Hilbert-Huang transform-based time-frequency analysis of speech signals for the identification of common cold

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—**The current advancements in machine learning research pertaining to speech and health are highly interesting. One aspect of speech-processing research that is gaining popularity is the use of computational paralinguistic analysis to evaluate a variety of health conditions. In this study, we have used the Hilbert-Huang transform (HHT) for the time-frequency analysis of speech signals for the identification of the common cold. The HHT is a time-frequency transform that is adaptive and ideal for non-linear and non-stationary signals. The HHT is a combination of empirical mode decomposition (EMD) and the Hilbert transform (HT). The HHT gives the time-frequency representation (TFR) matrix of the speech signal. Then, the entropy of each frequency component in TFR is computed and used as a distinguishing feature between cold and healthy speech. The efficacy of the proposed methodology is evaluated on the URTIC dataset using a deep neural network. The proposed features achieve UARs of 65.66% and 65.26%, respectively, on the develop and test partitions. The results of the study demonstrate that the time-frequency entropy features extracted using the HHT are effective in distinguishing between cold and healthy speech.**

Index Terms: Common cold, Empirical mode decomposition, Hilbert-Huang transform, Deep neural network.

I. Introduction

A speech signal is known to encompass a diverse range of data pertaining to the speaker. The data presented comprises the linguistic elements associated with the speaker's intended message and paralinguistic characteristics such as their present health and emotional state, age, and gender [1], [2]. The contemporary developments in machine learning research concerning speech and health are exceedingly interesting. The utilization of computational paralinguistic analysis to assess different health conditions is an increasingly popular area of interest in speech-processing research. The utilization of speech signals for the purpose of detecting pathology is becoming more prevalent due to their non-invasive nature and the ease with which they can be transmitted remotely. A common cold is characterized by nasal congestion, a runny nose, and a sore throat [3]. The common cold affects the nasal passages and the throat, affecting the person's ability to articulate words clearly. These speech alterations can be quantified and used as appropriate characteristics to identify a common cold.

The analysis and classification of cold speech may aid in the diagnosis of the common cold and its associated maladies. It could potentially provide valuable insights for the remote health monitoring of patients. Normal or healthy speech is commonly utilized for training speech recognition and speaker recognition systems. When these systems are tested using cold speech, their performance may suffer. Hence, cold speech analysis might be used to improve the efficacy of these man-machine interface systems [4], [5].

Researchers explored the common cold's impact on speaker recognition systems as well as the classification of healthy and cold speech. Tull et al. [6] noticed that there are distinguishable variations in the MFCC between cold and healthy speech. The INTERSPEECH 2017 Cold Challenge had the objective of identifying individuals suffering from upper respiratory tract illnesses such as the common cold using speech [7]. Cai et al. [8] recognized the common cold using the perception-aware spectrum. Gosztolya et al. [9] used the output of frame-level classification achieved using DNN to get utterance-level features for cold and healthy speech categorization. Suresh et al. [10] used a phoneme state posteriorgram (PSP) feature with a Gaussian mixture model (GMM) to classify common cold from speech. Huckvale and Beke [11] analyzed the performance of various voice quality features (VOI) features for discrimination between cold and healthy speech. Deb et al. [12] divided voice signal into a number of modes, from which various statistics were obtained and employed as a feature for the categorization of the common cold. Vicente et al. [13] used SVM to classify cold speech using the MFCC Fisher vector (FV). Warule et al. [14] utilized vowel-like regions (VLR) MFCC features for categorizing cold and healthy speech. Deb et al. [15] combined MFCC, LPC features, and DNN for

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

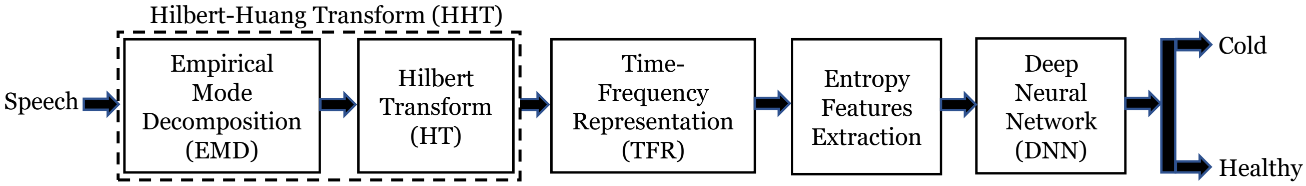

Hilbert-Huang Transform (HHT)
Speech
Empirical Mode Decomposition (EMD)
Hilbert Transform (HT)
Time-Frequency Representation (TFR)
Entropy Features Extraction
Deep Neural Network (DNN)
Cold
Healthy

