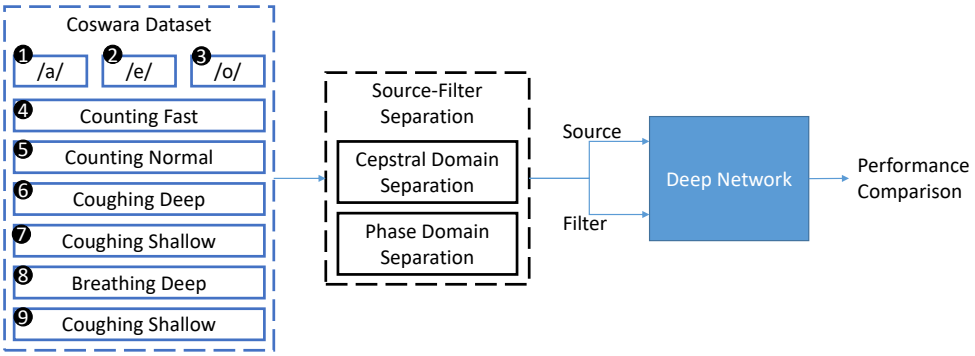

Fig. 1. Coswara data is available in nine different audio categories collected from COVID-19 and healthy subjects. In the present analysis, these are decomposed into source and filter components. These components' performance is compared using a neural network.

TABLE I

NUMBER OF SUBJECTS WITH DATA AVAILABLE IN EACH OF THE SEVEN CATEGORIES OF THE COSWARA DATABASE. THE TOTAL (COUNT) COLUMN INDICATES THE NUMBER OF SUBJECTS WITH DATA BELONGING TO THE HEALTHY AND COVID-19 CATEGORIES .

Audio Category	# Count	# Total
Healthy	1198	
No respiratory illness found	97	1372
Not exposed to respiratory illness	77	
Recovered	23	
Asymptomatic	14	
Mild positive	84	131
Moderate positive	10	

B. Phase Domain Separation

The authors of [16] propose phase domain (PD) separation of source and filter-based properties of a speech signal by passing the Hilbert transformed cepstral signal through a low pass filter. Further, the group delay functions of the LFCs and HFCs yield the filter and source components, respectively. In this study, the authors have mentioned that the CD separation leads to a loss of vocal tract information, where the PD separation performs better. The authors have improved the PD separation performance as explained in [17] by using a modified Hilbert transform where the log function is replaced by a generalized logarithmic function and a modified group delay function, where the sample difference operation is replaced by a regression filter.

C. Separation using a Zeros of Z Transform

More than a decade ago, the authors of [18] explained a method to separate the source and filter components of speech using zeros of the Z-transformed (ZZT) signal: They used a Discrete Fourier Transform (DFT) calculated from the zeros inside the unit circle for getting a vocal tract filter dominated spectrum and that from the zeros outside the unit circle for obtaining the glottal source dominated spectrum.

As explained by the authors of [18], this method is highly sensitive towards the glottal closure instance (GCI) synchronous windowing step. Also, as mentioned by the authors of [19], the method of decomposition using ZZT is functionally equivalent to the one exhibited by cepstrum based decomposition, where the later one is preferred for its high computation speed. For the reasons explained here, we are using CD and PD separation techniques in this paper.

III. DATABASE

The Coswara database [15] collected at IISc Bangalore, India, is used for the analysis presented in this paper. It consists of nine audio recordings from each participating subject, comprising of the three vowels /a/, /e/, and /o/, fast counting, slow counting, deep breathing, shallow breathing, deep cough, and shallow cough. The C19 infection status for each subject is given by one of the seven labels: 'healthy', 'no respiratory illness found', 'not exposed to respiratory illness', 'recovered', 'asymptomatic', 'mild positive', and 'moderate positive'. The data distribution among these categories is as shown in Table I.

For the binary classification of identifying C19 bio-markers, these seven categories merge to form two classes. The three categories, 'Healthy', 'No respiratory illness found', and 'Not exposed to respiratory illness' together form the "healthy" class. All other four categories belong to the "C19" class. As seen from the Table I, the two classes are highly imbalanced, with 131 subjects belonging to C19, and 1372 subjects belonging to the healthy class. Audio data augmenting techniques might lead to the loss of COVID-19 bio-markers, as they change the audio signal properties. Hence, only 10 % of the healthy class (comprising of all 97 subjects with the 'No respiratory illness found' label and 34 subjects with the 'Not exposed to respiratory illness' label) is used for classification, such that the classes are balanced.

