Empirical Mode Decomposition Based Detection of Common Cold using Speech Signal

Pankaj Warule
Department of Electronics Engineering,
SV National Institute of Technology,
Surat, India
d20ec007@eced.svnit.ac.in

Siba Prasad Mishra
Department of Electronics Engineering,
SV National Institute of Technology,
Surat, India
ds20ec005@eced.svnit.ac.in

Suman Deb
Department of Electronics Engineering,
SV National Institute of Technology,
Surat, India
sumandeb@eced.svnit.ac.in

Deepak Joshi
Department of Electronics Engineering,
SV National Institute of Technology,
Surat, India
d.joshi@eced.svnit.ac.in

***Abstract*—This study investigates the discrimination between cold speech and healthy speech using features based on empirical mode decomposition (EMD). The EMD is employed to break down the signal into several intrinsic mode functions (IMFs). From each IMF, various statistical values like minimum, maximum, mean, standard deviation, first, second, and third quartiles, skewness, kurtosis, and energy of each IMF are extracted and used as a feature for distinguishing cold and healthy speech. The T-test examines the importance of EMD-based features for classifying cold speech. EMD-based feature performance is assessed using the deep neural network (DNN) classifier. The findings show that EMD-based features effectively discriminate between cold and healthy speech classes. Combining Mel-Frequency Cepstral Coefficients (MFCC) characteristics with EMD-based features improves the performance for identifying healthy and cold speech classes. On the URTIC database, the combination of MFCC and EMD-based features achieve a UAR of 66.92%.**

Index Terms: Common cold, Empirical mode decomposition, Deep neural network.

I. Introduction

The speech signal contains linguistic as well as paralinguistic information. The computational paralinguistic deals with nonverbal aspects of speech, such as emotion, physical and mental health detection. Speech is produced by the brain and respiratory systems. Therefore a variety of respiratory and brain-related illnesses may affect the acoustic parameters of speech signals. These changes in the acoustic parameters of speech caused by different health conditions may be evaluated using relevant speech features for disease diagnosis [1], [2], [3]. Pathology detection based on speech signals is gaining popularity since it is non-invasive and may be transmitted remotely with relative simplicity. Nasal congestion, runny nose, and a sore throat are all signs of the common cold [4]. The common cold has an effect on both the nose and the throat, and this has a knock-on effect on speech produced during a common cold. Speech that is considered to be cold is the speech of a person who is sick with a cold.

Cold speech analysis and categorization might help with the diagnosis of the common cold and its accompanying ailments. It might give useful information for remote health monitoring of patients. In general, normal or healthy speech is used for training in speech recognition and speaker recognition systems. If these systems are evaluated using cold speech, their performance can deteriorate. Therefore, cold speech analysis may be utilized to enhance the effectiveness of these man-machine interaction systems [5], [6].

Researchers have investigated the the common cold impact on speaker identification systems and the categorizing healthy and cold speech. Tull et al. [7] noted that there are differences in the Mel-Frequency Cepstral Coefficients (MFCC) for cold and healthy speech. The INTERSPEECH 2017 Cold Challenge was organized to detect a person with upper respiratory tract infections utilizing speech [8]. Suresh et al. [9] utilized a phoneme state posteriorgram (PSP) feature to classify common cold from speech using a Gaussian mixture model (GMM). Cai et al. [10] utilized the perception aware spectrum to diagnose the common cold. Deb et al. [11] decomposed speech signal into number of modes, and from each mode, various statistics are extracted and used as a feature for the classification of the common cold. Warule et al. [12] extracted MFCC features from vowel-like regions of speech for diagnosis of the common cold. Deb et al. [13] employed MFCC, linear prediction coefficients (LPC) features, and deep neural network (DNN) for distinguishing cold and healthy speech. Warule et al. [14] analyzed the importance of voiced and unvoiced speech segments for distinguishing healthy and cold speech. Warule et al. [15] employed the sinusoidal model-based features for distinguishing healthy and cold speech.