Fig. 1: Proposed HHT-based framework for categorizing healthy and cold speech.

categorizing healthy and cold speech. Warule et al. [16] investigated the role of voiced and unvoiced speech segments for categorizing healthy and cold speech. Warule et al. [17] used sinusoidal model-based features for categorizing healthy and cold speech.

This study explored a novel feature extraction methodology utilizing the Hilbert-Huang transform (HHT) for distinguishing healthy and cold speech. The HHT has been shown to be effective in a variety of speech processing and categorization applications. The HHT is a new and strong time-frequency analysis theory that is effective in describing the local properties of non-linear and non-stationary signals [18], [19]. Karan et al. [20] utilized the HHT in order to automate the detection and assessment of speech associated with Parkinson's disease. Turan et al. [21] used the HHT to classify ingesting sounds captured by the throat microphone. Liu et al. [22] utilized HHT-based time-frequency analysis for depression identification in speech.

In this study, we have used a time-frequency representation (TFR) matrix of speech signals achieved using the HHT. The entropy of each frequency component in TFR is then computed and utilized as a distinguishing feature between cold and healthy speech. We have given some thought to the possibility that the information provided by the entropy of the frequency components in the TFR matrix of the speech signal can be used to distinguish between cold and healthy speech. The efficacy of the proposed methodology is evaluated on the URTIC dataset utilizing a deep neural network.

We used the upper respiratory tract infection corpus (URTIC) database in this investigation. The URTIC database has been used for the cold sub-challenge of the 2017 INTERSPEECH computational paralinguistics challenge [7]. Speech recordings from 630 people (382 men and 248 women) are available in the URTIC database. The database comprises a total of 28,652 speech samples, which have been categorized into cold and healthy classes. Only 10% of the samples belong to cold classes, indicating that the database is significantly unbalanced. The database has been partitioned into three subsets, namely train, develop, and test, comprising 9505, 9596, and 9551 speech samples, respectively.

This paper follows the format outlined below: In Section II, the proposed methodology for categorizing healthy and cold speech is described. Section III includes the results and an analysis of the findings. The study's conclusion is presented in Section IV.

II. Methodology

It includes the decomposition of the speech signal into several IMFs using EMD, Hilbert transform (HT), extraction of the entropy features from TFR achieved using HHT, and a DNN classifier for categorizing healthy and cold speech.

A. Hilbert-Huang transform

This section describes the evaluation procedure for the Hilbert-Huang transform. The HHT is a modern signal analysis tool that employs an empirical approach to signal processing. The technique encompasses two significant processes: empirical mode decomposition (EMD) and Hilbert transform (HT) [23]. The initial step in conducting the HHT involves executing the EMD technique to decompose the signal into intrinsic mode functions (IMFs). Subsequently, the Hilbert Spectrum of the IMFs is computed in order to derive the instantaneous frequency.

1) Empirical Mode Decomposition: The EMD algorithm was proposed by Huang et al. [24]. The EMD decomposes the signal into several modes known as IMFs. These IMF signals meet the following requirements:

- It is expected that the total number of zero crossings and extrema will either be equal or exhibit a difference of no more than one.
- The mean values of the envelope formed by local minima and maxima are zero at every given position.

The goal of EMD is to describe any signal using a set of IMFs $m_i(t)$ and the residual signal $r(t)$. The speech signal $s(t)$ is decomposed using EMD as

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

where $r(t)$ denotes the residual signal and $m_i(t)$ denotes the imf of the i^{th} mode. The signal is decomposed into IMFs by finding the extrema points of the speech signal and then forming the lower and upper envelopes by interpolating the extrema points. The mean of the lower and higher envelopes is subtracted from the original signal to get the initial IMF. The residual component obtained by subtracting the calculated IMF from the original signal is utilized as new data, and the method is repeated to compute the next IMF. The process is continued until the residual signal becomes a monotonic function.