IV. SOURCE FILTER PROPERTIES

In this section, we extract and compare the source and filter components obtained from two sources – the cepstral and the phase domain. As explained in Section II-A, the first 20 coefficients extracted using Mel scale filters are used as filter-component (vocal tract) features and later 20 coefficients as source-component (excitation) features. Similarly, the steps explained in Section II-B are applied for extracting source and filter components of the PD. In the PD, the source and filter components' feature vector length resembles 960. Of these 960, the initial 120 for the source as well as the central 120

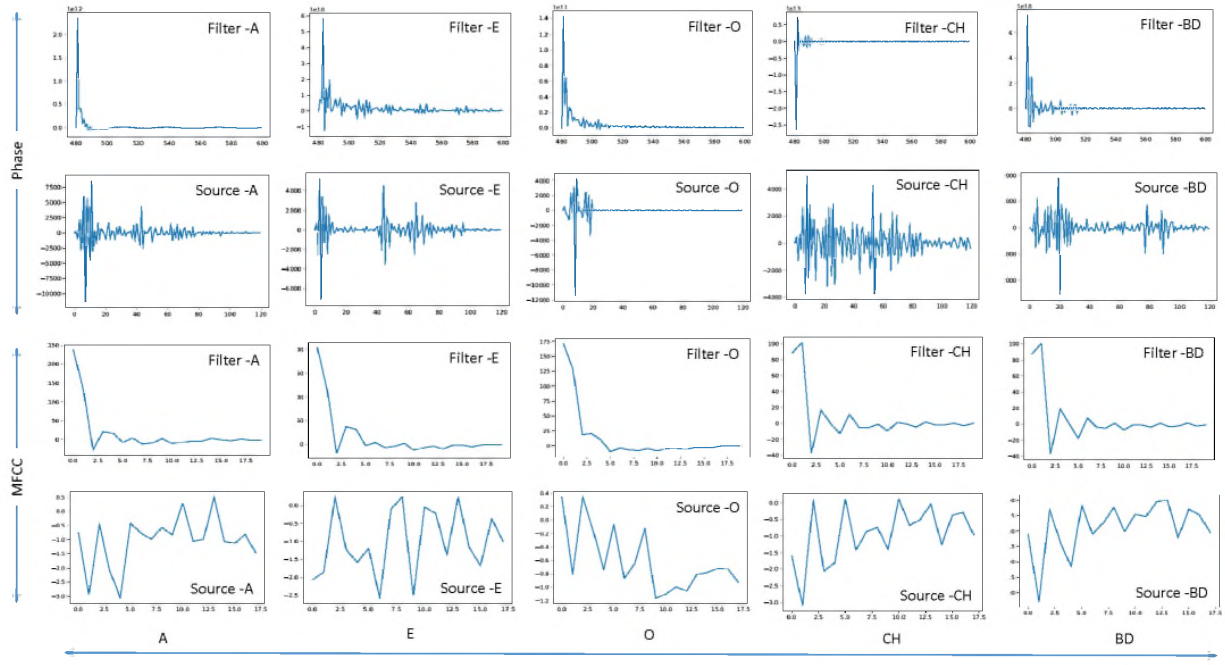

Fig. 2. Mean of Source and Filter components decomposed from 'moderate' COVID-19 infection status using phase (top two rows) and cepstral (bottom two rows) domain techniques. The audio categories are: A: vowel /a/, E: vowel /e/, O: vowel /o/, CH: Cough Heavy, BD: Breathing Deep.

from index 480 to 600 are found to carry useful information using principal component analysis. Figure 2 shows the source and filter components decomposed using the CD and the PD for vowels, cough, and breathing audio. These feature vectors of length 120 from the PD and 20 from the CD are fed into a neural network for binary classification between C19 and healthy subjects' audio. The neural network comprises of 1-dimensional convolution layers with 32 nodes followed by a Long Short-Term Memory (LSTM) layer of 16 nodes. This network has an output layer with 'sigmoid' activation. The network is trained using an Adam optimiser with a learning rate of 0.00008 for 55 epochs. The loss function used is 'binary cross-entropy'. The performance of the classifiers is measured in Area Under the Curve (AUC).