This study explored a novel feature extraction methodology utilizing empirical mode decomposition (EMD) for distinguishing healthy and cold speech. The EMD has been used successfully in various speech processing and classification applications. Khonglah et al. [16] used statistical characteristics based on EMD to distinguish between speech and music. Mainkar and Mahajan [17] employed EMD-based feature extraction for discriminating environmental sounds in the real world. Ravindran and Nair [18] classified pathological and normal speech using statistical features obtained from EMD and MFCC features. Sharma and Prasanna [19] explored the

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

effectiveness of EMD in characterizing glottal activity from voice signals.

The EMD decomposes the speech signal into the number of intrinsic mode functions (IMF). The literature review findings indicate that IMF provides important information for classifying speech signals. The IMFs or various frequency scales derived from the EMD of the speech signal provide discriminating information for differentiating the various speech classes. This motivates us to utilize EMD-based statistical measures to categorize healthy and cold speech.

In this study, we have used the Upper Respiratory Tract Infection Corpus (URTIC) database. The URTIC database was utilized for the cold sub-challenge of the 2017 INTERSPEECH Computational Paralinguistics Challenge [8]. The URTIC database contains speech recordings from 630 people (382 male and 248 female). The database has 28,652 speech samples with cold and healthy classes divided into train, develop, and test partitions. The train, develop and test partitions of the database consist of 9505, 9596, and 9551 speech samples, respectively.

This paper follows the following structure: Section II explores the empirical mode decomposition algorithm. The proposed methodology for categorizing healthy and cold speech is described in Section III . Section IV contains results and a discussion of the findings. The conclusion of the study is drawn in Section V.

II. Empirical Mode Decomposition

Huang et al. [20] proposed the empirical Mode Decomposition algorithm. The fundamental idea of EMD is to find appropriate time scales that reveal the physical properties of the signals and then decomposing the signal into modes known as IMF. These IMFs are signals that meet the following criteria:

1) The total zero crossings and extrema should be the same or differ by no more than one.
2) At any given point, the mean values of the envelope formed by local minima and maxima is zero.

The objective of EMD is to describe an arbitrary signal using a set of IMFs $m_i(n)$ and the residual signal r(n). Using EMD, the speech signal $s(n)$ is decomposed as

$$s(n) = r(n) + \sum_{i=1}^{M} m_i(n) \tag{1}$$

where M represents the total number of IMFs obtained, $r(n)$ is the residual signal and $m_i(n)$ is the imf of i^{th} mode. The algorithmic flow chart for EMD is shown in Fig. 1. The signal is decomposed into IMFs by identifying the speech signal's extrema points and constructing the lower and upper envelopes by interpolating the extrema points. The first IMF is derived by subtracting the mean of the lower and upper envelopes from the original signal. The residual component created by subtracting the computed IMF from the original signal is used as new data, and the procedure is repeated to determine the next IMF. The procedure is repeated till the residual signal turns into a monotonic function.

III. Methodology

Fig. 2 represents the block schematic of the proposed EMD-based framework for the healthy and cold speech classification. It includes pre-processing of the input speech signal,