2) Hilbert Transform (HT): It reflects the instantaneous energy distribution of IMFs in TFR [25]. The Hilbert transform of the signal $s(t)$ is given by

$$H(t) = \frac{Q}{\pi}\int_{-\infty}^{\infty} \frac{s(\tau)}{t-\tau} d\tau \tag{2}$$

where Q is the Cauchy principal integral value. The convolution of the signal $s(t)$ with $1/t$ gives the HT, as shown in Eq. (2). As a result, the HT can identify the local features of $s(t)$ [26].

The analytic signal $z(t)$ is represented by

$$z(t) = s(t) + jH(t) = \alpha(t)e^{j\phi(t)} \tag{3}$$

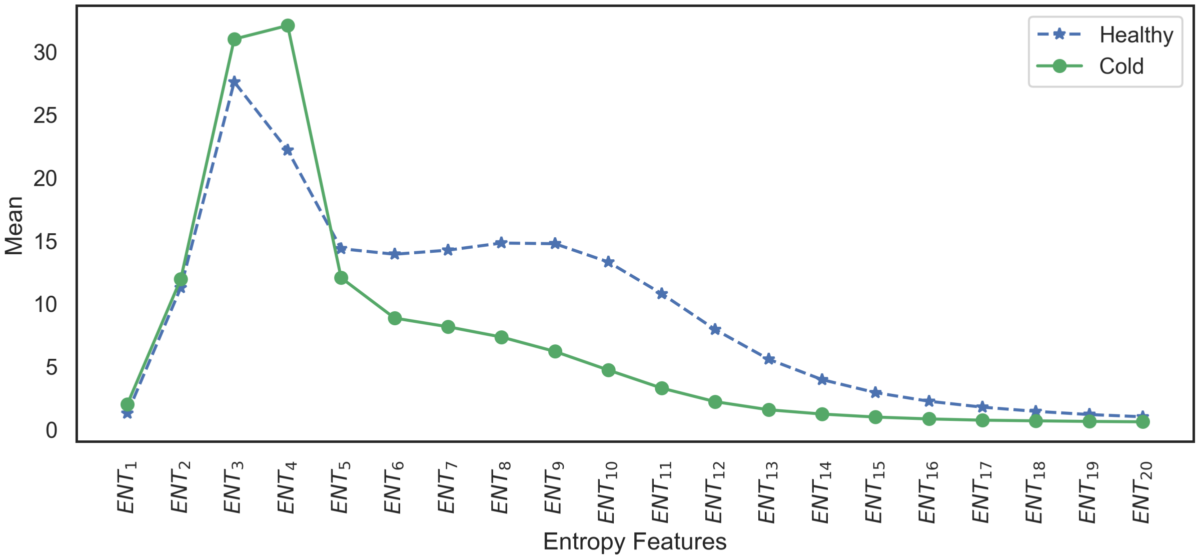


Fig. 2: Mean values of first 20 ENT features for healthy and cold speech classes.

where $\alpha(t)$ is the instantaneous amplitude, and $\phi(t)$ is the instantaneous phase can be given as

$$\alpha(t) = \sqrt{s^2(t) + H^2(t)} \tag{4}$$

$$\phi(t) = arctan\left(\frac{H(t)}{s(t)}\right) \tag{5}$$

The instantaneous frequency $\omega(t)$ is given by

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{6}$$

Finally, speech signal $s(t)$ can be expressed in terms of instantaneous amplitude and frequency as

$$s(t) = \sum_{i=1}^{M} \alpha_i(t) e^{j\int \omega_i(t)dt} \tag{7}$$

The amplitude distribution of the signal at various frequencies over time gives the time-frequency representation (TFR) of speech signals. This TFR, achieved using HHT, is used to extract the features for categorizing healthy and cold speech.

B. Feature Extraction

In this study, we have calculated the entropy of every frequency component in the HHT-based TFR of the speech signals and used them as discriminating features for cold and healthy speech classes. The histogram of frequency components is used to calculate the entropy of frequency components in the TFR. The probability value of the r^{th} frequency component is computed as [27], [28]

$$P_b(r) = \frac{h_b(r)}{\sum_{b=1}^{B} h_b(r)} \tag{8}$$

where $h_b(r)$ represents the histogram of the r^{th} frequency component and B is the total number of bins. The entropy of r^{th} frequency component is given by [29]

$$E_r = -\sum_{b=1}^{B} P_b(r) log_2\big[P_b(r)\big] \tag{9}$$

In this work, we have extracted entropy (ENT) features for the first 100 frequency components to form a 100-dimensional feature vector [ENT_1, ENT_2, ENT_3, ..., ENT_{100}] for each speech recording.