V. Results and Conclusion

As discussed in Section III, the data samples belong to one of the seven categories, which essentially maps to five distinct stages of C19 infection – 'recovered', 'asymptomatic', 'mild positive', 'moderate positive', and 'healthy'. The neural network explained in Section IV is trained with healthy samples as 'C19 negative samples' and all others as 'C19 positive samples' to detect healthy subjects from the subjects belonging to other stages of C19. The AUC for binary classification using different audio categories is as shown in Figure 3 (a). While comparing the performance of source and filter components, the filter components perform better than the source components for both MFCCs and the PD. As seen in the Figure, MFCCs appear more suited than the PD features across all nine audio categories, with an average improvement of around 16.6 % AUC. The combined feature set comprising of source and filter components of MFCCs performs better than their respective performance. However, for PD analysis, the combined feature set does not improve over the filter components. A maximum AUC of 88.1 % is achieved using the combined MFCC feature set for the vowel /e/, which is 31.6 % higher than that of using the combined PD feature set for the same vowel /e/. A minimum difference of 11.6 % AUC is found between the MFCC and PD performance using combined feature sets for heavy cough audio. Using PD analysis, the maximum performance of 66.1 % AUC is exhibited by filter components for the vowel /o/. The 95 % confidence interval calculated for AUCs across all the nine audio categories using PD analysis is $(55.3 - -65.4)$ and CD analysis is $(68.3 - -87.7)$. Another observation is, vowels exhibit the highest detection performance (maximum average AUC: 86 % using MFCCs combined, 63.4 % using the PD filter component) followed by breath signals (maximum average AUC: 77 % using MFCCs combined, 56.8 % using the PD filter component), and then counting (maximum average AUC: 70.1 % using MFCCs combined, 55.3 % using the PD filter component) & cough audio signals (maximum average AUC: 68.7 % using MFCCs combined, 55 % using the PD filter component). Figure 3 (b) shows the performance of binary detection of C19 healthy subjects from the subjects at each of the four C19 stages – 'recovered', 'asymptomatic', 'mild positive', and 'moderate positive'. Again, while comparing the performance of source and filter components, filter components perform better than the corresponding source components except for the moderate staged 'vowel /e/'. It is also seen from the Figure that C19 healthy subjects are classified with higher AUC from asymptomatic and recovered subjects as compared to moderate and mild subjects.