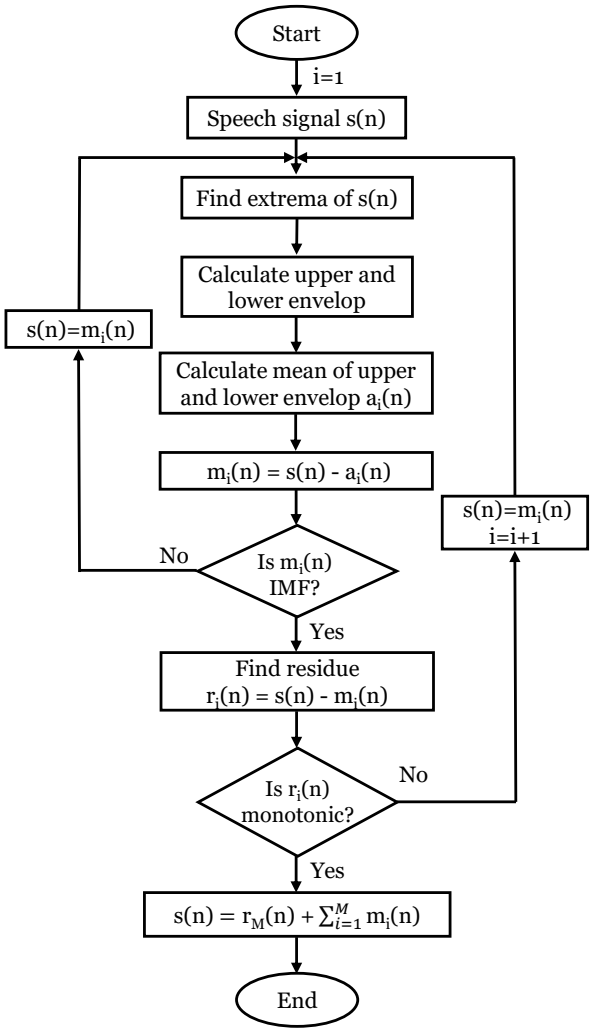

Fig. 1: The algorithmic flow chart for empirical mode decomposition.

Decomposition of the speech signal into several IMFs using EMD, statistical features extraction from each IMF, and DNN classifier for categorizing healthy and cold speech.

A. Pre-processing

Pre-processing comprises normalization, silence removal, framing, and windowing. Normalization of the speech signal is performed with reference to the maximum value. Then, utilizing short time energy, silence is eliminated from the speech signal [21]. Each speech is segmented into segments of 20 milliseconds with 10 milliseconds overlap. Then, the Hamming window is multiplied to each speech segment.

B. Empirical Mode Decomposition Based Feature Extraction

After pre-processing, EMD is performed on each speech segment as discussed in Section II to decompose it into a number of IMFs. Then from each IMF, following features are are extracted and used as a feature to classify cold and healthy speech.

1) Statistical features: The various statistical measures, including minimum, maximum, mean, standard deviation, first, second, and third quartiles, skewness, and kurtosis, are extracted from each IMF to capture the variations in IMF for

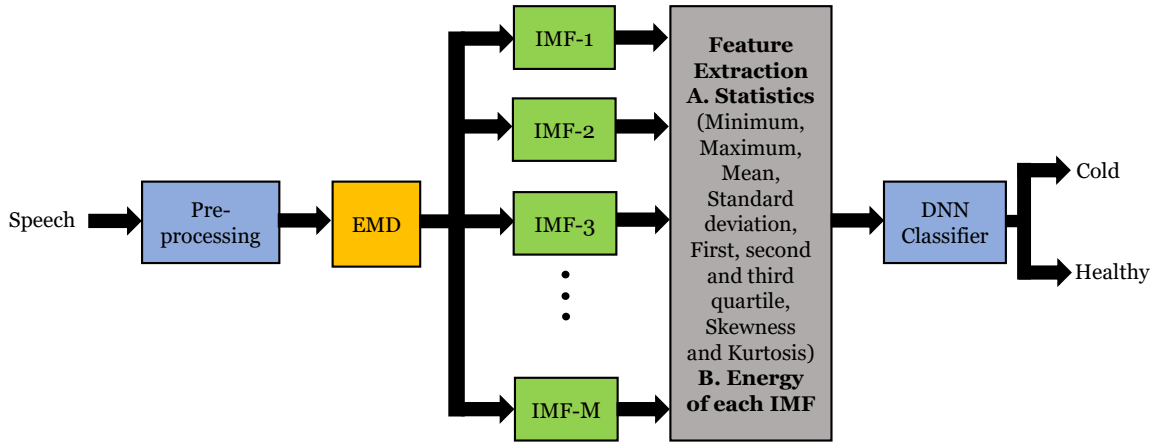

Fig. 2: Proposed EMD-based framework for distinguishing cold and healthy speech.

cold and healthy speech. Skewness is a statistical measure of the asymmetric distribution of data, and kurtosis is a statistic that helps identify whether the distribution has light-tailed or heavy-tailed compared to a normal distribution. Skewness S and kurtosis K are calculated as

$$S = \frac{1}{N} \frac{\sum_{n=1}^{N} (m_i(n) - \overline{m_i})^3}{s^3} \tag{2}$$

$$K = \frac{1}{N} \frac{\sum_{n=1}^{N} (m_i(n) - \overline{m_i})^4}{s^4} \tag{3}$$

where $m_i(n)$ is the i^{th} IMF, N denotes total number of samples in $m_i(n)$, $\overline{m_i}$ and s denote the mean and standard deviation of samples in $m_i(n)$.