C. Deep Neural Network (DNN)

The DNN has been found effective in speech related applications like natural language processing, speech recognition, and speech pathology detection [30], [31], [32], [33]. In this study, we have employed three hidden layers of DNN with 256, 128, and 64 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.

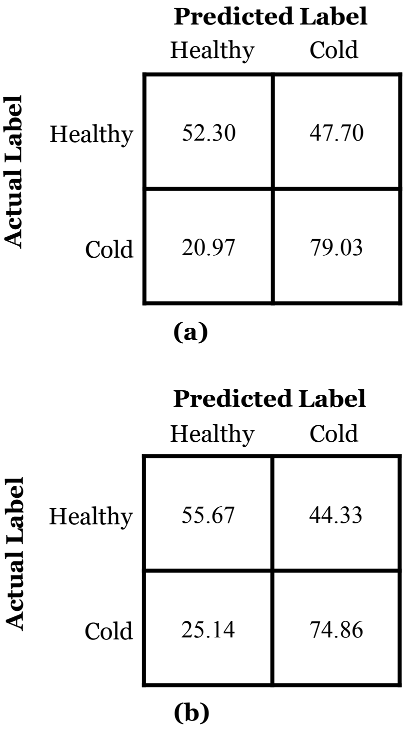
With the highly imbalanced nature of the URTIC database, accurate identification of both the cold and healthy classes is of vital importance, the performance of DNN is measured using unweighted average recall (UAR). The UAR is computed by determining the mean of the recall values for the cold and healthy classes.

TABLE I: Performance of HHT-based entropy features using DNN classifier on the URTIC database.

	Develop partition	Test partition
Healthy class recall (%)	52.30	55.67
Cold class recall (%)	79.03	74.86
UAR (%)	65.66	65.26

III. Results & Discussion

This section examines the efficacy of proposed HHT-based entropy features for categorizing healthy and cold speech classes. The efficacy of the proposed features is evaluated on the URTIC database using the DNN classifier. While evaluating the develop partition, the training partition is utilized for training, and both the training and develop partitions are utilized for training while evaluating the test



	Healthy	Cold
Healthy	52.30	47.70
Cold	20.97	79.03

	Healthy	Cold
Healthy	55.67	44.33
Cold	25.14	74.86

Fig. 3: Confusion matrix (%) of the DNN on (a) Develop partition, and (b) Test partition of the URTIC database.

partition. The obtained results are contrasted with the state-of-the-art (SOTA) method's results.

The mean values of the first 20 ENT features extracted from the train partition of the URTIC database are presented in Fig. 2 to examine the importance of the proposed features for categorizing healthy and cold speech classes. It is observed that there is a significant difference in the mean values of the proposed features between the cold and healthy speech classes. Therefore, these features can be employed to categorize these classes.

The results obtained utilizing the proposed features for categorizing healthy and cold speech are shown in Table I. The confusion matrices in % for the classification results obtained using the proposed HHT-based entropy features are depicted in Figs. 3a and 3b, respectively. The proposed features achieve an UAR of 65.66% with recalls for healthy and cold classes are 52.30% and 79.03%, respectively, on the develop partition. Similarly, it achieve the UAR of 65.26% with recalls for healthy and cold classes are 55.67% and 74.86%, respectively, on the test partition.

The performance comparison of the proposed method and SOTA methodologies is presented in Table II. Cai et al. [8] achieved 64.80% and 65.40% UAR, respectively, using constant Q cepstral coefficients (CQCC) and MFCC features on the develop partition. Gosztolya et al. [9] achieved 65% UAR using DNN-based frame-level features. Using PSP features and GMM, Suresh et al. [10] achieved a UAR of 64%. Huckvale and Beke [11] achieved UARs of 65.58% and 62.10%, respectively, on develop and test partitions using VOI features. Deb et al. [12] employed VMD-based features and achieved a UAR of 66.84%. Warule et al. achieved a UAR of 61.93% using VLR MFCC features on develop partition. Warule et al. [16] achieved UARs of 66.12% and 64.92% using statistics of MFCC features. In this study, we got comparable outcomes with the SOTA methods. The proposed features achieve UARs of 65.66% and 65.26%, respectively, on the develop and test partitions.