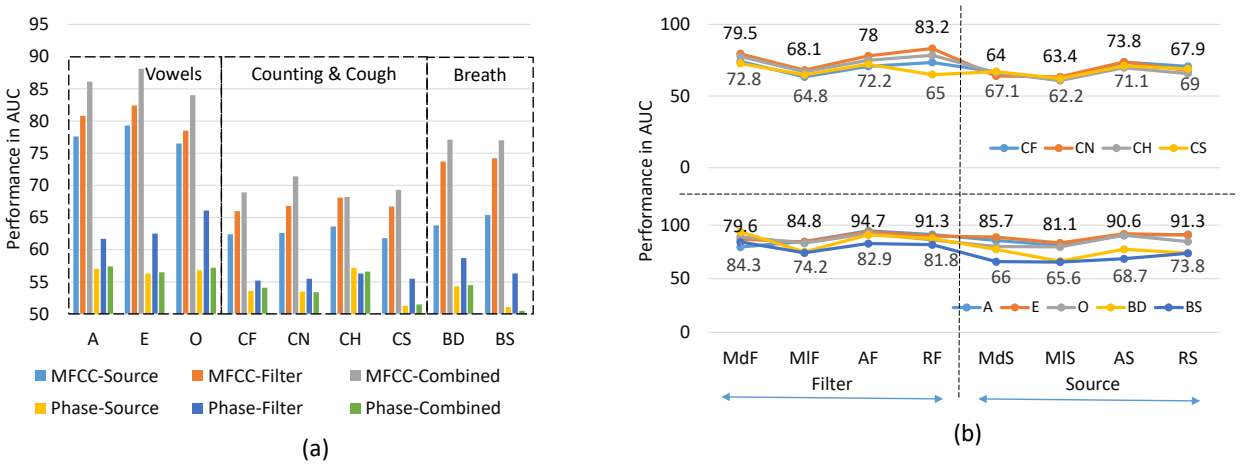

Fig. 3. (a) The binary classification results exhibited by the source and filter components decomposed using the cepstral domain and the phase domain. The audio categories are: A: vowel /a/, E: vowel /e/, O: vowel /o/, CF: Counting Fast, CN: Counting Normal, CH: Cough Heavy, CS: Cough Shallow, BD: Breathing Deep, BS: Breathing Shallow. (b) Binary classification results obtained between COVID-19 healthy subjects and other subjects at four different stages – MdF:Moderate Filter, MlF: Mild Filter, AF: Asymptomatic Filter, RF: Recovered Filter, MdS: Moderate Source, MlF: Mild Source, AS: Asymptomatic Source, RS: Recovered Source.

From the results, it is evident that C19 infection has a higher impact on the properties of vocal tract modulation than the source of excitation. This also reveals that the audio signals produced by asymptomatic and recovered subjects also carry the required bio-markers for C19 identification. More research is required to confirm these preliminary observations.

Acknowledgements

We thank all researchers, health supporters, and others helping in this crisis. Our hearts are with those affected and their families and friends. We acknowledge funding from the German BMWi by ZIM grant No. 16KN069402 (KIrun).

References

[1] W. Thorpe, M. Kurver, G. King, and C. Salome, "Acoustic analysis of cough," in *The 7th Australian and New Zealand Intelligent Information Systems Conference*. IEEE, 2001, pp. 391–394.

[2] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–541.

[3] V. C. Raykar, B. Yegnanarayana, S. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 751–761, 2005.

[4] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481–484, 2007.

[5] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America (JASA)*, vol. 126, no. 4, pp. 2061–2071, 2009.

[6] A. Bajpai and B. Yegnanarayana, "Combining evidence from subsegmental and segmental features for audio clip classification," in *IEEE Region 10 Conference TENCON*. IEEE, 2008, pp. 1–5.

[7] A. Chauhan, S. G. Koolagudi, S. Kafley, and K. S. Rao, "Emotion recognition using lp residual," in *IEEE Students Technology Symposium (TechSym)*. IEEE, 2010, pp. 255–261.

[8] R. Rajoo and R. A. Salam, "Performance of the vocal source related features from the linear prediction residual signal in speech emotion recognition," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 3-7, pp. 7–11, 2017.

[9] M. Meier, M. Borsky, E. H. Magnusdottir, K. R. Johannsdottir, and J. Gudnason, "Vocal tract and voice source features for monitoring cognitive workload," in *The 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2016, pp. 000 097–000 102.

[10] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system." in *The 20th annual conference of the international speech communication association (Interspeech)*. ISCA, 2019, pp. 739–743.

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

[12] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. Ji, T. Koike, X. Li, Z. Zhang, Y. Yamamoto, and B. W. Schuller, "An Early Study on Intelligent Analysis of Speech Under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety," in *The 21st annual conference of the international speech communication association (Interspeech)*. ISCA, 2020, pp. 4946–4950.

[13] M. B. Alsabek, I. Shahin, and A. Hassan, "Studying the similarity of covid-19 sounds based on correlation analysis of mfcc," in *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 2020, pp. 1–5.

[14] S. Deshmukh, M. A. Ismail, and R. Singh, "Interpreting glottal flow dynamics for detecting covid-19 from voice," *arXiv preprint arXiv:2010.16318*, 2020.

[15] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *The 21st annual conference of the international speech communication association (Interspeech)*. ISCA, 2020, pp. 4811–4815.

[16] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain," in *The 16th annual conference of the international speech communication association (Interspeech)*. ISCA, 2015, pp. 598–602.

[17] E. Loweimi, J. Barker, O. Saz-Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain." in *The 18th annual conference of the international speech communication association (Interspeech)*. ISCA, 2017, pp. 414–418.

[18] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, 2005.

[19] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," *arXiv preprint arXiv:1912.12602*, 2019.