2) Energy: The energy associated with each IMF is also considered as a feature for distinguishing cold and healthy speech. The energy of i^{th} IMF is given by

$$E_i = \sum_{n=1}^{N} |m_i(n)|^2 \tag{4}$$

where $m_i(n)$ is the i^{th} IMF and N is the samples in IMF.

C. Mel-Frequency Cepstral Coefficients (MFCC) Features

This study employs MFCC characteristics in addition to the proposed characteristics for classifying cold and healthy speech. The MFCC features are commonly used to distinguish between pathological and normal speech [22], [23], [24], [2], [14]. The morphology of the vocal tract influences which phonemes are produced during human speech. The MFCC features reflect the shape of the vocal tract for each spoken sound [25]. In order to extract MFCC characteristics from the speech, the number of speech frames that have a length of 20 milliseconds and an overlap of 50 percent are segmented. After that, the DFT is applied to every frame in order to obtain the power spectrum. After that, mel-scale filter banks are utilized in order to process the power spectrum. Finally, the discrete cosine transform (DCT) is used to get the MFCC values after converting the power spectrum to the log domain. In this study, the 13 MFCC coefficients and the first and second order MFCC difference (ΔMFCC & $\Delta\Delta$MFCC) are extracted from every single speech frame.

D. Deep Neural Network (DNN)

The DNN has been found effective in speech related applications like natural language processing, speech recognition, and speech pathology detection [26], [23], [27]. In this study, we have employed two hidden layers of DNN with 128 and 32 neurons. At the hidden layers of a DNN, the rectified linear unit (ReLU) activation is employed, whereas the sigmoid activation is used at the output layer.

The performance of DNN is evaluated using unweighted average recall (UAR). As the URTIC database is extremely unbalanced, the identification rates of both the cold and healthy classes are crucial. The UAR is calculated by taking the average of recalls of cold and healthy class.

IV. Results & Discussion

This section analyzes the statistical significance and performance of EMD-based features for categorizing healthy and cold speech classes. The effectiveness of the proposed framework is also evaluated using the combination of MFCC and EMD-based features. The achieved results are compared with the results of the state-of-the-art (SOTA) methods.

To analyze the statistical significance of EMD-based features, T-test [28] is conducted on EMD-based features. A T-test is a way to compare the means of two groups. The t-value and the p-value are calculated for each feature in the T-test. The feature has a greater t-value and a lower p-value (less than 0.0001), indicating that it is more relevant for classification. The t-values and the p-values between cold and healthy speech classes are calculated for the EMD-based features extracted from all speech samples of the train partition, as shown in Table I. The obtained t-values are higher, and the corresponding p-values are less than 0.0001, with a few exceptions. This reveals that the EMD-based features successfully categorize cold and healthy speech classes. Therefore, these features can be used for distinguishing cold and healthy speech classes.

To evaluate the proposed framework, EMD-based features and MFCC features are extracted from each speech signal of the URTIC database as discussed in Section III-B. The features extracted from the train partition are used for training, and feature extracted from the develop partition are used for testing the DNN.

TABLE I: The results of the T-test for the EMD-based features.

Fearure	IMF-1		IMF-2		IMF-3	
	t-value	p-value	t-value	p-value	t-value	p-value
Minimum	3.9898	<0.0001	5.3766	<0.0001	3.6488	0.000265
Maximum	4.1225	<0.0001	5.3394	<0.0001	3.2014	0.0013
Mean	3.9378	<0.0001	5.5140	<0.0001	4.4017	<0.0001
Standard deviation	3.4465	0.0005	5.6579	<0.0001	4.2056	<0.0001
First quartile	2.4377	0.0147	5.7004	<0.0001	5.1140	<0.0001
Second quartile	1.2643	0.2061	4.1236	<0.0001	4.1769	<0.0001
Third quartile	2.6001	0.0093	5.3744	<0.0001	4.5759	<0.0001
Skewness	4.5417	<0.0001	3.9955	<0.0001	7.7153	<0.0001
Kurtosis	2.4652	0.0137	6.9877	<0.0001	9.0467	<0.0001
Energy	0.3460	0.7292	4.4808	<0.0001	5.2208	<0.0001