In this investigation, we have used the TFR of speech signals achieved using HHT for categorizing healthy and cold speech. In SOTA methods, first a voiced speech region is detected, then voiced speech is segmented into frames, and features are extracted for classification. But, in this study, we have directly calculated the TFR of speech, and the entropy of each frequency component over time is calculated to get discriminating features for classification. This significantly reduces the complexity of the speech pathology detection system.

IV. Conclusion

The Hilbert-Huang Transform (HHT) is an efficient approach for analyzing nonlinear and nonstationary signals. In this investigation, we have used the HHT-based time-frequency representation of speech signals for categorizing healthy and cold speech. The results reveal that the proposed HHT-based time-frequency analysis method effectively classifies healthy and cold speech. This method reduces system

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

Research work	%UAR	
	Develop partition	Test Partition
MFCC features + GMM [8]	64.80	-
CQCC features + GMM [8]	65.40	-
MFCC features + DNN [9]	65.00	-
PSP features + SVM [10]	64.00	61.09
VOI features + DNN [11]	65.58	62.10
VMD features + SVM [12]	66.84	-
MFCC Fisher Vectors + SVM [13]	63.98	66.12
VLR MFCC features + DNN [14]	61.93	-
MFCC statistics+ SVM [16]	66.12	64.92
Proposed HHT-based entropy features + DNN	65.66	65.26

complexity because, unlike traditional speech processing algorithms, it doesn't include any pre-processing steps like silence removal, framing, or windowing prior to feature extraction. In future work, we will apply the proposed HHT-based method to classify other pathological conditions and speech emotion recognition.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "Dnn-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," 2017.

[10] 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.

[11] 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.

[12] 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.

[13] E. L. José Vicente and G. Gosztolya, "Using the fisher vector approach for cold identification," *Acta Cybernetica*, vol. 25, no. 2, pp. 223–232, 2021.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] X. Li, X. Zou, R. Zhang, and G. Liu, "Method of speech enhancement based on hilbert-huang transform," in *2008 7th World Congress on Intelligent Control and Automation*. IEEE, 2008, pp. 8419–8424.

[19] S. S. Hanna, N. Korany, and M. B. Abd-el Malek, "Speech recognition using hilbert-huang transform based features," in *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2017, pp. 338–341.

[20] B. Karan, S. S. Sahu, J. R. Orozco-Arroyave, and K. Mahto, "Hilbert spectrum analysis for automatic detection and evaluation of parkinson's speech," *Biomedical Signal Processing and Control*, vol. 61, p. 102050, 2020.

[21] M. T. Turan and E. Erzin, "Classification of ingestion sounds using hilbert-huang transform," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2017, pp. 1–4.

[22] Z. Liu, Y. Xu, Z. Ding, and Q. Chen, "Time-frequency analysis based on hilbert-huang transform for depression recognition in speech," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 1072–1076.

[23] N. E. Huang, *Hilbert-Huang transform and its applications*. World Scientific, 2014, vol. 16.

[24] 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.

[25] R. Sharma, R. K. Bhukya, and S. M. Prasanna, "Analysis of the hilbert spectrum for text-dependent speaker verification," *Speech Communication*, vol. 96, pp. 207–224, 2018.

[26] Z. Peng, W. T. Peter, and F. Chu, "An improved hilbert–huang transform and its application in vibration signal analysis," *Journal of sound and vibration*, vol. 286, no. 1-2, pp. 187–205, 2005.

[27] R. K. Tripathy, M. R. Paternina, J. G. Arrieta, A. Zamora-Méndez, and G. R. Naik, "Automated detection of congestive heart failure from electrocardiogram signal using stockwell transform and hybrid classification scheme," *Computer methods and programs in biomedicine*, vol. 173, pp. 53–65, 2019.

[28] 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.

[29] S. K. Ghosh, R. Ponnalagu, R. Tripathy, and U. R. Acharya, "Automated detection of heart valve diseases using chirplet transform and multiclass composite classifier with pcg signals," *Computers in biology and medicine*, vol. 118, p. 103632, 2020.

[30] 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.

[31] 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.

[32] 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.

[33] 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)*, 2023, pp. 93–98.