The confusion matrices in % for the classification results achieved using the EMD-based features, MFCC features, and combination of MFCC and EMD-based features are shown in Figs. 3a, 3b, and 3c, respectively. The EMD-based statistical features extracted from the IMFs achieve the UAR of 64.14% with recalls for healthy and cold classes are 59.84% and 68.44%, respectively. The MFCC features achieve the UAR of 65.28% with recalls for healthy and cold classes are 71.36% and 59.21%, respectively. It is observed that the

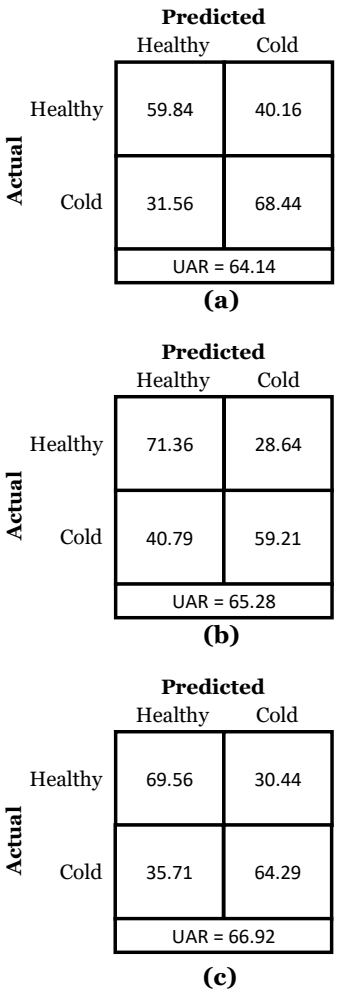

Fig. 3: The confusion matrices (%) of DNN for (a) EMD based features, (b) MFCC features, (c) MFCC + EMD based features.

TABLE II: Performance of proposed framework using DNN classifier on the URTIC database.

Feature	% UAR
EMD	64.14
MFCC	65.28
EMD + MFCC	66.92

UAR achieved using MFCC features is higher than EMD-based features. The EMD-based features provide a greater recall for the cold class, and MFCC features provide a greater recall for the healthy class. The achieved UAR is improved up to 66.92% using a combination of MFCC and EMD-based features. The combination of MFCC and EMD-based features gives 69.56% recall for the healthy class and 64.29% recall for the cold class. Table II shows the effectiveness of the proposed framework using the DNN classifier for distinguishing cold and healthy speech on the URTIC database.

TABLE III: The performance evaluation of the proposed framework with the SOTA methods.

Research Work	%UAR
ComParE features + SVM [8]	64.00
ComParE BoAW features + SVM [8]	64.20
MFCC features + GMM [10]	64.80
CQCC features + GMM [10]	65.40
PSP features + SVM [9]	64.00
MOD features + DNN [29]	67.95
VMD features + SVM [11]	66.84
VLR MFCC features + DNN [12]	61.93
Proposed EMD features + MFCC + DNN	66.92

The performance evaluation of the proposed framework with the SOTA methods is given in Table III. Using the ComParE and BoAW features, baseline performances of 64% and 64.20% UAR were obtained for INTERSPEECH 2017 Cold Challenge [8]. Cai et al. [10] obtained 64.80% and 65.40% UAR, respectively, using constant Q cepstral coefficients (CQCC) and MFCC features. Suresh et al. [9] obtained a UAR of 64% using PSP features and GMM. Using spectral modulation feature (MOD), Huckvale and Beke [29] obtained a UAR of 67.95%. Deb et al. [11] used variational mode decomposition (VMD)-based features to reach a UAR of 66.84%. Warule et al. [12] achieved a UAR of 61.93% using vowel-like region (VLR) MFCC features. In this research,

we got comparable outcomes with the SOTA methods. The combination of MFCC and EMD-based features provides a UAR of 66.92%.

V. Conclusion

In this investigation, we have used the EMD-based framework for categorizing healthy and cold speech. The EMD decomposed the speech signal into several IMFs. Then for each IMF various statistical parameters and energy of each IMF are calculated and used as features for classification. Statistical analysis using the T-test shows that the EMD-based features can discriminate between cold and healthy speech classes. The DNN classifier is used to evaluate the effectiveness of EMD-based features. The effectiveness of the proposed framework is improved when MFCC and EMD-based features are used together for distinguishing cold and healthy speech classes.

References

[1] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[2] S. S. Nayak, A. D. Darji, and P. K. Shah, "Machine learning approach for detecting covid-19 from speech signal using mel frequency magnitude coefficient," *Signal, Image and Video Processing*, pp. 1–8, 2023.

[3] P. Warule, S. P. Mishra, and S. Deb, "Time-frequency analysis of speech signal using chirplet transform for automatic diagnosis of parkinson's disease," *Biomedical Engineering Letters*, pp. 1–11, 2023.

[4] D. E. Pappas, "The common cold," *Principles and practice of pediatric infectious diseases*, p. 199, 2018.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[7] R. G. Tull and J. C. Rutledge, "Analysis of "cold-affected"speech for inclusion in speaker recognition systems." *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2549–2574, 1996.

[8] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[9] A. K. Suresh, S. R. KM, and P. K. Ghosh, "Phoneme state posteriorgram features for speech based automatic classification of speakers in cold and healthy condition." in *INTERSPEECH*, 2017, pp. 3462–3466.

[10] D. Cai, Z. Ni, W. Liu, W. Cai, G. Li, M. Li, D. Cai, Z. Ni, W. Liu, and W. Cai, "End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum." in *INTERSPEECH*, 2017, pp. 3452–3456.

[11] S. Deb, S. Dandapat, and J. Krajewski, "Analysis and classification of cold speech using variational mode decomposition," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 296–307, 2017.

[12] P. Warule, S. P. Mishra, and S. Deb, "Classification of cold and non-cold speech using vowel-like region segments," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2022, pp. 1–5.

[13] S. Deb, P. Warule, A. Nair, H. Sultan, R. Dash, and J. Krajewski, "Detection of common cold from speech signals using deep neural network," *Circuits, Systems, and Signal Processing*, pp. 1–16, 2022.

[14] P. Warule, S. P. Mishra, and S. Deb, "Significance of voiced and unvoiced speech segments for the detection of common cold," *Signal, Image and Video Processing*, pp. 1–8, 2022.

[15] P. Warule, S. P. Mishra, S. Deb, and J. Krajewski, "Sinusoidal model-based diagnosis of the common cold from the speech signal," *Biomedical Signal Processing and Control*, vol. 83, p. 104653, 2023.

[16] B. K. Khonglah, R. Sharma, and S. M. Prasanna, "Speech vs music discrimination using empirical mode decomposition," in *2015 Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[17] S. D. Mainkar and S. Mahajan, "Emd based efficient discrimination of real-world environmental sounds using svm classifier," in *2015 International Conference on Information Processing (ICIP)*. IEEE, 2015, pp. 272–277.

[18] P. Ravindran and V. V. Nair, "Analysis of vocal tract disorders using mel-frequency cepstral coefficients and empirical mode decomposition based features," in *2015 International Conference on Control Communication & Computing India (ICCC)*. IEEE, 2015, pp. 505–510.

[19] R. Sharma and S. M. Prasanna, "Characterizing glottal activity from speech using empirical mode decomposition," in *2015 Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[20] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[21] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[22] A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Bhat, S. R. Chetupalli, S. Ganapathy, S. Ramoji *et al.*, "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[23] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66 749–66 776, 2020.

[24] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam, "Vocal fold disorder detection based on continuous speech by using mfcc and gmm," in *2013 7th IEEE GCC Conference and Exhibition (GCC)*. IEEE, 2013, pp. 292–297.

[25] S. P. Mishra, P. Warule, and S. Deb, "Deep learning based emotion classification using mel frequency magnitude coefficient," in *2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)*. IEEE, 2023, pp. 93–98.

[26] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: a preliminary study," in *2017 international conference and workshop on bioinspired intelligence (IWOBI)*. IEEE, 2017, pp. 1–4.

[27] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[28] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology*, vol. 68, no. 6, pp. 540–546, 2015.

[29] M. A. Huckvale and A. Beke, "It sounds like you have a cold! testing voice features for the interspeech 2017 computational paralinguistics cold challenge." International Speech Communication Association (ISCA), 2017.